TECHNISCHE UNIVERSITÄT BERLIN
FAKULTÄT FÜR ELEKTROTECHNIK UND INFORMATIK
LEHRSTUHL FÜR INTELLIGENTE NETZE
UND MANAGEMENT VERTEILTER SYSTEME

Understanding Benefits of Different Vantage Points in Today's Internet

vorgelegt von
Jan Böttger (M.Sc.)
geb. in Berlin

Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
– DOKTOR DER INGENIEURWISSENSCHAFTEN (DR.-ING.) –
genehmigte Dissertation

Promotionsausschuss:

Vorsitzender	Prof. Dr. Axel Küpper, Technische Universität Berlin
Gutachterin:	Prof. Anja Feldmann, Ph.D., Technische Universität Berlin
Gutachter:	Dr. Walter Willinger, NIKSUN (USA)
Gutachter:	Prof. Dr. Odej Kao, Technische Universität Berlin
Gutachter:	Prof. Dr. Jean-Pierre Seifert, Technische Universität Berlin

Tag der wissenschaftlichen Aussprache: 24.10.2016

Berlin 2017

Eidesstattliche Erklärung

Ich versichere an Eides statt, dass ich diese Dissertation selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

Datum Jan Böttger (M.Sc.)

Abstract

Measuring the Internet is indispensable to a better understanding of its current state and trends, but obtaining measurements is difficult both qualitatively and quantitatively. The challenges of "Internet Measurement" are manifold due to the nature of the beast, namely its complexity, distribution, and constant change. The Internet grows continuously and since it consists of many interdependent autonomous systems, there is no ground truth regarding how it looks – not even retrospectively. Nonetheless, we rely on a fundamental understanding of its state and dynamics to approach a solution. Since it is impractical to understand such complex systems at once – research on complex systems is older than the Internet itself – in this study we focus on a better understanding of the key players on the Internet by measuring Internet service providers (ISPs), Internet exchange points (IXPs), and systems running the Internet, such as routing, packet exchange, and the Domain Name System (DNS).

We describe our methodology of passively measuring large amounts of network traffic data at different vantage points, and discuss the challenges, solutions, and best practices that we experienced in the course of our work. Our measurements require an understanding of the vantage points in terms of their characteristics, the systems we measure, and the data we obtain. In the course of the work, we do not exclusively rely on passive data collection and its analysis. Instead, combining our active and passive measurements helps us to improve the understanding of the data in the domain of Internet network operation.

Our primary findings regard the role of IXPs in the current Internet ecosystem. We find that IXPs are understudied compared to their importance as hubs for exchanging Internet traffic, some of them handling traffic volumes comparable to major ISPs. We identify and describe different models of IXPs' operation specific to marketplaces, namely Europe and North America. We make use of different kinds of publicly available data and proprietary data collection of Internet traffic to which we have been granted access. Our measurement results show that the Internet peering complexity is higher than anticipated in previous publications, and that IXPs are the key to this unexpected complexity. This highlights the importance of IXPs and the role they play in today's Internet ecosystem.

To further improve our understanding of global players' operation in the Internet, we use a DNS protocol extension (EDNS0) to reveal the mapping of users to servers for one of the early adopters of this extension. The elegance of this particular measurement is in its ability to run a global crawl from a single vantage point without the need to access proprietary data or a significant amount of infrastructure.

We find it useful to examine both dominant and emerging Internet components to gain a better understanding of how the Internet changes and how it is used. It is critical to measure the Internet's driving forces, but this is a difficult task and comes with technical and legal restrictions. In order to make the best use of the data we have, it is possible and practical to combine measurement methods. As the Internet evolves constantly and rapidly, the quest to understand becomes more

challenging by the hour. However, even without access to private data it is possible to find exciting details regarding how this large system is operated.

Zusammenfassung

Um den Zustand und die Entwicklung des Internets einzuschätzen, bedarf es Messungen, die erhebliche Herausforderungen bezüglich Qualität und Quantität darstellen. Die Hauptproblematik liegt in den Komponenten, die das Internet erst erfolgreich machten: Komplexität, Verteilung, Wachstum und ständige Weiterentwicklung. Aufgrund seiner Partitionierung in unabhängige kooperierende Autonome Systeme, ist es nicht möglich den Ist-Zustand dieses verteilten Systems zu irgendeinem Zeitpunkt festzustellen oder im Nachhinein zu rekonstruieren. Für eine sinnvolle Weiterentwicklung des Internets ist es jedoch unabdingbar, die treibenden Kräfte des Internets zu verstehen. Wir konzentrieren uns auf die großen Hauptakteure, Internet Service Provider (ISPs) und Internet Exchange Points (IXPs), sowie auf einige der Kerntechnologien des Internet, beispielsweise Routing, Paketvermittlung und DNS.

Wir legen unsere Methodik der passiven Datengewinnung an verschiedenen strategischen Messpunkten dar, arbeiten die wichtigsten Herausforderungen (und Lösungen) heraus und beschreiben unsere Verfahrensweisen. Die Arbeit erfordert ein exaktes Verständnis der Messpunkte, der gemessenen Systeme, der Messsysteme und der Daten. Im Verlauf der Arbeit beschränken wir uns nicht ausschließlich auf passive Messungen, sondern kombinieren passive und aktive Messungen. Damit gewinnen wir ein besseres Verständnis über Funktion, Prozesse und Betrieb des Internets.

Die Hauptergebnisse liegen in den Erkenntnissen über Internet Exchange Points, die, gemessen an ihrer Bedeutung und Funktion im Internet, bislang nicht ausreichend erforscht wurden. Einige der großen IXPs bewegen täglich Datenvolumen, die denen der großen Internet Service Provider entsprechen. Wir stellen die verschiedenen Betriebsarten von IXPs heraus, die sich geografisch stark unterscheiden – namentlich in Europa und Nordamerika. Für die Erkenntnisse nutzen wir einerseits öffentlich zugängliche Daten und andererseits eigene Messungen, die uns im Rahmen von Forschungskooperationen möglich gemacht wurden und die selbst nicht-öffentlich sind. Die Ergebnisse zeigen, dass die Peering-Komplexität signifikant höher ist als bislang angenommen und dass IXPs dafür verantwortlich sind. Ebenso bedeutsam ist der Einblick in die Art und Weise wie Route-Server funktionieren und zu einer besseren Skalierbarkeit mit vielen Peering-Parteien beitragen. Diese Skalierbarkeit kommt jedoch auf Kosten von Sichtbarkeit.

Um ein besseres Verständnis einiger Hauptakteure im Internet zu gewinnen, benutzen wir eine Erweiterung des DNS-Protokolls, EDNS0. Wir identifizieren die Zuordnung von Nutzer-IPs zu Servern für einen bestimmten Inhalteanbieter (Content Provider) auf globaler Ebene. Die Eleganz des Verfahrens liegt in der Zugänglichkeit der Information, deren Gewinnung keine proprietären Daten oder eine dedizierte Infrastruktur zur Messung voraussetzt.

Publications

Pre-published Papers

Parts of this thesis are based on the following peer-reviewed papers that have been published or have been accepted for publication already. The thesis includes the author's versions of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in ACM conferences.

International Conferences

N. Chatzis, G. Smaragdakis, J. Böttger, T. Krenc, A. Feldmann,
On the Benefits of Using a Large IXP as an Internet Vantage Point,
Internet Measurement Conference 2013, Barcelona [65].
`http://doi.org/10.1145/2504730.2504746`

F. Streibelt, J. Böttger, G. Smaragdakis, N. Chatzis, A. Feldmann,
Exploring EDNS-Client-Subnet Adopters in your Free Time,
Internet Measurement Conference 2013, Barcelona [146].
`http://doi.org/10.1145/2504730.2504767`

P. Richter, G.Smaragdakis, N.Chatzis, J.Böttger, A. Feldmann, W. Willinger,
Peering at Peerings: On the Role of IXP Route Servers,
Internet Measurement Conference 2014, Vancouver [136].
`http://doi.org/10.1145/2663716.2663757`

Contents

1 Introduction 15
- 1.1 Challenges . . . 15
- 1.2 Internet Players and Network Topology . . . 17
- 1.3 Contributions . . . 20
- 1.4 Structure of this Thesis . . . 22

2 Background 25
- 2.1 Active and Passive Measurements . . . 25
- 2.2 Legal Implications . . . 28
- 2.3 Data Sources . . . 28
- 2.4 Supplementary Information . . . 30
- 2.5 Data Formats for Representation of Network Traffic . . . 31
- 2.6 EDNS-Client-Subnet DNS Extension . . . 32
  - 2.6.1 EDNS Protocol Specification . . . 33
  - 2.6.2 Challenges in Enabling ECS . . . 33

3 Vantage Points in the Internet 35
- 3.1 Overview and Facets of Vantage Points . . . 36
  - 3.1.1 Internet Traffic and Communication Models . . . 37
  - 3.1.2 Measurement Data . . . 44
- 3.2 Internet Exchange Points . . . 46
  - 3.2.1 IXP 101 . . . 46
  - 3.2.2 Under the Surface . . . 52
  - 3.2.3 Surprised by the Obvious . . . 54
  - 3.2.4 Non-profit Matches Profit . . . 55
  - 3.2.5 Challenges and Opportunities . . . 56
- 3.3 Internet Service Providers . . . 58
  - 3.3.1 Description . . . 58
  - 3.3.2 Typical Services . . . 59
  - 3.3.3 Information Availability . . . 61
  - 3.3.4 Layout of an ISP Eye-ball Network . . . 62
  - 3.3.5 Data . . . 62
  - 3.3.6 Measurement Opportunities . . . 63
  - 3.3.7 Challenges and Limitations . . . 64
  - 3.3.8 Scope of the ISP Vantage Points . . . 65

4 Monitoring Internet Vantage Points 67
- 4.1 General Aspects of Building and Operating a Network Vantage Point . . . 67
  - 4.1.1 Cooperation and Data . . . 67
  - 4.1.2 Ground Truth . . . 70

4.1.3 Access . . . 70
4.1.4 Technical Parameters . . . 73
4.1.5 Documentation . . . 81
4.1.6 Monitoring and Alerting . . . 86
4.1.7 Synchronization . . . 89
4.1.8 Planning and Future Scalability . . . 90
4.1.9 Measurement Operation . . . 93
4.1.10 Maintenance Operation . . . 98

5 Measuring the network of a large European ISP **103**
5.1 Cooperation and Data . . . 103
5.2 Ground Truth . . . 104
5.3 Access . . . 104
5.4 Documentation . . . 105
5.5 Monitoring and Alerting . . . 105
5.6 Synchronization . . . 106
5.7 Planning and Future Scalability . . . 107
5.8 Measurement Operation . . . 108
5.9 Maintenance Operation . . . 109

6 Measuring the Network of a Large European IXP **111**
6.1 Cooperation and Data . . . 111
6.2 Ground Truth . . . 111
6.3 Access . . . 112
6.4 Technical Parameters . . . 113
6.5 Monitoring and Alerting . . . 115
6.6 Synchronization . . . 115
6.7 Planning and Future Scalability . . . 116
6.8 Measurement Operation . . . 117
6.9 Maintenance Operation . . . 118

7 Measuring from an Active Vantage Point at a University **119**
7.1 Cooperation and Data . . . 119
7.2 Access . . . 121
7.3 Planning and Future Scalability . . . 123
7.4 Summary . . . 123

8 The Pulse of Today's Internet at an IXP **125**
8.1 IXP as Rich Data Source . . . 125
8.1.1 Available IXP-internal Datasets . . . 125
8.1.2 Methods for Dissecting the IXP's Traffic . . . 126
8.1.3 Available IXP-external Datasets . . . 128
8.1.4 IP Server Meta-data . . . 129

8.2 Local yet Global 130
8.2.1 On the Global Role of the IXP 130
8.2.2 On the IXP’s Dual Role 132
8.2.3 On the IXP’s “Blind Spots” 134
8.3 Stable yet Changing 136
8.3.1 Stability in the Face of Constant Growth 136
8.3.2 Changes in Face of Significant Stability 140
8.4 Beyond the AS-level View 141
8.4.1 Alternative Grouping of Server-IPs 141
8.4.2 New Reality: ASs are Heterogeneous 142
8.4.3 New Reality: Links are Heterogeneous 145

9 Peering at Peerings 149
9.1 Trends in IXP Operations 150
9.1.1 Background and Motivation 150
9.1.2 The IXP Route Server as Enabler 151
9.1.3 On the Deployment History of Route Servers 153
9.1.4 BIRD: A Popular Router Daemon for RSs 153
9.1.5 Looking Glasses and Route Servers 154
9.2 IXP Specific Data Sources 155
9.2.1 Profiles of three IXPs 155
9.2.2 IXP-internal Data: Route Server 156
9.2.3 IXP-internal Data: Traffic 156
9.2.4 Public IXP-specific Data 157
9.3 Connectivity: Bi-/Multi-lateral 157
9.3.1 Connectivity: IXP-provided Data 157
9.3.2 Connectivity: Public BGP Data 159

10 Exploring EDNS-Client-Subnet Adopters 161
10.1 Introduction 161
10.2 Data Sets 162
10.2.1 Network Prefixes 162
10.2.2 Content Provider Datasets 163
10.3 Methodology 164
10.4 Evaluation 164
10.4.1 Uncovering Infrastructure Footprints 165
10.4.2 Uncovering DNS Cacheability 166
10.4.3 Uncovering User-to-Server Mapping 170
10.4.4 Summary 171

11 Conclusion 173
11.1 Outlook 175

Acknowledgements 177

List of Figures **179**

List of Tables **181**

Bibliography **183**

1 Introduction

The Internet has become an essential part of our everyday lives. Large parts of our society already depend on its reliability and security. This involves not only technical interaction, e. g., using end-user services on the Internet, but also direct inter-personal communication, be it text based, speech, or video. Other and less obvious dependencies are business relationships via virtual market places, e. g., customer-to-customer (C2C), business-to-customer (B2C), and business-to-business (B2B). It also includes hidden dependencies, such as the financial markets and politics.

1.1 Challenges

The Internet is not the only technical system on which we rely, but it is a particularly complex one. Due to this complexity, the Internet is not yet well understood. Network technology, technical deployments, economy, and other aspects continuously change at rapid speed regarding multiple aspects, e. g., connectivity, business relations, services, and security. Moreover, an increasing number of devices are becoming connected to this global network every single day.

Research on the Internet is continuously trying to keep up with the newest developments. Among them are economics, players, vantage points, and network traffic, which are the focus of this thesis. They influence and depend on each other, and it is thus impossible to understand them in isolation. Our goal is therefore to understand some of their interactions, for example by identifying relevant and/or new players required to understand economics. This also requires an understanding of network traffic and vantage points, because the big players do not necessarily dominate all network traffic at all vantage points. In fact, some of these big players (e. g., Google and Akamai) localize their network traffic; consequently, the more successful they are, the smaller is the amount of traffic we might expect to see at some Internet vantage points. Therefore, we need to understand their business in order to understand their network traffic. In turn, the network traffic can help us to understand their vantage points and their business.

However, because the Internet consists of many subsystems that are controlled in a distributed way, it is possible that our biggest challenge is that we have no obvious and/or central points from which to measure all aspects of the Internet. Moreover, when we find vantage points, access to them may be restricted, and the data may be confidential. In addition, there are too many instances of such vantage points, e. g., various Internet service providers (ISPs), Internet exchange points (IXPs), and data centers (DCs), and it is physically impossible to measure them all.

Identifying Internet players is feasible if one has access to data that these players exchange. This requires the mapping of network traffic to players, which is circular depending on the interesting players that are identified. Still, we can start from a few well-known players, examine their traffic at different vantage points, understand this traffic, and infer their business. With this information we can identify patterns and, as a result, add additional players to the list. Our starting point is the assumption that the Internet is a hierarchy of connected ISPs. With time, new services and business models are introduced and some of these evolve into full-grown major players equal or superior to some of the major ISPs. For example, content delivery networks (CDNs), search engines, mail providers, content providers, social media companies, and IXPs have joined the Internet ecosystem. The list continues to grow and in addition to new players entering the market, the importance and influence of such players can change.

Selecting a representative set from even the major players on the Internet is difficult. The most obvious metric is the transfer volume of each of the players. Besides the traffic volume, the question is whether this traffic also represents the most influential traffic and what "influence" refers to. For example, video traffic is known to be one of the top contributors of traffic volume. However, it is difficult to determine its contribution compared to that of social media or news portals, which contribute significantly less traffic. Once we have identified the "most influential players" with any metric, how do we measure their traffic? The models, how traffic is exchanged, and how content is delivered do not only change over time, but also differ with regard to participating players. For example, a centralized way of delivering content has lost importance because of scalability issues. It has been replaced by distributed content delivery, including traffic engineering, and is supported by CDNs. The delivery model for content continues to change towards cooperation of major players, which begin to employ technical systems for traffic engineering for mutual benefit. For example, ISPs and CDNs cooperate for the purpose of service quality and traffic engineering. Another challenge is the access to traffic exchange models and to the underlying business models. This information is typically confidential because it is the basis for the economic success of these players. In particular, the players that are among the "most advanced" (e. g., Google, Facebook, and Akamai) are difficult to study because they use proprietary infrastructure. Not only the big companies shape the Internet, however: although they each contribute a negligible amount of traffic, in sum the end-users' demand and traffic are the driving forces of the Internet business.

Selecting the data from which we infer the role and the importance of each Internet player is a complex challenge. We differentiate data by means of a control plane and a data plane. The classification of traffic into one or the other plane is not straightforward. Consider, for example, the Domain Name System (DNS) network traffic. In the context of DNS-based content delivery and traffic engineering it belongs to the control plane. On the other hand, however, it belongs to the data plane when considering traffic volumes, traffic types, and services. Another example is VPN traffic and other kinds of tunnels. Other traffic is easy to classify: BGP and other routing traffic are always classified as belonging to the control plane, while video or social media content clearly belongs to the data plane. While identifying control traffic (i. e., the border gateway protocol (BGP) and other routing protocols) is easy, identifying content type of data plane traffic is a research field of its own.

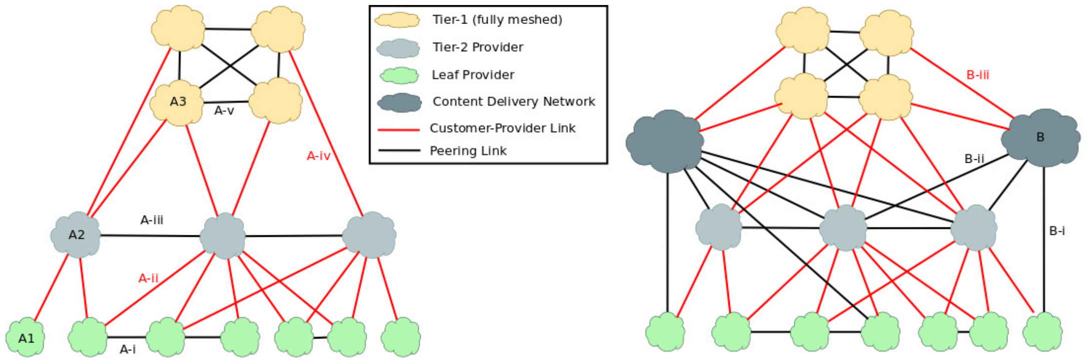


Figure 1.1: Internet Topology Evolution: Transition from a pure hierarchical model of connectivity to a free peering with major content providers (CDN).

In this thesis, we report on work that has been conducted to understand some of the above challenges.

1.2 Internet Players and Network Topology

End-users

End-users are a diverse group of players, and we refer to them as those individuals who use the Internet. There are many reasons for this diversity. First, individuals have different interests and use different kinds of technology. Thus, even among the same type of connectivity, e. g., among residential networks or among campus networks, the traffic mix of two entities can differ significantly. Second, each individual can fill different a role when using the different e. g., networks. Typically, the same person uses the Internet differently on his private residential network compared to on his company network, or on mobile networks. Moreover, the end-user is the entity that consumes and drives the demand for content. Related work has evaluated user behavior based on user traffic inside ISPs [113] and campus networks [107]. Work has also been conducted to evaluate connectivity parameters, e. g., available speed, settings, and quality of service (QoS) within residential networks, such as Netalyzr [102], Bismark [147] and SamKnows [56]. Other work [143], [99] focuses on application traffic analysis.

Internet Service Providers (ISP)

Networks of end-users are typically connected to an ISP, which relays the users' traffic to the rest of the Internet. To differentiate between end-user networks, e. g., company or campus networks, and ISPs, in this study we use the ownership of an autonomous system number (ASN), which is required to operate a network as ISP.

Not all ISPs are the same size, and there are a variety of business cases. With regard to routing, one major distinction between ISPs is their class, i. e., the *tier level* in which they operate. Tier 1

ISPs are those top few players that do not pay anybody for transit, that is to route traffic on their own and their customers' behalf. By definition, those ISPs interconnect (peer) with each other without payment and are paid by their customers, which are mostly tier-2 ISPs. Tier 2 ISPs pay tier 1 ISPs for transit and are paid by their customers, who are on a lower tier level. This is (not strictly) a hierarchical model, but since an ISP's tier level is determined by business agreements, it is not always possible for us to discover the true relationships. An ISP is classified as a leaf ISP if it provides neither paid nor unpaid transit to other ISPs.

Figure 1.1 (left) illustrates our starting point of looking at the Internet topology. It is the AS-level view of interconnected networks that uses different tiers and peering relations, e. g., peer-peer and customer-provider. ISPs typically operate complex networks, but as ASs they appear as a single entity in the global routing system. ISPs can also offer different classes of service, including connectivity to residential and mobile networks, and network applications such as IPTV. Moreover, they can run data centers, for example for housing and co-location, and they can offer content. Although there are specialized ISPs for dedicated services, e. g., hosting (Hetzner, AS24940), ISPs typically offer a set of services on the market. Based on the number of actively used Autonomous System Numbers (ASNs), the number of operating ISPs worldwide, is currently estimated to be $47K$.

Content Delivery Networks (CDN)

A class of Internet service has emerged from the need for global, efficient, reliable, and scalable distribution of websites and other content. Although this business often includes the operation of an AS, there are two reasons to consider those players from a different angle. First, the nature of CDNs is to operate a network and other infrastructure across ASs and to change their configuration with demand and available resources. The first reason is that CDNs do not necessarily own an ASN. These are called "Meta-CDN" and are typically distributed across different ISPs, e. g., CDN77[1] or MetaPeer[2]. Unlike ISPs, ASs of CDNs typically do not offer transit for sale. The second reason is that CDNs can operate multiple ASs, but are observed to use Internet Protocol (IP) address space that is assigned to foreign ASs, e. g., Google-Caches [64]. These properties break the pure hierarchical topology model of the Internet, since those ASs are willing to peer more than traditional ISP ASs. Figure 1.1 (right) shows how CDNs' ASs peer with ASs of all tier levels

Data Centers

Data centers are another class of players. This stems from the observation that especially cloud computing services and cloud storage services rely on them. Data centers contribute a significant amount of traffic to the Internet traffic volume. In particular, the communication between data centers [112] is recognizable in the network traffic. The operation of data centers creates a significant amount of machine-to-machine communication, e. g., for load-balancing and redundancy.

[1] CDN77: http://www.cdn77.com
[2] MetaPeer: http://www.metapeer.com

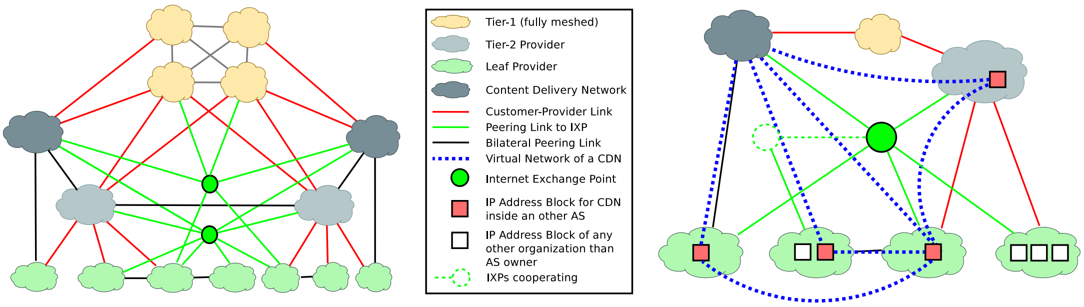


Figure 1.2: Internet Topology Evolution: Introducing IXPs into the picture, leading to more alternatives in connectivity compared to the former hierarchical models in Figure 1.1

Data center operation itself has been investigated by [54], [134], [156], and [48]. However, the impact of inter-data center communication is not yet well understood

Internet Exchange Points (IXP)

IXPs are the "elephants" in the Internet inter-domain traffic exchange business. At the time of this writing, there are at least 211 IXPs operating worldwide in 149 different areas[3]. Still, the number of IXPs is small compared to the reported number of assigned ASNs at the same time, which is greater than $45K$. To date, this particular class of Internet players has been less of a focus of research than others, e. g., CDNs and ISPs. One reason is the assumed lack of availability of public information on IXPs. Some of the first information in this field comes from our cooperation with a major European IXP, which we established in 2009. As a first step, the focus of IXP research was set by Ager et al. [43] with exploring the anatomy of such an IXP. The results point to a highly central and global role of IXPs in interconnecting ASs on the Internet. Figure 1.2 (left) shows the interconnecting role of IXPs in the inter-AS context. It illustrates the increased opportunity for interconnection among lower-tier ASs and their waned dependencies on big tier-1 ASs.

Address Space

A particularly interesting facet of Internet address space, e. g., ASs and IPv4/v6, is the usage of IP blocks (prefixes) within and among organizations and ASs. An AS is recognized as a single organization that is assigned an AS number and that owns a fraction of the public routable IP address space. IP address space is organized in blocks for managing routing information. Currently, the perception of the Internet is, on a high level, that it is a decentralized and distributed "network of ASs", and Internet scale traffic models typically refer to an AS-level view of the Internet. This view is about to change on the organizational and the technical level. For example, from the research on CDNs [65], [82], it is understood that organizations that may or may not own an AS and IP address space make use of prefixes in other ASs. Ownership transfer of AS numbers and

[3] According to http://www.datacentermap.com/ixps.html by July 2014

IP prefixes is another reason for deviations from the assumed 1:1 mapping of ASs and organizations. For example, according to the Cymru whois service [34], Akamai had been assigned at least 22 AS numbers by October 2013, and Google at least 12. These Internet players typically make creative use of their assigned ASs, and of their assigned and leased prefixes, e. g., by deploying overlay networks for traffic engineering and/or security [108]. We illustrate this with an example in Figure 1.2 (right).
Since CDNs try to place their infrastructure as close as possible to large concentrations of "eyeballs", they try to deploy them directly within eyeball ISPs. Moreover, for traffic engineering purposes, CDNs try to deploy equipment and services into strategic places on the Internet (Figure 1.2 right). Running a network service requires the assignment of publicly routable IP addresses. If the CDN needs to keep control over routing for these addresses, it requires the control over publicly routable IP address blocks. As a result, the CDN uses IP address blocks that it does not own. By examining the traffic, domain name, and whois information in this study, we were able to identify such networks. The analysis shows that CDNs run virtual networks across those prefixes, which enables them to perform traffic engineering beyond the limitations of BGP.
Another observation is the remote peering practice at IXPs. Members who connect to an IXP that enables such a service can peer with members of a facility of the same IXP or a cooperating IXP. As a consequence, this enables particularly smaller players to increase the density of peering throughout the Internet even further.

1.3 Contributions

Operating and Monitoring Vantage Points

Our contributions stem from operating multiple different vantage points throughout the Internet and relate to ISPs, IXPs and campus networks. We document guidelines for planning, deploying, maintaining, and operating our measurement deployments. Our focus is on methodology, scalability, and usability. Moreover, we provide recommendations based on our best practices for Internet measurements.

Characterization of Vantage Points

ISP: ISPs are well understood with regard to their operation and business since they have been studied in the past. However, operating vantage points is not trivial nor well documented. Therefore, for the ISP that we use as a vantage point in this study, we provide information on routing and on network traffic that is observable at different levels and portions in its network, and compare it to the IXP later on.

IXP: To date, IXPs have been understudied and, as we show, underestimated with regard to their impact on the Internet ecosystem. We collected and analyzed data over a time span of more than two years. By mid 2014, the IXP connected 600^{+} participants on the peering platform with $1,100^{+}$ ports in service of a capacity of up to 100 Gb/s per port. Table 1.1 presents some additional metrics of that IXP.

	Oct. 2011	Oct. 2013	growth in %	avg growthper month in %
Members	400	580	45%	1.55%
Traffic/day	$10.3PB$	$19.1PB$	85%	2.6%
Samples/day	$2.9 * 10^8$	$8.9 * 10^8$	205%	5%

Table 1.1: The growth of one of the observed IXP within the last 2 years demonstrates its increasing importance in the Internet.

With this data, we characterize the business and the operation of IXPs by investigating public information and internal IXP data. We analyze quantities of traffic, e. g., traffic volumes across links and address spaces, and of the whole platform. Moreover, we analyze connectivity matrices and traffic matrices, and conduct a traffic classification.

Active Measurement: For the active measurement observation point, we provide a similar overview to that of the ISP. It is important to understand the opportunities that each vantage point offers and to overcome some of their limitations by combining measurements across them.

Evaluation of Publicly Available Information

We analyze different types of ISP and IXP network measurement data, conduct active measurements, and explore publicly available information. Moreover, we compare our measured data from different vantage points and of different types to the publicly available information. This enables us to identify opportunities and limitations of the vantage points, particularly with regard to routing information and DNS, and we demonstrate the combination of public and private information for benefit. This provides a better understanding of how far we may trust the public data and how to judge research results based on those data. In particular, we verify some ideas about AS connectivity, prefix usage, and traffic flows. As a result of our work, we improve some of these ideas. Moreover, we point out some business models, identify big players, and connect traffic patterns with business models.

Data Formats and Sampling

Measuring at an ISP and at an IXP requires different methods. While at the ISP we can afford to take unsampled full packet traces of up to almost 10Gb, the network traffic measurement data at the IXP comes in SFLOW format and is thus already sampled by rate and by size. In order to compare what we can observe at an ISP and an IXP, we apply the IXP sampling to the ISP data and analyze the extent to which we can trust the sampled data from the IXP. Our findings show that the sampling rate does not impact the representativeness of the IXP data regarding the analysis in which we are interested.

AS versus Player

We overcome the pure AS-level model for the modeling of Internet traffic. We identify Internet address space that is used in contradiction to the still-dominating AS-level model by measuring at an IXP. In addition, we investigate a /24 prefix granularity of that traffic and classify the application of the exchanged traffic with an accuracy of 95%. This allows for a classification of prefixes into server-only, client-only, and mixed. By examining the Internet players, network traffic, and the involved network elements more closely, we identify the purpose of some of the affected traffic, which allows us to link it to business models.

Public Routing Information

By analyzing and comparing publicly available routing information with the routing information at a major European IXP, we can show the limitations of both the public data and the IXP data. For the publicly available information, we use and compare various sources, e. g., RIPE RIS, RouteViews, and Packet Clearing House. For the IXP, on the other hand, we rely on two sources. On the control plane, we analyze information from the IXP's route servers. On the data plane, we find additional AS links and we assign traffic volumes to AS links. We use this traffic matrix information to estimate the significance of differences in the public and private data. Moreover, we identify unknown and confirm available routing policies that some of the players use.

Active and Passive EDNS Measurements

As an example of an active measurement that can be enriched with passive measurement data, we examine the consequences that EDNS deployment have.
The recently proposed DNS extension, EDNS-Client-Subnet (ECS), has been quickly adopted by major Internet companies, such as Google, to better assign user requests to their servers and improve end-user experience. We show that the adoption of ECS also offers unique but likely unintended opportunities to uncover details about these companies' operational practices at almost no cost. A key observation is that ECS allows everyone to resolve domain names of ECS adopters on the behalf of any arbitrary IP/prefix on the Internet. We are able to (i) uncover the global footprint of ECS adopters with little effort, (ii) infer the DNS response cache-ability and end-user clustering of ECS adopters for an arbitrary network on the Internet, and (iii) reveal the mapping of users to server locations as practiced by major ECS adopters.

1.4 Structure of this Thesis

The thesis is structured as follows. First we provide background information on some of the fundamentals on which we depend in Chapter 2. In Chapter 3, we characterize the role and operation of three location types on the Internet where we collect and analyze data. In particular, we emphasize the possibilities for active and passive network measurement opportunities. We show vantage point limitations and elaborate on possibilities to cope with and/or overcome them. Moreover, we describe the roles and operations of the vantage points that we access and use throughout the thesis, which are ISP networks, IXP facilities, and active measurement points. We highlight that IXPs are important, major, but to date understudied players in the Internet ecosystem.

In Chapters 4–7, we describe the setup and the measurement methodology that we used at relevant vantage points for our work in technical, operational, and organizational means, because running those kinds of vantage points is neither a common operation nor well known among researchers. In Chapters 8, 9, and 10, we present examples of the data, analysis, and results of our research. The thesis concludes with a discussion and suggestions for future work in Chapter 11.

2
Background

.

2.1 Active and Passive Measurements

Active Measurements

With active network measurements, we refer to methodologies in which traffic is induced or modified for the purpose of measurement. These methodologies have the advantage of designing traffic for a measurement, such as defining the source, destination, type of traffic, payload, time, and network location. This allows for sophisticated probing of the network for its condition with a particularly crafted method. On the other hand, the measurement itself changes the state of the studied object, namely the network and the end-systems to measure. Active probing causes an additional load not only on the links but on all involved systems. Especially capacity measurements in operational networks are problematic because they significantly change network states by traffic engineering and other mechanisms, and as such by definition bias the measurement results.

Passive Measurements

With passive network measurements, we consider all activities that entirely rely on observing network traffic without interfering with it or inducing it in the first place. Such measurements are usually taken by duplicating traffic via optical splitters. The advantage of correctly conducted passive measurements is the unbiased observation of the object of study. The large disadvantage is the reliance on traffic that is unpredictable and uncontrollable. For instance, the effect of congestion can only be measured if the network is naturally congested. Moreover, additional context information on observed network traffic is often not available.

Enriching Passive Measurements

Whenever possible, we rely on passive measurement for our study. However, in some cases we need to enrich measurement data with additional active measurements. The first and most important requirement is not to interfere with the subject and the medium that we measure. Another critical requirement is a close distance between the correlated measurements with regard to time. The quantification of "close" depends on the type of measurement and the dynamics that are observed. This property of synchronization imposes online data analysis, at least to a degree where the parameters for the supporting measurements can be extracted. Examples of active methods of enriching passive measurement data are DNS lookups of observed network traffic, and X.509 certificate fetching while passive measurement data are analyzed. Active probing based on passive measurement data must be the source of bias, i. e., in the case of DNS and due to DNS-based resource allocation.

Tools for Passive Network Traffic Measurements

Simple Traffic Recording and Low Level Analysis

The most basic and common tool for network data analysis is tcpdump. There are also a variety of tools available, such as tshark, wireshark, tcpflow, tcpstat, and ipsumdump. For flow processing, we rely on sflowtool, which we have modified to fit our needs. On the systems that run with specialized network recording equipment, we rely on the open vendor specific "DAG" capturing format and recording mechanisms. While some low-level tools already support the format, it must often be re-written into PCAP format. In this process, meta-information can be lost, such as time precision and link id. We use a set of "fishing-gear" tools to overcome those limitations.

Bro Network Intrusion Detection System

The Bro [131] Network Intrusion Detection System (NIDS) was initially designed and implemented as a versatile intrusion detection and prevention system. The core of this open source tool is a protocol detector and analyzer that is capable of reconstructing data flows on different network layers, including tunneling. The authors of this system describe the focus as follows:

> *While focusing on network security monitoring, Bro provides a comprehensive platform for more general network traffic analysis as well. Well grounded in more than 15 years of research, Bro has successfully bridged the traditional gap between academia and operations since its inception. As by today, particularly scientific organizations rely on Bro in terms of monitoring and securing their network infrastructure. Bro's user community includes major universities, research labs, supercomputing centers, and open-science communities.*

Indeed, we and our partners use Bro for both operational and research purposes. It has become a standard tool in the research community. The authors of Bro also point out some key features that contribute significantly to Bro's success:

- **Separation of concern:** Bro separates the protocol analyzers from the reporting. On the one hand, the analyzers are written in C++, BinPAC/BinPAC++ [128] for a high performance, and contribute to the event engine. The event engine itself is neutral to any application of

Bro. On the other hand, reporting is done by implementing policies in the Bro scripting language [129]. Bro acts upon events thrown by the event engine. The result can be no, one, or multiple actions. The actions can be log writing; event throwing, possibly resulting in another action; alerting; command execution; internal state change; sending mail; and more. For our purposes, we mostly use the log functionality on which we base the analysis.

- **Adaptation:** The concept of a scripting language makes Bro adaptable to tasks of highly different natures. While Bro is conceptually a network security monitor, we mostly use it as a protocol analyzer and for such different roles as timing analyzer, caching analyzer, content classifier, traffic mix analyzer, and link utilization monitor.
- **Efficiency:** Bro targets large deployments, i. e., company and university networks and uplinks that operate at high speeds. There is a collection of mechanisms that make Bro scale, such as clustering [152] and the "time machine" [117].
- **Flexibility:** Bro does not rely on a single detection approach. It can make use of port identification, content signatures, and protocol behavior and analysis.
- **Forensic capability:** Bro supports customizable logging of network activity and has recording capabilities. We mostly rely on the logging for all network layers and also for timing analysis.
- **In-depth analysis:** Bro can analyze traffic even if tunneled, provided the protocol analyzers are implemented and activated. It is capable of semantic analysis at the application layer, i. e., for characterization of AJAX [143], mobile device applications [115], and other Web technologies.
- **State:** Bro keeps extensive application layer state information about the monitored networks. This feature is most relevant in network security and we do not use it for our work.
- **Interfaces:** Bro can exchange information in real time via the Bro Client Communications Library (Broccoli) [101]. For our passive analysis, we do not use this feature.

Bro has matured and grown in both stability and features over time, and has been re-implemented several times. It has also grown with extensions, such as the time machine, cryptographic certificate support, and syslog2bro. It works for live network data as well as on stored traces, and it can handle compressed file formats.

Compared to other tools with which we have worked, Bro is easy to extend for protocol analyzers and policy definitions independently. Examples of additional Bro protocol analyzers are Bittorrent [142] and NNTP [99] analyzers.

Other Network Traffic Analyzers

We have also tested other available open source tools, but have found none matching the capacities and support of Bro. The following are two examples of such tools.
Snort [137]: Snort is signatue-based. However, a pattern detection mechanism is already part of Bro, as in former versions it came with a signature conversion tool (snort2bro) [1].
Tstat: The "TCP Statistic and Analysis Tool" (tstat) can analyze traces for the occurrence of several protocols and their effects. However, one of the main disadvantages compared to Bro is

[1] Bro online documentation: https://www.bro.org/sphinx-git/frameworks/signatures.html

the interlacing of mechanism and reporting. We use it when we need features that are not available in Bro, e. g., overall statistics of a trace. We find the tool difficult to extend, adapt, and modify.

Tools for Active Measurements

In the area of active measurement tools we try to rely on open source software. In the area of active measurement tools, we rely on open source software. The repertoire includes *ping*, *traceroute*, *dig*, *openssl*, *nmap*, and *hping*. All of these tools are well known to the community and thus we do not describe them here. In addition, we rely on *our own framework* that we have made publicly available for EDNS measurements [146].

2.2 Legal Implications

The legal conditions for our measurements are covered by two German laws: the “Telekommunikationsgesetz” (TKG) [62] and the “Bundesdatenschutzgesetz” (BDSG) [64]. The TKG is a federal law that regularizes the competition in telecommunication. It covers the subjects of registration obligations, eavesdropping and interception, market regulation, privacy protection, data retention for future reference, content blocking, right of way, aspects of consumer protection, and personal data provision (“Bestandsdatenauskunft”). On the other hand, the BDSG regulates the handling of data of a personal nature that are processed in information and communication systems. Other laws in this domain are the media federal state laws (“Landesmediengesetze”) and privacy protection federal state laws (“Landesdatenschutzgesetze”). Which of these applies depends on the nature of the vantage point. Whereas analyzing and monitoring traffic of ISPs or IXPs is clearly governed by the TKG, the BDSG applies for privately operated vantage points. The most important practical difference is that the BDSG allows for exceptions from some regulations for the purpose of teaching and research in §40[2]. Finally, independent of the applied law, it is strictly forbidden to draw conclusions from data from natural persons except for billing.

2.3 Data Sources

Flow Data

Flow data summarize network traffic on a connection level. Such a flow data measurement point consists of the connection tuple, source IP address, destination IP address, source port number, destination port number and protocol (transport layer in the OSI model), and attributes for flow description. Those attributes may include flow start, flow duration, number of IP packets, MPLS information, source and destination AS number, and more. Flow data do not include information beyond that obtained from packet header information and counting, and in particular, no payload is collected. One of the most popular flow collectors is NETFLOW (by Cisco).

[2] Verarbeitung und Nutzung personenbezogener Daten durch Forschungseinrichtungen, §40 BDSG

Flow Samples

In contrast to flow data, the term “flow sample” refers to a measurement practice in which real packet samples of limited size and number are collected. The format that we use for monitoring large amounts of volume over weeks and months continuously is SFLOW [144]. SFLOW is a mechanism that allows for the probing of a device (e. g., a switch or router), the collection of statistics on network traffic, and the reporting of it to a number of proxies and collectors. Proxies are used for reformatting and filtering the flows. The current SFLOW version is 5, and it has the status of a memo[3]. The former versions 2-4 are described in RFC3176 [130].
The sampling comes with a sampling rate and a sample representation. The rate we use is $1 : 2^{14}$. A network device collects multiple packet samples and reports them along with counter statistics in the form of datagrams. Flow sample information is not to be confused with flow information, which represents network traffic as the summary of connection-level “flow” analysis. The term flow describes network traffic that shares a tuple of properties, e. g., source and destination IP, transport protocol, and transport protocol source and destination (“port”). In contrast to flow information formats such as IPFIX [70] and NETFLOW [69], the flow sample format does not contain information about flow length and flow size. The advantage compared to flow data collection is the presence of a limited amount of payload in flow samples.

Packet Header Traces

Packet header traces contain per-packet information and as such provide more detailed information than flow information and flow sample information. One can think of packet header traces as flows samples with a sampling rate of 1. This also includes in-connection timing information. It allows activity across a flow or connection to be determined. One limitation of this format is the sampling in the dimension of data volume: header traces contain only a limited number of bytes from each frame or packet. The sampling restricts the sample size to a fixed maximum of octets, or it can use header analysis in order to determine the sample size. The advantage of the former method (flow samples) is that it requires less computation effort, while the disadvantage lies in the chance of storing too few or too many octets. Along with the packet headers, meta-information is usually stored, such as timestamp of collection, original packet size, and sample size.

Full Packet Traces

Full packet traces contain the complete information of each packet or frame, and as such represent the maximum information for passive network traffic measurements, including all header information. While the advantage is the completeness of information, the disadvantages lie in recording and storage requirements and in the application of confidentiality and privacy protection. Each of the other capturing formats (flow samples, flow information, and packet header traces) can be compiled from full packet traces.

[3] http://sflow.org/sflow_version_5.txt

Server Logs

Server logs contain application-specific information and thus overcome some limitations of network traces. Similarly to network packet traces, server logs are typically not a publicly shared resource and legal and technical efforts are usually required to gain access to them. Although server logs can be collected on one's own servers, the problem is that the more “interesting” logs are owned by big players, e. g., Facebook, Google, and Akamai.

Data Feeds

Network data are collected by many parties that do not share the resources of their vantage points directly but are willing to give out information to cooperators and researchers. We refer to information that is provided in a timely manner and either continuously or in intervals as a “feed”. Examples of feeds that we use are a routing table state feed, packet traces containing routing updates, a feed of the routing state of devices, a spam feed, and a daily snapshot of the .net and .com DNS top level domain (TLD) zones. Data feeds can be active, meaning that the source pushes information. This suggests the employment of unreliable data transfer so that the source is not affected if the destination does not cooperate reliably. The alternative is passive feeds, where the source of information is queried by the destination. In this case, reliable data transfer is common practice, such as sftp, tftp, and http/https.

Auxiliary Information

Auxiliary information is high-level information on the operation of objects or participants in the area of study. It involves information on business practices of parties in the network, ground truth information, and all information that contributes to an understanding of the network and traffic observed. Such material comes in the form of written documentation and databases, or even occurs informally by talking. Especially the latter is a great source and we encourage researchers to make use of this type of information, which is often neglected or underestimated.

2.4 Supplementary Information

Internet research cannot be restricted to a single type of information, such as traces or flows. In many cases, it has to be linked to additional information to put the primary information into context. The best example is the DNS. When considering traffic, we mainly see the connection end-points, namely IP addresses. In order to assign the traffic to an organization or another Internet player, we rely on the DNS as the service to deliver structured information about address usage. It is obvious that the domain name information cannot be extracted from the observed traffic most of the time. The reasons are manifold, i. e., caching, path selection to DNS servers, and traffic sampling. This limitation requires the active use of the DNS, and this in turn requires an understanding of how the DNS is used currently. Indeed, the DNS is used not only for plain name and address resolving, but also for traffic engineering, including by major CDNs such as Akamai and Google.

Another kind of supplementary information that we use is the X.509 certificate infrastructure. The advantage compared to the DNS is the reliability of X.509 information. In the DNS information is unreliable by default. Especially reverse lookups are misleading at times, when the information is either outdated or incomplete. As HTTP and HTTPS contribute the most traffic volume, and as HTTPS is evolving even faster than HTTP, it is logical to rely on information provided by HTTPS to assign traffic to organizations. HTTPS itself is encrypted, so we cannot even extract the requested uniform resource idendifier (URI) – either in the request or in the answer. What we can do, however, is check for the certificate delivered by an IP on the regular HTTPS (TCP) port 443. Besides cryptographic parameters, certificates contain information on sets of host names for which they are valid. In addition, legal information, such as organization name and country, is included in the certificate and improves the quality of server assignment to organizations.

2.5 Data Formats for Representation of Network Traffic

A few data formats are common for storing different types of network traces. The main differentiation is among packet capturing formats and flow collection formats.

Packet Capturing Formats

Packet capturing formats are used for a detailed snapshot of network traffic. Each passing frame is copied off the network and is stored alongside some meta-information on size, type, and timestamp. Traces of this type of data collection can even be used to replay network traffic. Due to the high volume of information, packet capturing is often used with filtering or trimming. Sampling is less common when collecting packet-level traces. The two most commonly used formats are PCAP and the Extensible Record Format (ERF).

PCAP: is the industry standard and supported by almost all software for network traffic capturing. Its limitation is a restricted accuracy of timing information, but it matches the accuracy of software-based measurements. The format itself is described in many publications, so it will not be described here. The main limitation of PCAP is a limited meta-information capability, which is restricted to 16-bit timestamp, sample size, and frame-type. In addition, PCAP requires a particular file header, which makes it inconvenient to operate different traces on a file basis.

ERF: on the other hand, is a data format produced mainly with Endace DAG network monitoring equipment. Its main advantage is its extensibility by arbitrary information and increased time accuracy of 32 bits. The time accuracy improvement is in the fact that timestamps are inserted by the hardware and can be synchronized among multiple capturing devices with nano-second precision. This makes it feasible to tap links synchronized per direction and among multiple links, as is necessary in channel bond (aka link aggregation) setups. In addition, there is no file header. Mandatory and additional meta-information is added to each sample in the trace. Thus, it is better suited to multi-link and high-speed measurements. On the other hand, ERF's main limitation is the increased storage requirement due to the per-sample and extensible meta-data capability. Descriptions of the format are publicly available online from vendors using the format, such as Emulex[4] and open knowledge bases[5].

Flow Collection Formats

In contrast to packet formats, flow formats are used to collect a higher-level view of network traffic. Flows are unidirectional and identified by a tuple that commonly consists of source IP, destination IP, source port, destination port, and protocol id. There are two commonly used formats (NETFLOW and SFLOW), and one upcoming one (IPFIX).

SFLOW: provides samples on an Ethernet frame basis. Instead of providing a complete set of high-level information for all flows, it provides two types of information: counter samples and flow samples. The collection process involves rated sampling and only considers individual frames – per-flow information is not included.

SFLOW counter samples are provided on a periodical basis. The network device submits the number of packets that have traversed a device (e. g., a switch or a port) per time unit, unsampled. **SFLOW flow samples** on the other hand, are collected by layer-2 frame sampling – usually at a high sampling rate that is adapted to the network speed and processing capabilities. In our SFLOW setup, the sampling rate is $1 : 2^1 4$ (16384). In contrast to NETFLOW formats, SFLOW does not summarize flow information. However, it does provide shortened frame samples and counter statistics.

NETFLOW and IPFIX: NETFLOW and its more powerful and IETF[6] standardized successor IP Flow Information Export (IPFIX) [61] are formats in which to store information on a per-flow basis. This implicates state processing of the selected flows for the collector, and imposes computational complexity on the devices. We restrict ourselves to a short description of IPFIX here, since it is a flexible and complex format and protocol. The type of information provided is configurable and thus flexible, but it usually does not change in a particular setup. IPFIX is capable of including data description in addition to the data themselves. Each message (see Figure 2.1) consists of the message header (8 octets) and sets. A set either describes a set definition (template) or contains an instance of a set. In addition, there is an IPFIX extension PSAMP [71], which uses the IPFIX protocol for exporting SFLOW-like flow samples.

2.6 EDNS-Client-Subnet DNS Extension

The EDNS-Client-Subnet (ECS) DNS extension [75] was introduced to tackle problems of mis-locating the end-system from which a DNS request originates. The main problem is that the

[4] Emulex website http://www.emulex.com
[5] Wireshark wiki: https://wiki.wireshark.org/ERF
[6] Internet Engineering Task Force

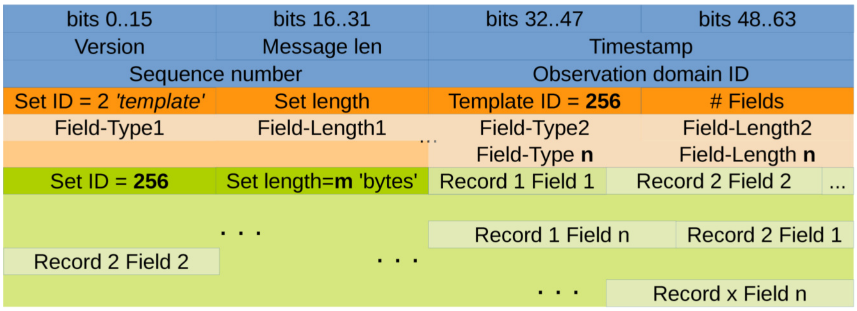


Figure 2.1: IPFIX message format: multiple data set types per message possible

end-system's IP information is typically hidden from the authoritative name server. With ECS, client IP information is forwarded by all ECS-enabled resolvers to the authoritative name server in the form of network prefixes.

2.6.1 EDNS Protocol Specification

ECS is an EDNS0 DNS extension [153] proposed by the IETF DNS Extensions Working Group. EDNS0, which is also needed for DNSSEC, uses an "ADDITIONAL" section in DNS messages to transfer optional data between name servers. Since all sections of a DNS query are present in the DNS response, bidirectional data transfer is enabled as the responder can modify this section. Name servers that do not support EDNS0 either strip the EDNS0 OPTRR in the ADDITIONAL section or forward it unmodified.

An example ECS-enabled query and response are shown in Figure 2.2. The ADDITIONAL section includes an OPTRR resource-record containing the ECS header and data. The ECS payload consists of the address family used, i. e., IPv4 or IPv6, prefix length, scope, and client prefix. To protect a client's privacy, ECS recommends the use of prefixes less specific than 32. In the request, the scope must be zero and is a placeholder for the answer.

The answer from an ECS-enabled DNS server differs in only one byte from the request, namely the scope that is needed for DNS caching. The answer can be cached and used for any query with a client prefix that is more specific than or equal to the prefix, as specified by the scope. We note that the response may contain a different scope than the query network mask, and we have indeed observed larger as well as shorter scopes than prefix length in our measurements. In the example, the query prefix length is 16, while the returned scope is 24. The scope is the essential element that allows us to infer operational practices of ECS adopters.

2.6.2 Challenges in Enabling ECS

While ECS is transparent to the end-user, it requires significant efforts by the DNS server operators, mainly because all involved DNS servers have to at least forward the ECS records. Among

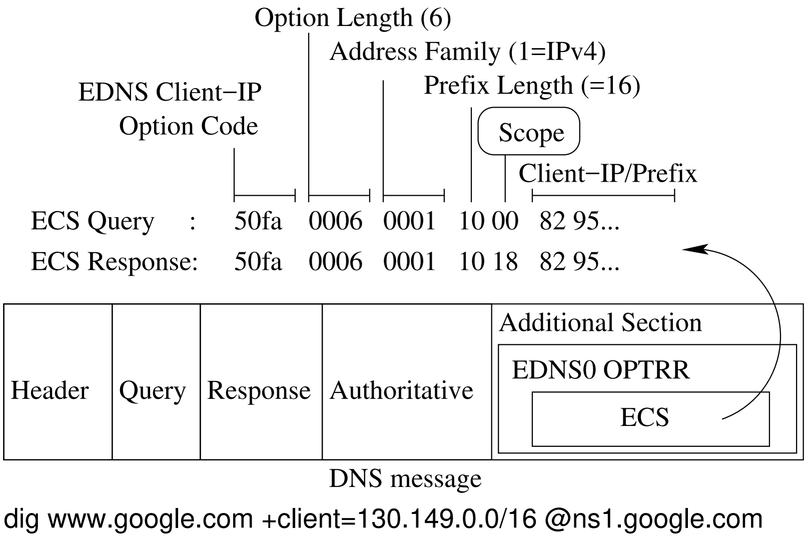

Option Length (6)
Address Family (1=IPv4)
Prefix Length (=16)
EDNS Client−IP Option Code
Scope
Client−IP/Prefix
ECS Query : 50fa 0006 0001 10 00 82 95...
ECS Response: 50fa 0006 0001 10 18 82 95...
Header
Query
Response
Authoritative
Additional Section
EDNS0 OPTRR
ECS
DNS message
dig www.google.com +client=130.149.0.0/16 @ns1.google.com

Figure 2.2: Example of ECS query and response.

the major obstacles are that (i) ECS-supported server software is not readily available[7]; (ii) all involved DNS servers need to be upgraded[8]; and (iii) third-party resolvers are not necessarily sending ECS queries by default. To change the latter, an engineer of Google Public DNS or OpenDNS, for instance, has to manually check the authoritative name servers and white-list them as ECS compliant.

Moreover, appropriate cache support has to be added to the DNS resolvers. Here, ECS with its notion of scope introduces another problem. For each query, the resolver has to check whether the client's IP address lies within the scope of any cached result. If not, the query has to be relayed with the appropriate prefix information. Consider the extreme scenario: a scope of "32". In this case, the resolver would keep a separate entry per client, making caching largely ineffective.

Overall, handling ECS is complicated, as the draft requires DNS forwarders to forward the ECS information sent by the client. A forwarder may modify the prefix mask to a less specific prefix. If no ECS information is in the DNS request, the forwarder may add an OPTRR record based on information from the socket. This is the rule followed by Google's public DNS servers, for example. Note that until Google enabled EDNS0 for DNSSEC support, it stripped the ECS records. Now, this information seems to be forwarded unmodified to the back-end servers.

[7] Currently, open source ECS patches are limited to clients, e. g., dig [2] and dnspython of OpenDNS [25].

[8] CDNs may internally use multiple resolution levels, e. g., Akamai [124]

3
Vantage Points in the Internet

Although "Internet" is a commonly used term in everyday life and research, experience working in the computer network researching area convinces us that we need to clarify the terminology that we use and to describe our focus.

Communication networks interlink to one another on different layers, based on decisions of their operators. On an abstract level, networks consist of nodes and edges connecting those nodes. Regarding the Internet, what we consider a node or an edge depends on the particular model. The clarification is important once we discuss vantage points. The term *vantage point* seems to refer to a single physical location, and is often considered a node or a link. Depending on the layers we consider and the model we choose, it can be one or the other.

We refer to the IXP as a node when considering the connectivity in the link layer. However, on the AS level in the network layer, for example, the service of an IXP is link provision.

However, "vantage point" can also refer to a particular network layer, a network protocol, an application, or other entities. In this thesis, we use "vantage point" first and foremost for a specific point on operators' networks.

The Internet provides great opportunity to study how distributed and inhomogeneous systems can work reliably, particularly under the condition of different, often competing operators and participants' interests, which influence network design and operation [72]. Since there is no single point of administration, control is distributed among operators and across different network layers. Thus, we do not have access to complete information on any of the Internet components. The best we can do to understand Internet operation is to use measurements. Similarly to administration, there is no possible single point that can claim true representativeness of the whole system or of one of the components. Moreover, in this highly inhomogeneous system, it is difficult to create a representative set of vantage points even for single aspects, e. g., routing, traffic characterization, and user behavior. In addition, a view on a single aspect of the system "Internet" can look different from different locations and different points in time, which applies for example to routing information but also to the DNS, as we discussed in Section 2.6.

Despite the difficulties due to decentralization, distribution, and diversity, for example, we can identify types of locations in the network that can serve as vantage points. We classify locations by the role and placement of network elements, by roles of the players, and by known business models, e. g., ISPs, IXPs, and CDNs. Although "measuring the complete Internet" is unfeasible, we can measure certain aspects of behavior at selected locations in order to improve our understanding of how the system works, and in which condition its parts are, and we can identify ongoing changes. By combining and comparing those pieces of information, we are able to infer changes in Internet businesses' models, operation, network traffic, and topology.

While we have to face the fact that we cannot provide complete information on the Internet's state and operation, we can collect different pieces of information and samples from many possible sources in order to obtain a valid overview of how the Internet is operated and used today. Although a reasonable number of independent researchers and companies are working on "understanding the Internet", there are still single events and upcoming opportunities that surprise us with new insights and lead to the extension or revision of our knowledge.

This chapter presents general observations about Internet vantage points. Subsequently, we describe the vantage points that we use throughout this thesis. We focus on results that are not necessarily common research knowledge, namely details on IXPs, ISPs, and active measurement vantage points.

3.1 Overview and Facets of Vantage Points

Important properties and challenges to take care of in the Internet regarding to measurements are: Important properties and challenges to consider in the Internet regarding measurements are the following:

- There is no central, multi-purpose and representative vantage point, although there are natural special purpose observation points with limited capability, i. e., DNS root-servers, border routers, and inter-AS links.
- There are a large variety of orthogonal and overlapping interacting mechanisms, i. e., routing, switching, virtualization, name resolving, and traffic engineering.
- The Internet design enables a large diversity regarding network layouts, technology and control, network devices, types, and roles
- We observe cyclic and acyclic longitudinal changes in network traffic and resource allocation, e. g., "follow the sun/moon" models in data center operation, or load-balancing.
- Distributed control is in fact a feature for reliability and scalability in key aspects, e. g., routing and resolving. At the same time, however, that makes it difficult to conduct and synchronize measurements regarding those aspects.
- Regional differences and longitudinal changes of legal situations restrict measurements.
- Business and non-commercial activities are permanently changing the Internet.

3.1.1 Internet Traffic and Communication Models

Models are tools for abstraction purposes, allowing us to restrict ourselves to a subset of aspects in complex systems by ignoring other aspects and details. By definition, models cannot reflect reality. Common abstractions in the communication networks domain are:

- Layer models, e. g., ISO/OSI or TCP/IP network stack
- Topology models, connectivity, and peerings
- Physical infrastructure, e. g., lines, switches, and routers
- Geo-location versus network location of infrastructure and services
- Service models, e. g., Web, email, and online social networks
- Content delivery, including information source, transport, destination, and delivery policy.

Each of these models can contribute to a better understanding of specific aspects of Internet traffic, and we need to refer to a particular model in order to explain where and what we measure. This is necessary to place results in the right context, and it also prevents inappropriate generalization. Therefore, we rely on some models for abstraction when it comes to measuring and interpreting our results. In the following, we describe the models we use in this thesis.

Layer Model

As the communication protocols in the Internet are designed and implemented in layers, we can use that approach to classify the measurement on a technical level. Information on the layers in our measurement determines what kind of information to expect. The TCP/IP model uses four layers.

Physical layer: Access to the physical layer requires a network tap on a link. Typically, no information is used on this level and data analysis starts on the link level. The information we care about regarding the physical layer concerns whether it is operating and in what condition the medium is, i. e., signal strength of optical links and media conditions in wireless environments. We have to deal with the fact that some monitoring technologies, e. g., optical and electrical wire tapping, interfere with the system that we measure as we insert measurement devices or enable measurement features on existing infrastructure elements. Thus, our measurement can add signal depletion, signal noise, signal delay, or even signal disruption.

Link layer: The link layer provides us with the lowest usable level of network traffic analysis. Commonly used technologies are Ethernet, PPTP, ATM, and FDDI, although in our measurements we focus mostly on Ethernet frames and some ATM. The link layer provides us with the most complete information of network traffic as it implies visibility of all upper layers, unless an upper layer uses encryption. When we refer to “full packet traces”, link layer information is included. On this layer we can identify the involved devices of the link or collision domain, and we are able to observe layer-2 specific techniques, e. g., link aggregation. Data formats for this level include PCAP and ERF (Section 2.5). Link-layer level information can be collected at media taps, where the methodology is referred to as “sniffing”. Typical methods are optical line tapping, Ethernet switch-port monitoring, and wireless signal collection. The original signals are not interfered with unless active measurement techniques are used, e. g., spoofing.

Network layer: In the network layer, we focus on control plane information. More precisely, what we seek is routing information on intra-domain and inter-domain scope. Routing information tells us that (at least a part of) network traffic is sent towards a prefix. In intra-domain routing, the information is restricted to AS-internal routing information and misses the inter-AS component. Thus, the level of detail focuses on the local portion of the network, e. g., the router addresses within the observed AS. In inter-domain routing, on the other hand, information is restricted to the high-level view and shows the complete AS path to travel to the destination. Inter-domain routing data provide the global AS view, but miss all AS-internal routing information. To operate a network, inter-AS and intra-AS routing information must be coordinated.
The type of data with which we work in the network layer are routing tables and routing messages. Routing tables can be extracted from routers, looking glasses, and route servers. The latter two offer the advantage of including multiple routers routing tables, but they are generally limited to "best route" information. Different data providers on the Internet collect and share inter-AS routing information, e. g., the RIPE route information service (RIS)[1]. Unfortunately, there is no service that includes all inter-domain routing information. Direct access to routing messages requires either access to the involved routers or access to the exchanged traffic, as routing messages are generally not encrypted. With direct access, all routing data of a router, including all links, are available. Conversely, indirect access (tapping) restricts us to the routing messages on the observed link.
Popular formats for storing routing tables are the Multi-Threaded Routing Toolkit (MRT) [60] and structured American Standard Code for Information Exchange (ASCII) representation. Routing messages are stored either in pcap format or in a routing protocol-specific ASCII representation.

Application layer: The application layer can contain both data plane information and control plane information. The DNS, for example, is considered to be control plane information because it is used to map address spaces on behalf of other applications, e. g., websites and email. Web traffic and email are typical examples of data plane information.
Application layer information can be gathered from different sources. Two common methods are log file analysis and network traffic capturing. In client-server applications, log files across all client-server interactions are available to the server operator. Although application information is generally also available on the client side, its collection is difficult and only includes the application information of that single client. Typical formats to store application layer data are structured ASCII or databases. In peer-to-peer scenarios, the information is available from both peers, but we face the same restrictions as in the case of client-side information collection.

AS Tier Centric Model

If we consider the Internet as a "network of networks", we can model it with nodes and edges. The highest technical level of administrative organization of which we are aware is the AS level,

[1] http://www.ripe.net/data-tools/stats/ris/

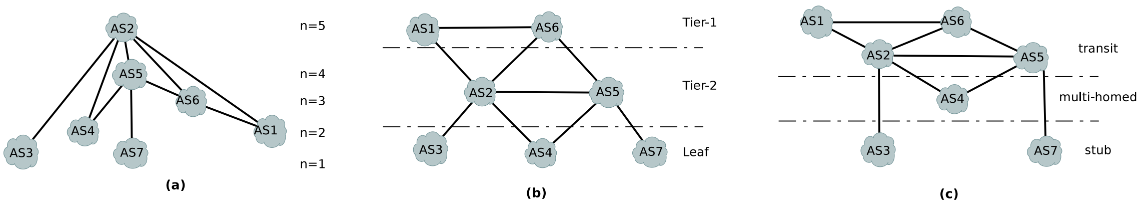


Figure 3.1: AS level views of a topology: different ways of AS classification in the same AS topology. While we can generate (a) and (c) from routing information, (b) requires payment information for traffic exchange.

ASs being the nodes and the inter-AS links constituting the edges. ASs can be attributed roles and network location. Typical hierarchical locations for networks are attributed a tier level, e. g., tier 1 (highest), tier 2, . . . , and leaf. The roles are distinguished into stub, multi-homed, and transit. The stub AS only has a single connection to another AS, and uses that as an upstream. A multi-homed AS has more than one upstream, but all traffic on the upstream links necessarily belongs to the address space of that AS. A transit AS is an AS that transports the traffic of other ASs on their behalf

Although this classification can provide a first impression of connectivity, it does not distinguish between all facets of different transit ASs. The classification simply does not include the magnitude of exchanged traffic or the fraction of transit traffic. After all, most transit ASs also need to originate and sink traffic. ASs of large national telecommunication companies with a significant end-customer basis are often also transit ASs, but big carriers are transit as well. Moreover, the tier level does not reliably indicate which traffic, transit or its own, dominates. While in some tier 1 the transit clearly dominates, e. g., AS 3356, “Level 3”, in others the transit traffic is not dominating, e. g., AS 3320, “Deutsche Telekom AG”. Another limitation is the classification of CDNs and IXPs in that scheme.

Figure 3.1 shows three different views of AS level connectivity: (a) degree of connectivity, (b) tier level; and (c) stub vs. transit. Although all three visualizations display the same topology, only (a) and (c) can be generated on a technical (measurable) level. To generate view (b), we rely on mostly non-public information about business relations. More specifically: from the traffic observable on the Internet and with regard to the AS level, we can identify AS business models and their relation to other ASs. This requires a mapping of AS to a prefix, which is contained in the inter-domain routing information.

Between tier 1s, the observable traffic and routing information is more diverse and should include sources and destinations from the entire public routable address space, because it contains a large fraction of transit traffic. At tier 2, on the other hand, the information is more localized to the provider, although transit traffic information is observable. Finally, information from leaf tier providers may be local because all traffic includes at least one IP that belongs to that provider.

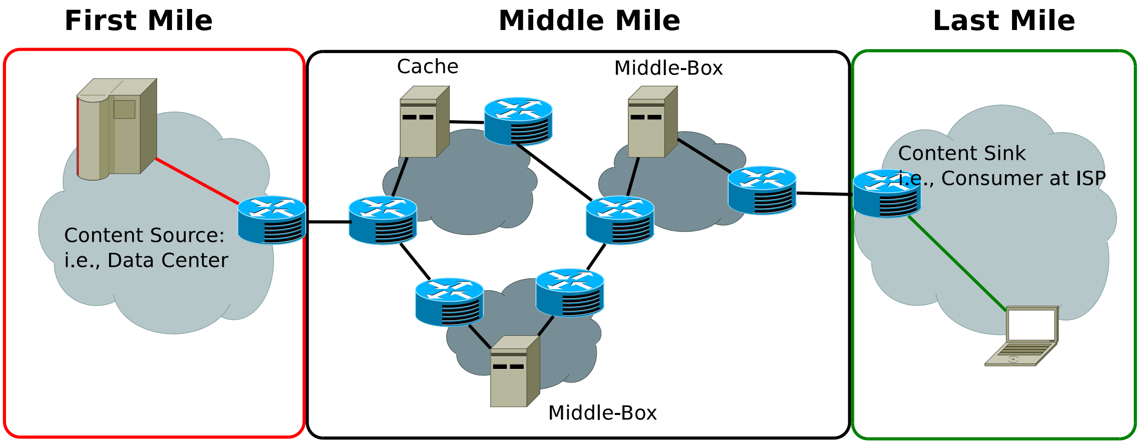


Figure 3.2: Definition of zones for content distribution via the Internet

Content Delivery Centric Model

First Mile

When content is generated centrally, e. g., by a commercial provider of video streaming, the first mile refers to the data center hosting the infrastructure needed to provide the content. Those data centers can be dedicated to a specific and highly customized entity, as we know from the case of Google data centers. However, content is also generated in data centers of cloud providers (e. g., Amazon Cloud, Microsoft Azure) or hosting providers (e. g., Equinix, OVH). The major properties of what we call the "first mile" are

- large hardware deployments such as in data centers
- engineered for scalability, reliability
- operated with high internal and external network bandwidth
- commercially used to serve content to users

Overall, we can state that the first mile is usually a highly sophisticated deployment and operation of infrastructure under a centralized management.

Middle Mile

The term middle mile refers to the large-scale deployments of network equipment at the core of the Internet. Although this infrastructure is trivializingly referred to as the "middle mile", it is the most complex. The commonly used techniques are inter-domain routing, including peering, transit, and traffic engineering. In contrast to the first and the last mile, the middle mile is decentrally controlled by multiple parties, mostly businesses, with competing objectives and different capabilities for operating the Internet's backbone:

- Competition: operators compete with each other, but have to cooperate at the same time.
- Decentralization: there is no centralized control of operation.

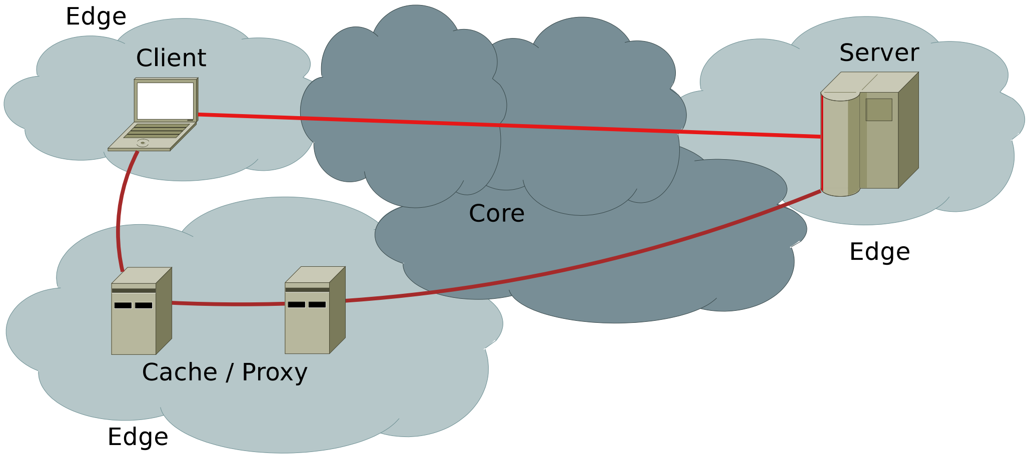


Figure 3.3: Definition of the edge and core by connection end points: While on the content level the cache is a part of the transit, on a connection level it is an end-point. All traffic that involves the cache terminates at the cache's IP address.

- No reliability: there is technically no guarantee of path availability.
- Change: There is permanent and fast development and engineering of components, yet stable overall behavior.

In the physical layer, the middle mile mainly relies on optical fiber transport, which provides the necessary capacity for Internet backbone traffic demands. The middle mile is usually connected to the first mile (data centers) with higher bandwidth links. It is connected to the last mile "eyeball" networks too.

Last Mile

Following the path of content delivery, we call the carrier network that is next to the end-user's network the "last mile". Last mile networks use access technologies, e. g., synchronous or asynchronous DSL, cable deployments, power line, and mobile technologies, e. g., UMTS or LTE. Usually, the end-user's ISP is the last mile network.

Without context, terminology can seem ambiguous. For example, see Figure 3.3, the terminology of "end-points" and "transit" is vague. In the connection model, the end-points are the clients, caches, and servers, but in the content centric model, clients and servers are the end-points while the caches are considered transit.

End-point Centric Network Types

Edge Networks

The definition of *edge network* is not commonly agreed on, even though it is used in various models in computer networking. It is disputable whether a stub AS is considered to be an edge network (and in the AS view of the Internet it certainly is), or whether only networks that connect end-point devices are edges.

In fact, the concept of selling Internet connectivity has changed over time. Initially, residential customers with single devices and a single IP behind the line paid less than companies that had a network with ISP address space connected. Today, on the other hand, it is common for residential users to operate their own network with multiple devices.

As such, we define the term “edge network” to be a network that connects a non-transit network to the ISP’s “last mile”, yet is not under control of the ISP. Examples are home networks, small business networks that are not operated by the ISP, and university networks. They vary hugely in size and complexity, from a single device up to networks with intra-domain routing and IP address space larger than /16, which in turn can provide services just as the data centers in the first mile do. This also illustrates that the roles are not as clearly separable.

Core Networks

Core networks are transit networks, e. g., ASs and AS backbones that are used for transit. A relevant distinction is that core networks must not change or modify traffic except to block or shape it for security and for traffic engineering purposes. Thus, a core network does not contain elements that actively modify content or connections.

Data and Control Model

Data plane: Data plane information can generally be collected anywhere in an operator network, and it can be represented in different formats. Access to the data plane typically requires access to the network traffic, e. g., links or end-points, or to network traces. Commonly used formats are packet traces, flow information, and log files of events.

Control plane: Control plane information is used to operate a network, e. g., for structuring and configuring. Thus, control information in itself represents the configuration of the network or a particular aspect of it, e. g., routing. This information is typically available at network elements that actively participate in the network, e. g., switches, routers, and caches. The format in which the control information is represented is typically structured in ASCII, XML, tables, and databases. Control plane information can be collected from configurations, e. g., the configuration file of a switch or router, as well as from collectors, e. g., the SNMP data of a network device or routing information in looking glasses. The two common representations in routing context are routing tables and routing update messages.

Overlaps: Computer networks are designed to be able to self-organize. However, the extent to which they can do so mainly depends on the operator and is significantly determined by the size and complexity of the network. Complex networks tend to be organized on much higher levels of abstraction and automation. Self-organization implicates that the network itself is used for its own configuration by in-band signaling of control data inside payload data. As a consequence, some control plane information can be extracted from the data plane, that is when control information is exchanged between network elements, e. g., SNMP or routing traffic. The classification of network traffic to either the control plane or data plane may even depend on the type of measurement. As an example, DNS messages can be considered to be control plane traffic when investigating DNS-based traffic engineering, but data plane traffic when extracting an application mix of network traffic.

Limitations of our Models

Although we can now refer to a vantage point more precisely, there are limitations in characterizing a vantage point in technical terms for the following three reasons:

1. The Internet is constantly changing.
2. The role of some network devices and mechanisms is unknown, meaning that it cannot be determined by measurements.
3. The purpose of some network traffic remains unknown.

Regarding the first reason, the constant changing of the Internet, business models and traffic patterns are created all the time. Although we do have an idea of what the main models are nowadays and how they are applied to the Internet, we do not have ground truth regarding the intended role of particular traffic and how that role may currently be changing. We illustrate this with the example of the DNS. While the DNS was initially only designed and used as a mapping mechanism for names to numbers, its role has changed. At this point, the use of the DNS has been extended to the implementation of traffic engineering. Some other game changers are NNTP [99], HTTP, and peer-to-peer traffic for content distribution purposes.

We illustrate the second reason with the example of content distribution. Caching is a mechanism that can take place on different layers, and it can also include recoding of data formats on demand. The processing at such “middle-boxes” can be complex and involve operations on more than one layer. Merely observing network traffic of such a device does not show us the truth regarding its role if we do not know about its purpose in advance.

A typical cause of the third reason is anonymization, which aims to hide an identity behind communication by using encryption.

At times, multiple limitations apply. For example, the TOR [77] and JAP [21] anonymization services both rely on relaying traffic through a cascade of middle-boxes and use encryption. That makes it impossible to infer the type of communication and on the true end-points.

Type	Bytes	% in a 1500 frame
PCAP	20	1.3
ERF	16+	1.1+
Ethernet	14	0.9
IPv4	20 - 60	1.3 - 4.0
IPv6	40+	2.6+
UDP	8	0.5
TCP	20 - 60	1.3 - 4.0

Table 3.1: Overhead components in packet trace data

3.1.2 Measurement Data

The present work deals with different data types or formats coming from direct or indirect (i. e., pre-computed or compressed) measurement outputs. In most cases, the format is bound to the type of measurement and the underlying technology. Over the course of the work on this thesis, we focus on the following:

1. Packet traces
2. Flow traces
3. Logs
4. State (BGP)
5. Events
6. Configurations
7. Bulk data
8. Structured information

Packet Traces

Storing network packet traces involves two types of information: the network data themselves, already including (network) protocol headers, and the meta-information that is stored alongside each packet. While the data section includes everything from (usually) layer-2[2] up to the bulk data of application-specific information, the meta-information contains the snap length, system time of capturing, and link information.

There is still information that is not included in the data and meta-data in PCAP and ERF files: the high-level meta-information concerning the trace. In our experience, a lack of high-level information makes it difficult to use foreign and historical data sets. It also makes it difficult to put data in context, and in fact renders them entirely useless, particularly if data are anonymized. High-level information that should be documented along the trace is described in Section 4.1.5.

[2] in the ISO/OSI layering model

In terms of correctness and representativeness, packet traces contain the most complete information. Most other formats, such as flow information and events, can be compiled from that format.

Flow Traces

A highly different methodology for capturing network traffic information is the collection of flow traces. In general, we refer to a flow as a (unidirectional) sequence of packets that share the same tuples of properties. Depending on the scenario, the properties can vary. The most commonly used tuples in TCP/IP networking environments are source IP, destination IP, transport protocol, source port, and destination port. The purpose of those flow formats is to provide a high-level overview of network traffic without payload recording. For different purposes, flow formats have been standardized; the most popular formats are SFLOW and NETFLOW. The formats provide different types of information. A selection of commonly used flow-representing formats (SFLOW, NETFLOW and IPFIX) are described in Section 2.5.

Logs

Logs are a method of storing pre-processed data in a structured or semi-structured format. The generation of logs requires computational resources (compared to storing flow or trace formats) and provides a much more high-level view of the data set. Data are usually stored line-based (in file terminology) or as entity (in database terminology). The type of information is generally unrestricted, and such an entity can usually be considered to be an event.

State

In contrast to events, state refers to information that describes a system (or subsystem) at a particular time. Examples of state information are the following:

- routing state
- operational state
- resource utilization

Configurations

Flexibility is a major design goal of many operational systems. This flexibility is achieved by changing the configuration and consequently the behavior and appearance of devices, processes, and systems. For usability reasons, configurations are often flat-structured in key-value pairs (e. g., in the DNS or mail system configuration), but hierarchical data structures (e. g., in Cisco IOS) are also common. The configuration of a system can be considered to be part of its state because it covers details of the operational behavior.

Structured Data

Structured text is a convenient type of data to process on a computer system. The structure can be explicit (such as in log files) or implicit (such as in XML format). The format does not have to be line-based, yet it often is – if only for human readability. Line-based formats are common for configurations as well as logs, such as the DNS zone definition or Web server process logs. Other information, such as email, is also structured but is not stored line-based. In this example, our options for storing are modifications of the raw data by:

1. making the structure explicit by describing the format with some sort of meta information; or
2. making the structure implicit and recognizable by structuring into XML.

Bulk Data

Bulk unstructured data are more difficult to process than structured data. In fact, sometimes "bulk data" is simply wrongly referred to when data structure is unknown (i. e., proprietary) or difficult to parse. Examples of bulk data are "blobs", i. e., network information that cannot be parsed because of encryption or files in a proprietary format or unstructured text. Bulk data are of limited use without meta-information that describes at least the context in which they are created, collected, or used.

3.2 Internet Exchange Points

3.2.1 IXP 101

A current and commonly used definition of the term IXP is given by [1]: *a network infrastructure with the purpose of facilitating the exchange of Internet traffic between Ass, and operating below layer 3. The number of connected ASs should be at least three, and there must be a clear and open policy for others to join.*

The term IXP as such describes a technical role as well as a business type in the Internet context. There are IXPs with different business models, operations, service offerings, architecture, shapes, and sizes. However, most of them share common aspects and properties around the world, which makes it possible to assign Internet players the role of "IXP".

Historical Context and Annotations

IXPs originated in the mid 1990s, when the National Science Foundation Network (NSFNET) moved towards opening its backbone to private business. Back then, the network devised the plan of establishing Network Attachment Points (NAPs) in the four major US locations of New York/New Jersey, Chicago, Washington D.C., and California, all operated by a different telecom provider (telco) (Sprint, Ameritech, MFS, and Pacific Bell). The purpose was to let commercial ISPs connect at those network facilities and exchange traffic with each other and the NSFNET backbone.

This is commonly considered to be the birth of today's Internet: a network of networks. It started out as a government-funded academic experiment, and then a diverse set of players interconnected in order to use the network for business purposes, especially selling services to business partners and other customers. The four former NAPs have been replaced over time by over 80 IXPs (as they are now called). Geographically they are distributed over all major cities and metropolitan areas in the US. Worldwide, we know of more then 300 IXPs. While most of them in the US area are operated for commercial profit, there are a few that follow an open and non-profit model and usually apply a donor model or low fees to connect and exchange Internet traffic, in the tradition of the first four nodes back in 1995. Both terms, "NAP" and "IXP", are used to describe the same business type and technical role in the Internet. Which of both terms is used in a context depends mostly on its geographic region. While most players in the business stick to NAP or IXP, a major category of IXPs refer to themselves as IBX (Internet Business Exchange) data centers, interconnection services, and co-location. Examples are Equinix and Telehouse, which base their business on connectivity but charge mostly for other services. Thus, they are not as forthcoming about their network strategies, operations, and interconnection capacities.

Europe has an entirely other tradition and, as a result, another model in the IXP business. The oldest NAP in Europe dates back to the late 80s: CIXP, the CERN IXP in Geneva. Then, in the early 1990s similar facilities became operational, especially today's biggest players in the locations of Amsterdam, London, and Frankfurt. Internet traffic business began to develop quickly and European ISPs soon realized that connecting to each other involved network paths that went overseas and back at tremendous transatlantic bandwidth costs. This required a commonly shared infrastructure to interconnect, and so the European IXPs became the local interconnection places in the Europe region. This is one of the major reasons why the business model developed in a different way and is not selling-based as much cooperative. In fact, DE-CIX, one of the three largest IXPs in Europe and currently the one with the highest volume of traffic, is owned and driven by ECO, the German association of Internet business.

The European cooperative IXP model has evolved faster and to be larger than its US counterpart model over the last 15-20 years. We counted more than 190 European IXPs by late 2015. The variability of business models and size is high, and ranges from small local players (e. g., B-CIX, MAE, AIX, GIX, IXNN) to critical regional players (e. g., GRIX, UA-IX, ESPANIX, PLIX, France-IX, ECIX, VIX) to global players (e. g., AMS-IX, DE-CIX, LINX). In terms of daily traffic volume and peaks, the largest European IXPs keep up with the largest global tier-1 ISPs. The reason is the large number of networks (600^+ with the biggest IXPs by mid 2015) that interconnect at the IXP facilities. A particularly noticeable fact is the openness of many IXPs that follow the "European model". Over websites and other resources, they share information on memberships, traffic volume statistics, infrastructure details, and services offered.

Basic Operations and Services

The definition of IXPs [1] permits a flexible interpretation of the businesses, facilities, and infrastructures to which the term can be applied. In the simplest case, an IXP can consist of a single small, unmanaged Ethernet switch in a building's basement. In fact, most of the IXPs started out their operation in this way. At the other end of the scale are huge deployments, such as the DE-CIX Apollon platform that interconnects members with up to 100 Gb/s single line speed. Its distributed four super-nodes[3] employ the ADVA's flagship FSP 3000 scalable optical 142 transport solution. The super-nodes are distributed throughout highly resilient and secure data centers across Frankfurt am Main. The platform is built on a switching layer and is supported by Alcatel-Lucent's 7950 Extensible Routing System. Figure 3.4[4] presents an overview of the network topology; it shows the different levels of redundancy and aggregation. The backbone consists of DE-CIX2, DE-CIX6, DE-CIX6, and DE-CIX7, which are fully connected to each other. Each of those nodes is centered by an Alcatel-Lucent 7950XRS20 core node, and connects a) fully meshed over ADVA FSP3000R7 Interconnection-Connections, and b) to the local 7950XRS40 edge nodes. The edge nodes in turn connect to a) customers on-site, b) co-location customers, and c) customers at remote locations via pairs of FSP300R7s.

As seen from the DE-CIX example, the setup can be distributed over multiple locations. In fact, the locations can even span larger geographical areas and allow for remote peering, e. g., in the case of ECIX. Global players even have facilities around the globe, such as Equinix Internet Exchange, which operates 19 locations in 17 metro areas. Another model is the “IXP of IXPs” model, such as Netnod in Scandinavia, or AMS-IX in Amsterdam and Hong Kong. The interconnections can be based on different technical and business models. In some cases, IXPs connect to each other, and again we observe different business models.

Examples of interconnecting IXPs are Lyonix/Topix (France-Italy) and BalkanIX/InterLAN (Bulgaria/Romania). IXPs also interconnect locally to each other, such as France-IX/SFINX (both Paris, France). Independent of size and business model, all IXPs aim for high resilience. This is achieved by using sophisticated architectures that can continue operation even when facing component failure in their backbone. In addition, a high level of reliability stems from data center standards, such as full UPS backup, heating and ventilation mechanisms, and a high level of access control and monitoring for security. Some of the larger IXPs are considered “critical for national cyber defense”, e. g., LINX [145].

IXPs generally operate below the network layer, and mostly rely on Ethernet switching networks. DE-CIX is currently an exception, using MPLS instead. All participating networks (members) connect on that layer and have to fulfill a few requirements to interconnect:

1. The general terms and conditions of the IXP have to be accepted. This often contains restrictions that, if violated, lead to totally blocking the member on the IXP's platform, or other strong consequences.
2. Each member must own an ASN.

[3] The announcement is to scale it up to 10 of those super-nodes

[4] source: DE-CIX

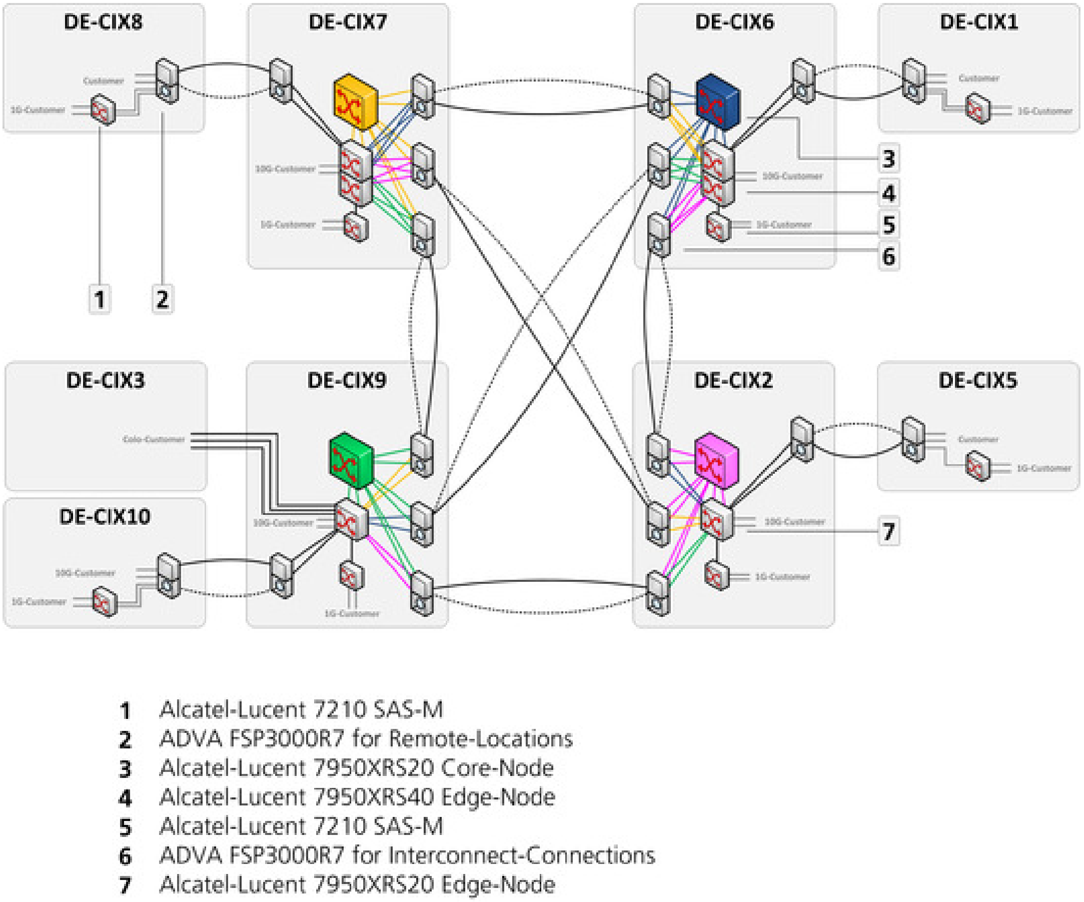


Figure 3.4: Example of an IXP topology: DE-CIX in Frankfurt. [Source: DE-CIX]

3. Each member deploys a router in at least one of the IXP's facilities. The router connects to the link layer network of the IXP as well as to the member's WAN ports, which must terminate on the member's network infrastructure. Physically, this can be anywhere in the world, often by traversing other carrier networks or dedicated fibers. Some IXPs support remote peering, where the member's traffic is exchanged via Ethernet-over-MPLS and the members connect to some transport network.
4. All routers exclusively run BGP as routing protocol towards the IXP's network. Common practice is to filter out other routing protocols or block a member's port if unwanted routing traffic is detected.

Connecting to an IXP and to publicly peer at the IXP involves a number of steps and requirements (each of which comes with some costs). First, the connecting network must find its network link to the IXP. This can be a direct fiber or involve a local POP that is able to connect to the IXP. Most smaller networks do not buy a dedicated link all the way to the IXP. Once the circuit to the IXP is established (and paid for monthly) and a one-time connection fee to the IXP has been paid, the port is available. A port comes with a certain capacity limit, and the fee for the port is paid monthly according to its capacity. Exchanging traffic with other peers over the port is not subject to fees, be they volume fees or any other fees. In general, the parties are free to exchange traffic according to bilateral conditions. This can even involve money flow from one member to another ("paid peering"). Charging for transit over an ISP is also a common business practice, i. e., by resellers. The time it takes to connect to an IXP in the first place depends mostly on the time it takes to obtain a link to that IXP's facility and to comply with the IXP's technical requirements in the GTC. Here, the practice is to connect a starting member to a test network and switch him over to the peering platform when the traffic complies with the IXP's rules. The procedure can take as little time as days. However, we have also heard of cases in which it took longer than two years.

One of the main reasons for IXP's existence is to keep local traffic local. However, there are good reasons for businesses not to connect to the local IXP, or to only use a subset of the services of the IXP. For one, connecting to an IXP involves a fixed cost – paid not only to the IXP but also to the carriers that connect a business to an IXP. It may be cheaper to connect upstream to one or multiple ISP(s). ISPs' fees are usually based on traffic volume and this may pay off for low traffic volume business. Moreover, peering at an IXP requires the ability to run a network for this environment. Some businesses do not consider this to be their competence, and therefore connect via ISPs that can handle these requirements for their customers. Another reason is QoS. Buying services from an ISP is easier than negotiating with all IXP peering partners. However, the QoS that a network can obtain via an IXP may well be better than that of an upstream ISP due to routing efficiency, path length, etc. A huge incentive to connect to IXPs is the presence or proximity of major content providers and content delivery networks at the IXP. Most of them participate in IXPs and provide enough network traffic capacity.

Another good reason for joining one or more IXP(s) is that not only existing peering partners are potentially present, but a number of other willing peering partners are also already there. The traditional way of peering is private peering, where two parties agree on a peering contract, usually set up routing, and exchange traffic. This private peering can happen at the IXP without other IXP members noticing. This service is called "private interconnect", and using such a dedicated service provides a direct link of high bidirectional capacity and stability because state of the public peering platform does not affect the private interconnect. However for most of the smaller member networks, the IXP is interesting because they are offered public peering with many new partners that already are at the IXP (or join later). The more, larger members there are at an IXP, the more interesting the IXP becomes for other networks. The effect is comparable to "gravity", where larger-sized objects increasingly attract the "attention" of other objects, and this in turn increases attractiveness again. The IXP can offer additional services to its members either for free or for additional payment.

IXPs are expected to provide additional services, such as statistics for each member, the operation of route-servers, and service level agreements (SLA). An increasing number of IXPs meet these demands. Operating a RS at the IXP relieves participants from establishing and maintaining peering sessions with each peer. Considering full BGP meshing of all r routers of all members, the amount of peering sessions for each peer decreases from $(r-1)$ to 1. The number of all sessions at the IXP in turn decreases from $r*(r-1)/2$ to r. However, the largest advantage for each member is not having to establish one more extra peering session with each joining router of a member.

Other services that help to decide whether to connect to an IXP are resellers. Instead of carrying the burden of connecting to the IXP and paying a fee that cannot be justified, a party connects to a nearby re-seller or member of a partner program. In this way, links to the IXP can be shared among the parties and the cost is split by any metric that the reseller chooses. Those third-party resellers and partners are usually certified and permitted to operate such business by the IXP. Examples of resellers are Atrato and IX Reach, both of which connect customers to multiple IXPs. The same services offer “remote peering”, a service that makes it possible to peer at an IXP even over large geographic distances and at all of the IXP’s facilities. It also relieves a peering party from buying equipment into each facility where it wants to peer. For example, AMS-IX provides such a service, so every participant in AMS-IX Amsterdam can also peer with any participant who connects at AMS-IX HK in Hong Kong. Networks with low traffic demand can profit from such programs and gather the IXPs. Other services that IXPs offer are mobile peering (i. e., peering of GSM/UMTS networks [6]) and black-holing, where members can announce a prefix to a black hole destination if they do not want to receive traffic from that network. This stops some types of misbehavior, misconfiguration, and attacks at the IXP, and does not even reach the members. A few IXPs choose to offer service for the “greater good of the Internet”, such as network time service (NTP), speed test reference points [37], or DNS root-servers.

Business and Operational Models

The worldwide market of IXP business distinguishes between “for-profit” and “non-profit” models. In turn, the non-profits are separated into “managed” and “cooperative” models. The for-profit ones are mainly in the US, while the non-profits have evolved mostly in Europe. Nevertheless, the US market is not exclusively run by for-profit IXPs, and the European market does not exclude for-profit IXPs.

US-model: US model: The Network Attachment Points’ (NAP), i. e., IXPs, business in the US market is dominated, yet not totally owned, by the for-profit model, also referred to as the “US model” later on in this thesis, where IXPs sell connectivity to customers for profit. The first and foremost goal of this economy is to generate profits for shareholders. The business as such evolves mainly for monetary incentives. The information on the infrastructure, operation, statistics, business, and services of those players is rare compared to that on players that operate the “European model” (next paragraph). Thus, our understanding of the North American marketplace is lower than our understanding of the European IXP market. There are exceptions in the US market, namely operators such as Equinix and Telehouse America, which do not advertise their role and services under the term IXP, but rather focus on deployment and management of big data centers and co-location facilities around the globe. However, they also offer IXP-specific services. The main business is to connect customers worldwide. Equinix [30], for example claims to connect peers at 19 different IXP locations in 17 global metro areas. Their customers are 750^{+} networks, among them content providers, content delivery networks, and cloud providers.

European model: European model: The “European model”, on the other hand, is based on openness and resource sharing for “the greater good of the Internet”. This business is run by a cooperation of participants (not members or customers) to provide them with services instead of accumulating revenue for shareholders. To illustrate the model, we use DE-CIX. Fees cover the operation and the scaling of the IXP in the participants' interests. DE-CIX is owned by ECO, the association of German Internet business, which is also the world's largest non-profit association of Internet industry. Decisions are made by ECO, to which the DE-CIX management reports. The association drives the development and operation of the IXP. ECO member meetings and voting are how participants influence the direction that DE-CIX takes as well as DE-CIX's direct operation. Enlightening here is the observation that the three largest IXPs worldwide, LINX, AMS-IX, and DE-CIX, run exactly this model. A major advantage to the US model is, that due to the openness regarding facts and data, high-level information about the operation of the European model IXPs is easier available. They include outreach activities, extension plans and status, network architectures, specifications and policies, descriptions, lists of connected members, aggregated statistics, offered services, and pricing information.

There are other IXP markets in the world besides the US and EU, but they also use one of the models discussed above. Africa and Latin America have mostly adopted the European model. In Asia, both models co-exist, but the "for-profit" one is concentrated in the more developed countries, while the "non-profit" model is more prevalent in less developed countries. Several attempts have been made to enforce specific behavior at IXPs (i.e. by governments), but this has not improved IXPs' participation. One example is the rule of "forced peering", according to which an IXP member must peer with all other members.

3.2.2 Under the Surface

The IXP business has been highly successful in the last 15-20 years. This section discusses some of the success factors that helped to increase the popularity and attraction of this business in the Internet ecosystem. For this, we rely on public information regarding these IXPs.

Know your Data Sources

To study the worldwide IXP business, a set of well-known and commonly used data sources are useful. Most are publicly available. However, we want to point out that these data need to be read with care. Information can be incomplete, inaccurate, or simply out of date. Naming can be misleading and research results have been published based on a wrong understanding and interpretation of such data. In our experience, especially with IXPs, it is advantageous to speak to the data collectors to clarify what is in the data and what is missing on the one hand, and to clarify wording in the data description on the other hand. Different instances of data sources often use different data sanitization, filtering, and wording, which, to compare data sources to each other, have to be unified first. Another kind of bias is the handling of outdated information by data providers. Packet Clearing House (PCH)4, a non-profit research institute, for instance, is known to not delete members and data from its lists once they are no longer members. Instead they mark them as "inactive" and so, over time, data quality worsens.

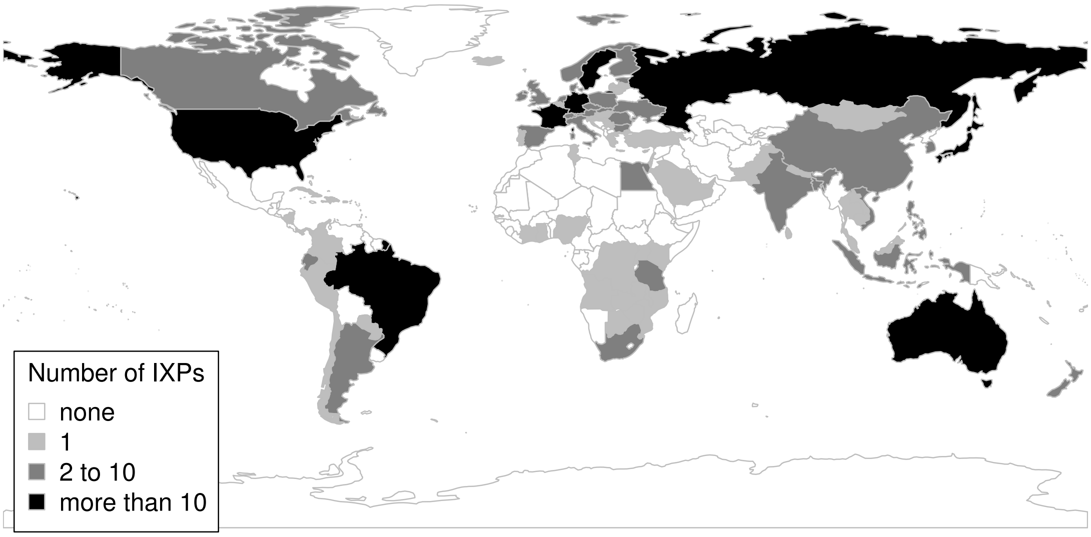


Figure 3.5: Number of IXPs per country (data from PCH).

That said, we use data from the following sources in the next step of our IXP operation and business analysis:

- **Peering-DB** [29] is a widely used peering database maintained by PCH [27], and is an often-cited IXP directory with associated meta-data, detailed lists, repositories, and reports. All of this information is provided by non-profit organizations, e. g., Euro-IX [15], the European Internet Exchange Association. The information must be handled with care, because from cross checks with other sources (i. e., IXP-proprietary measurements that reflect actual traffic, peerings, policies, etc.) [43] it is clear that the data are not entirely reliable, although it is considered to be the industry standard database. In fact, the data source is known for being relied on by network operators [17].
- **Data Center Map** [10] is an IXP database provided as a free Web service. This information is considered to be the link between data providers and customers in the data center industry. Although Data Center Map provides information regarding business and some relations, it is incomplete. This is because the data are published voluntarily and on the terms of the data providers. Nevertheless, we consider the data valid.

Managed non-profit IXPs often publish reliable information. Their motivation is the fact that the number of members and the amount of exchanged traffic are a success indicator among those operators. From looking into the publicly available data and into flow data over a period of more than one year, we can attest to there being correct information published for that particular IXP. However, not all information of IXPs is public. The reason for that is the protection of member interests: preserving business relationships and operational details. Thus, most IXPs do not publish information about the peering matrix (who exchanges traffic with whom) or the traffic matrix (how much traffic is exchanged on particular links and between members). This is infor-

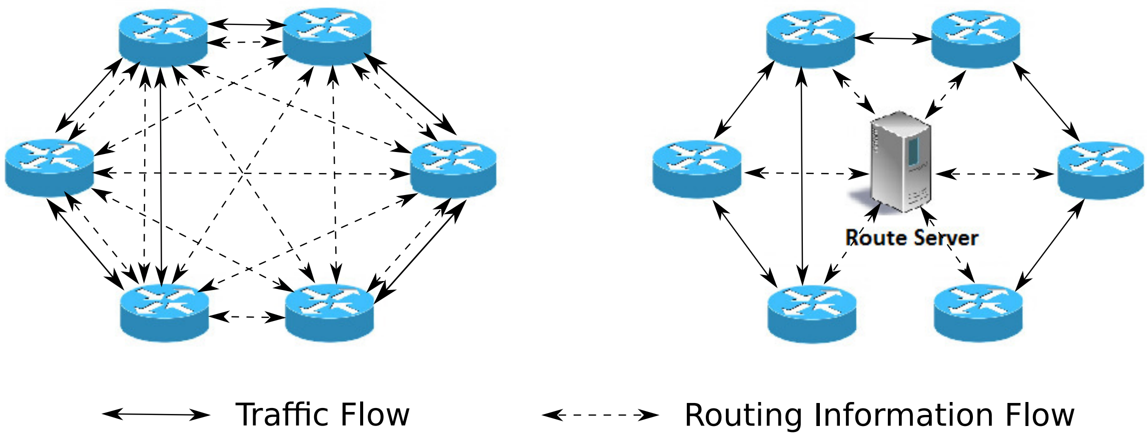


Figure 3.6: Multi-lateral peering in an IXP without a route server (left) vs. corresponding traffic and routing information flow in an IXP with a route server (right).

mation that is of interest to the Internet research community but is not publicly available. Without this information, the network research community must rely on less accurate published information. This information, such as AS-level interconnectivity and traffic flows, is usually gathered by using BGP measurements that are publicly available, such as RIPE RIS [31] or Oregon RouteViews [32]. This information is known to be incomplete, and so it was not surprising when in 2012 Ager et al. [43] revealed numbers regarding Inter-AS links at a single IXP which exceeded the information of all worldwide publicly available data sets combined. The network research community did know about the underestimating of the peerings from non-public information, and termed the phenomenon “invisible links” [126].

3.2.3 Surprised by the Obvious

When the number of 400^+ connected networks and a peering density of over $60^+\%$ was published in [43], the network research community was astonished. In fact, the analysis revealed more peerings ($50K^+$) in this single data set from a single IXP than the number of peerings that were known throughout the entire Internet from all other sources combined. This drastically changed the view on the peering business. While the ratio of customer-provider links to peer-to-peer links was estimated to be 3:1, it now seems to be closer to 1:2 or 1:3, which means much more peering. The impact was that our economic view of the peering business was reversed. Even by conservative estimations, there are easily more than $200K$ public peering links in today’s Internet.

Knowing that there are so many peerings at a single location raises the question of how the peering overhead is handled, namely the BGP sessions that a member has to maintain in order to connect to 700^+ other ASs. The answer is in the usage of route servers, which was barely known to the network research community, mostly because the extension of public peering was underestimated.

Year	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
Average Traffic (Gb/s)	13.4	24.9	58.1	110.3	193.3	288.5
Peak Traffic (Gb/s)	22.0	47.1	119.6	220.0	374.2	608.8
Number of Members	178	211	234	253	290	317
Year	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
Average Traffic (Gb/s)	443.9	623.5	815.6	1057.0	1357.6	1789.0
Peak Traffic (Gb/s)	856.6	1186.0	1458.3	2042.7	2516.4	3320.1
Number of Members	349	388	469	555	591	677

Table 3.2: Annual traffic and member statistics for AMS-IX.

Route servers are the tool to reduce the complexity of peering session maintenance. The way in which a route servers works is that a member replaces all (or a selection of) BGP sessions to other members with a single session with the route server. Figure 3.6 depicts the native BGP operation among peers without RS compared to BGP operation including a RS. Indeed, to avoid a single point of failure, most IXPs offer more than one RS per peering location, and most peers make use of that redundancy. The RS collects the information from all peers, applies filtering, and distributes the routes to the other peers, much like multicast The filtering [7] can include sanity checking of the announcements, and the unified use of the BGP community feature allows a peer to select destinations of route announcements. As a result, the information that a border router learns from the RS equals the information that it would otherwise have to collect from multiple peering routers. The route service is given away for free to the members, and members are encouraged to use the feature because it facilitates multilateral peering. In addition, the RS operation does not prevent direct peering among members. More insight on the operation of a RS is given in the RS analysis in Chapter 9.

3.2.4 Non-profit Matches Profit

European IXPs have grown significantly over the last years. The number of connected networks has increased, as has the total exchanged traffic volume. In this regard, we observe some common developments. Year after year, the exchanged traffic volume increases by about $30-100\%$, and the largest European IXPs (DE-CIX and AMS-IX) had membership in the range of $600-750$ by the end of 2015. Moreover, the number of connected networks grows by about $10-20\%$ per year. The additional bandwidth demand is met by higher port speeds that are sold to existing and new customers[11, 13, 5], as well as by improved core infrastructures. Table 3.2 illustrates this with the numbers of AMS-IX for the years 2003 to 2013. The magnitude of exchanged traffic volume per business day has become equal to those of major ISPs, such as AT&T (33PB [8]) or Deutsche Telekom (16PB[12]). AMS-IX traffic, peak, and average are plotted in Figure 3.7.

The non-profit IXPs are strictly neutral and open to doing business with any network and independent company [36]. However, in order to expand their business and reachability geographically and strategically, they rent at locations that are run by for-profit organizations (even IXPs), such as

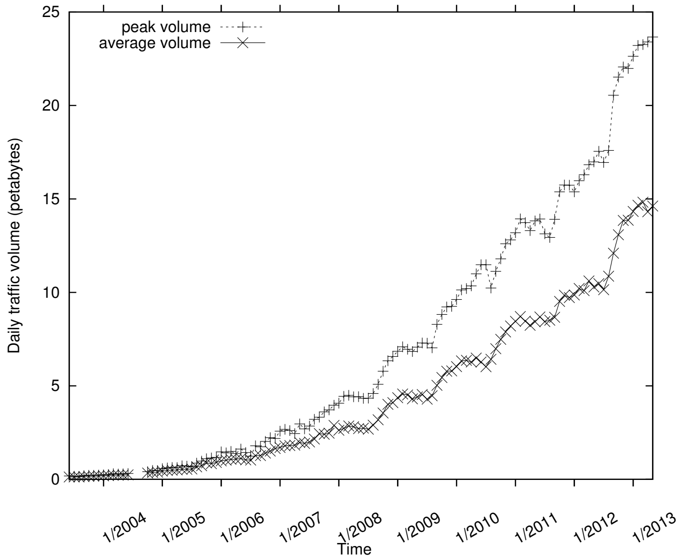


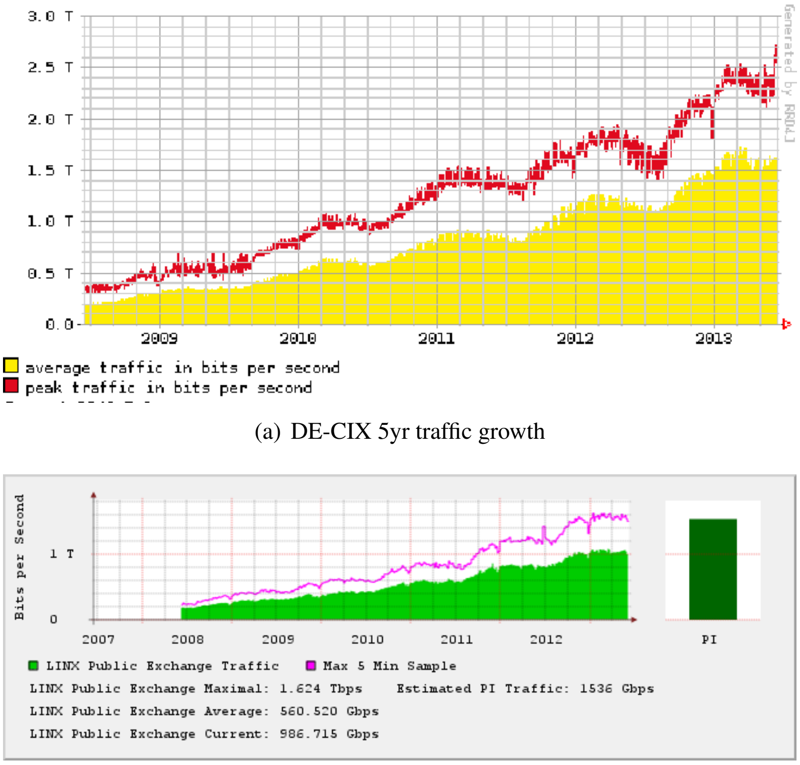
Figure 3.7: Daily traffic volumes of AMS-IX in Amsterdam (2003 – 2013).

commercial co-locations, data centers, and carrier-neutral facilities. Well-known for-profit organizations running those facilities all over the world are Interxion, Equinix, and Telehouse. Those organizations welcome the business from non-profit IXPs for several reasons. The most important is that the for-profit IXPs in fact profit from the presence of the non-profits because they bring customers to the facilities who likely make use of co-location, which is the main business of for-profits. Thus in the market of non-profits, the data center space increases in value with the presence of non-profit members. This is also why the for-profits prefer those non-profits as customers: they bring along many members as data center customers. This model is followed by Equinix and Telehouse America's Global Interlink service, among others.

The business relations between the complementary business models (for-profits and non-profits) increasingly tend to turn into join-ventures. Along those lines, the German non-profit IXP DE-CIX announced in a press release a partnership with Equinix by connecting three more Equinix co-locations (Points of Presence). In this way, Equinix gains access to DE-CIX members and DE-CIX makes use of its high performance "Apollon" platform by selling more ports and capacity to Equinix as well as to its own members. This can be seen simply as a new direction in the continued growth of DE-CIX. For Equinix, on the other hand, this partnership results in better connectivity to potential customers. However, we do not expect that more information will be published about the non-profit member base or the non-profit IXPs' data flows. Such visibility will remain limited.

3.2.5 Challenges and Opportunities

Having access to operational data of even one single IXP offers good chances of enhancing research knowledge, as shown by Ager et al. [44] in 2012. These data were truly eye opening to the



(a) DE-CIX 5yr traffic growth

(b) LINX 5yr traffic growth

Figure 3.8: IXP traffic growth.

network research community, and since then more research has been conducted in order to make use of these understudied Internet players.

The most interesting types of information that we have explored to date are:

- Connectivity of ASs (traffic matrix)
- Reachability of prefixes
- Prefix role classification
- Traffic Engineering
- Traffic Classification
- AS-level limitations
- Routing information

The observed information extends the current Internet knowledge drastically by reversing the estimated ratio of customer-provider to peer-peer routing links, by the sheer amount of observed peerings, and by how routing and traffic exchange are handled at IXPs.

One particular insight that surprised us was the verification results of publicly available peering and routing information. It appears that sources, such as RIPE RIS and RouteViews, have to be handled carefully because they are incomplete, inaccurate, and lack freshness [125].

The first challenge of conducting research in the IXP ecosystem lays foremost in the access to data. As we pointed out, non-profit IXPs are much more approachable for cooperation and information sharing than their for-profit counterparts are. We have found some IXP operator communities are surprisingly open in sharing information and working towards cooperation on an operational data level. The second challenge is the sheer volume of data that is available at the operational level of an IXP. High-data-rate, low-level information must be processed online and in an anonymized fashion in order to comply with contracts and laws. Here, the particular challenge is to extract information that is aggregated to a level that allows storing without violating the terms of cooperation on the one hand, and having the aggregated information in a form that allows for consistent analysis at a later date on the other.

Limitations

There are clearly limitations to what is technically and administratively possible with an IXP a vantage point. First, the most complete information naturally comes as full packet traces. For public research, this is a) not the type of information to obtain in the first place, and b) would require huge network and processing capacities just to pre-process and record it. Considering the data volume of 21 PB/day and peak transfer rates of 3^+ Tb/s, which were measured in January 2014 at DE-CIX, this would require a bandwidth and processing speed of the peak data rate. Second, the data formats on which we rely are a considerable source of limitation. In practice, two formats of data plane information are commonly in use. SFLOW at least provides samples of the data plane, but it does not provide any real flow information (size, distribution, duration), whereas IPFIX/ NETFLOW gives (sampled) flow-specific information but does not contain any traffic samples for a traffic mix characterization. Obtaining both formats at the same time would be preferable, but is not realistic in the business of real network operators. Third, there are parts of business at an IXP that are clearly not subject to information sharing, i. e., private interconnects. Such data remain unknown and inaccessible by all means to us as well as to the IXP in some cases. Therefore, we face incompleteness due to a whole branch of invisible operation at the IXP. Yet another limitation is the access to a restricted amount as well as a restricted type of IXPs. While we managed to establish cooperation with a number of IXPs running the “European model”, and of different sizes and geographical locations and distributions, the initiation of cooperation takes time (up to half a year until contracts are signed and equipment is deployed and operational) and comparison of the data has still not been completed.

3.3 Internet Service Providers

3.3.1 Description

ISPs are mostly profit-oriented companies[5]. They offer technical platforms and support for accessing the Internet for the purpose of consumption or provision of content and other services to

[5] There are non-profit models, such as community driven ISPs but the majority of ISP is commercially operated.

other participants on the Internet. In addition, ISPs have different sets of services that they offer to customers and partners. However, an ISP cannot be exclusively reduced to a set of technology platforms that it runs: it is also defined by its interplay of those services that make the ISP a business, such as peering relations and traffic optimization capabilities.

While the Internet is a distributed and de-centrally managed system, a single ISP's network can be distributed as well, but only in the sense of physical locations: it runs subnetworks and connects in a coordinated way to other networks to conduct business. For example, Deutsche Telekom has business in Europe, Asia, Africa, and America[6]. Yet, the company runs a single AS and is supposed to act as a single entity "AS3320" towards the other players on the Internet. From an administrative point of view, ISPs networks are centrally managed, since an ISP is either part of a larger business or it is a business on its own. ISPs take care of the coordination and optimization of their own network, the offering and improvement of services, and manage the relations with other Internet players. The operation of an ISP business requires the ownership of one or more AS(s).

3.3.2 Typical Services

There is no well-defined set of services that defines an Internet business as an ISP. However, there are a few categories of services that contribute to its business. As the Internet and its network services evolve, the services of ISPs also have to increase to meet the demands of customers in the market and to enable cooperation with partners. Typical classes of ISP service are providing connectivity, providing or delivering content, and running or hosting services.

Physical Connectivity

A common service is the provision of connectivity for customers to other parts of the Internet by enabling IP packet transit over a technical infrastructure. There are different methods and technologies available, including the following:

- Consumer-grade broadband connectivity, e. g., xDSL, cable, FTTx
- Leased lines and PoPs
- Radio frequency technologies, e. g., radio links or satellite
- Short-range wireless technologies, e. g., wireless access points

Most industry-grade ISPs, e. g., Telefonica, British Telecom, and AT&T, operate extensive access platforms and backbones to connect their customers to the rest of the Internet via multiple links to other ISPs, CDNs, carriers, and exchange points. Those ISPs are referred to as eyeball ISPs if they connect a large base of private and business end-customers, such as single persons and businesses that are characterized by demand or consumption, instead of providing services.

[6] according to their website `http://www.telekom-icss.com/ournetwork`

IP Connectivity

Networks, including ISPs, cannot directly connect to all other network operators in the world. Thus they have to cooperate and transport traffic on each other's behalf to relay traffic of networks that are not directly connected, called transit. Transit is generally based on one of two business relationship models in peering: customer-provider or peer-to-peer (P2P). In the customer-provider relationship, the customer pays the provider to be reachable and to be able to reach a defined portion of the different Internet address spaces. On the other hand, the P2P model connects the networks of two parties to exchange traffic between them. While P2P was initially realized without payment and "settlement free", other "paid peerings" have evolved, which involve charging, for instance based on the ratio of transmitted and received traffic.

Content Provision

In the early stages of the Internet, the main business model was based on traffic volumes that were exchanged and on connectivity that could be offered (see above). Content and advertisement business has replaced this model. Popular content providers, such as Google ($10.7 billion net income in 2013), have surpassed major ISPs, such as AT&T ($7.3 billion net income) and British Telecom ($2 billion net income). The market and technologies for content delivery mostly focus on dedicated content delivery networks (CDN), which is a transition from concentrated yet decentralized models into highly decentralized and distributed delivery models. For example, Google and Akamai drive initiatives to have their content delivery equipment deeply deployed inside the networks of eyeball ISPs. They also distribute an increasing amount of formerly central located computing resources to other businesses networks.

Hosting

In order to deploy computing resources in the first place or to place computing resources into network segments, hosting is a way to buy services instead of distributing hardware. Usually, the hosting model involves renting hardware and network capacity from an ISP "hoster" and running services on that infrastructure. We have even seen hybrid CDNs that use hosted resources, as shown in Section 3.2. Some hosters offer to deploy customer-owned hardware, but the extent of this sort of service is small compared to the overall housing business (see below). Hosting is highly dynamic and allows the hoster ISP to optimize the utilization of resources within its own data center. Another recent development is the offering of cloud services, which is the improvement of hosting towards selling virtualized resources, such as network connectivity, computation power, and storage capacity.

Co-location and Housing

When service providers need physical presence within or in direct reach of other networks, they can buy space, power, and connectivity in a co-location facility. The main difference compared to hosting is the ownership of the hardware – and as a consequence its administration. In many cases, co-location and housing facilities have connectivity that is not provided by the hoster itself. Co-location and housing are offered by businesses ranging from small local dedicated companies to IXP-grade businesses, e. g., Equinix and Interxion.

Other ISP Services and Challenges

Next, we discuss some additional challenges that eyeball ISPs have to tackle. Although net neutrality requires the same treatment of all traffic types, some services require a particularly optimized network behavior. QoS has recently attracted public interest because ISPs tend to migrate their voice and media platforms to an all-over-IP model to save cost and complexity. Of course, if television (TV), phone, and bulk internet traffic use the same physical line, the interference can impact Voice over IP (VOIP) and TV up to a point where people become annoyed (Quality of Experience). Another direction of development within a major European ISP is the convergence of platforms. Having different technologies, e. g., mobile Internet over LTE, UMTS, or GPRS, on the same technical platform as xDSL and FTTx not only simplifies deployment, but also enables the combination of access technologies. Most of the ISPs' concerns regard some sort of network optimization, be it within their own network or in the interplay with other networks. A particular problem that ISPs face is the distribution of money. Since the model of payment on the Internet has changed from selling and buying traffic to selling and buying service, those ISPs that focus on selling network capacity struggle to exist.

3.3.3 Information Availability

Due to the nature of business and competition, in the vast majority of cases the operation of an ISP is not publicly available. Although the ISPs publish business facts and public relation information, none of the tier-1 or tier-2 ISPs offer details regarding the operation or data to the wide public. This makes it difficult to obtain ground truth. Necessary technical and organizational data include:

- **Network topology**: How is the network organized internally in order to meet traffic demands and reliability?
- **Peering details**: Where and in which conditions does the ISP peer with business partners, and who are those partners?
- **Traffic characterization**: What services are popular, and how are the traffic flow and utilization of deployments?
- **Cooperation models**: Which services are run within an ISP that are affected and influenced by third parties?

Some ISPs cooperate with researchers. In those cases, some ground truth information is provided along with non-disclosure contracts. As a consequence, the research results can hardly be verified or replicated by other parties. However, some general technologies and practices are common among operators and are shared with the community, such as (high-level) network layouts from access, aggregation, and core networks.

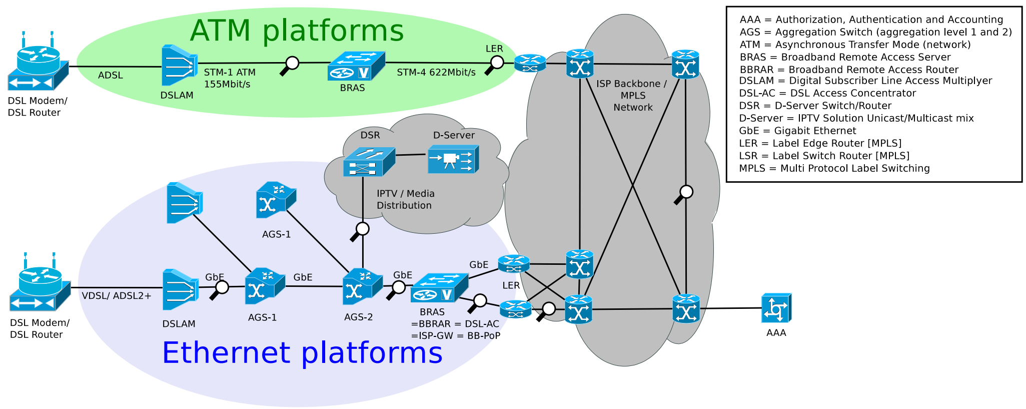

ATM platforms
DSL Modem/
DSL Router
ADSL
DSLAM
STM-1 ATM
155Mbit/s
BRAS
STM-4 622Mbit/s
LER
ISP Backbone /
MPLS
Network
DSR
D-Server
IPTV / Media
Distribution
AGS-1
DSL Modem/
DSL Router
VDSL/ ADSL2+
DSLAM
GbE
AGS-1
GbE
AGS-2
GbE
BRAS
=BBRAR = DSL-AC
=ISP-GW = BB-PoP
GbE
LER
AAA
Ethernet platforms
AAA = Authorization, Authentication and Accounting
AGS = Aggregation Switch (aggregation level 1 and 2)
ATM = Asynchronous Transfer Mode (network)
BRAS = Broadband Remote Access Server
BBRAR = Broadband Remote Access Router
DSLAM = Digital Subscriber Line Access Multiplyer
DSL-AC = DSL Access Concentrator
DSR = D-Server Switch/Router
D-Server = IPTV Solution Unicast/Multicast mix
GbE = Gigabit Ethernet
LER = Label Edge Router [MPLS]
LSR = Label Switch Router [MPLS]
MPLS = Multi Protocol Label Switching

Figure 3.9: Typical telco access network structure: in this case the assumption is a MPLS backbone of the ISP, alternatives are possible.

3.3.4 Layout of an ISP Eye-ball Network

Eyeball ISPs are dedicated to serving their customers' needs for Internet connectivity. The largest eyeball ISPs constitute the formerly largest national ISPs, e. g., France Telecom, Deutsche Telekom, British Telekom, and AT&T. Those networks do not only operate as end-user (i.e., consumer) networks, but also offer a large variety of different services. There are smaller and much purer eyeball ISPs, mostly highly localized ones, which run for profit as well as non-profit. Examples are NetCologne (Germany), AngloINFO Brittany (France), San Francisco Municipal Wireless (USA), and Freifunk[7] (Germany).

For our research, we have cooperated with one major European ISP. The environment where we have taken measurements includes:

- An access network for customer DSL connectivity at several aggregation levels
- Access to a single line of a trunked international link

For this, we have access to physical locations of vantage point deployments in two major European cities, which we operate over the management network of the ISP. At the locations, we obtain optically tapped full packet data of several link types and instances (see above).

Figure 3.9 provides a representative idea of how such an access networks looks.

3.3.5 Data

We obtain copies of bidirectional traffic in the access network. Unlike in IXPs, the routing is symmetric at this level of infrastructure. In other large network vantage points, the problem is to technically gain access to bidirectional traffic because of that asynchronous routing and because

[7] Freifunk can be seen as overlay ISP or virtual ISP: http://freifunk.net/

of trunking. The line speed is in the dimension of 10 Gb/s, but a few 1 Gb/s links are also operated and tapped.

In the two measurement locations, we observe data from about $10K$ and $12K$ DSL lines at the highest aggregation level per access network vantage point by October 2013. The data were measured using pseudomized radius information. The number of lines did not change drastically over 12 months.

3.3.6 Measurement Opportunities

Our measurements are viable for different cases, including IP geo-location reliability [133], analysis for possible ISP and CDN cooperation models [132], analysis for P2P locality improvement inside an ISP [46], traffic engineering inside ISPs [83], and security analysis inside the ISP [114]. Furthermore, the measurements make it possible to look behind the borders of the ISP's network[116] and to analyze the mobile device traffic of residential networks [115], the impact of evolving technologies [143], and traffic characterization [135]. Other analyses have not been published on delay measurements, TCP connection, or timing analysis.
The unique opportunity compared with other vantage points is the availability of unsampled, full packet, bidirectional, and multi-line network traffic, which allows in-depth analysis of traffic regarding timing, end-points, content, and network equipment influence. In addition, we have routing and NETFLOW information, which enriches the IP address information with inter-AS routing paths, as seen from the ISP's network.

Timing: Time-based metrics, such as delay, jitter, inter-flow and intra-flow, inter-arrival time, and the amount of parallel connections per aggregation level, are relevant for debugging networks to ensure quality of experience (QoE). Interesting examples include video codec performance under certain circumstances [122], influence of error correction mechanisms in TCP, or multi-media streaming platforms.
Endpoints and locality: The questions of interest are how much of the internet address space is exchanged traffic on the end-users' behalf, and, by connecting the information with timing, how big the influence of the amount of ASs or geo-location is on the performance of network traffic.
Content and locality: How does CDN traffic perform from a user's perspective? Starting with these questions, we can analyze how satisfied end-users may be with a certain service and how different the service "feels" when it is delivered from a specific portion of the Internet or the ISP's network. Parameters to consider are the frequency of use, service location, and network path, as well as network level quality [92] of streams and bulk traffic [143]. Since we do not have direct user experience feedback, measurements can be evaluated either with QoS models or by emulating measured network behavior to test users in laboratory environments.
Network equipment influence: Network traffic crosses not only links but also active equipment, such as routers, switches, and tunnel-encapsulation devices. Each of those devices has an impact on the traffic regarding measures for QoS and traffic engineering, as well as low-level impact, e. g., router buffering. However, not only single devices impact network behavior and performance. Other components, such as protocol selection and product combination, contribute to the behavior of network segments in terms of reliability or performance, e. g., delay or jitter. From the

ISP's perspective, it is vital to know about the behavior and impact of routers and other network components in different load scenarios. However, the whole picture also needs analysis regarding how much external networks or internal caching contribute to network performance.

3.3.7 Challenges and Limitations

The challenges and limitations regard two issues: visibility and capability.

As described in Chapter 1, the observation of all parameters that contribute to network behavior and performance is unfeasible at this scale. Although we have all information regarding network traffic on a link, we do not have full knowledge of the components that also influence the interesting aspects of network traffic, i. e., user experience. Such components are local caches (i.e. used in the DNS), state of user-side applications (i.e. Web apps) or server capacity, computational load (i.e. inside data centers, CDN deployments, and single application servers), and the state of the network outside of the ISP. Some of these are indirectly measurable, such as DNS caching behavior, but information on other components remains invisible to us, e. g., the link load and utilization on paths outside of the ISPs network.

Besides this incompleteness of information, the capacity of the measurement equipment restricts the effectiveness and completeness of our picture. First, the measurement equipment is restricted in storage and computational resources. Second, we do not have measurement equipment on each and every link or device. Even if we did, the complexity of computation is difficult to coordinate. Thus, given our measurement points, we can analyze problems and how to solve them only for similar environments. In particular, we can overcome this restriction of vantage point incompleteness with the observation of similarity of users' network traffic behavior. This makes it possible to first approach those network effects that affect most of the users, most of the applications, or the most critical applications, such as VOIP or video.

Limitations of storage and computational resources are first and foremost due to the high-speed links, and thus the speed of inbound information. We can approach the issue of space by limiting ourselves to the least amount of information that it is feasible to collect. This bears the problem of knowing all of those parameters in advance and finding ways to effectively handle space and processing. In some cases it pays off to collect full information over a technically and legally feasible time span and use it to tailor different kinds of analysis. We can use this most complete traffic sample to filter for the relevant information to keep for a particular measurement or research problem, and consequently collect only limited information but over a significantly longer time span.

An ISP is less localized than so-ho networks, for instance, but direct visibility is still local to that particular network. The ISP can extend this view by providing information on the rest of the network, e. g., flow information from border routers and inter-domain routing information. Still, the picture misses the traffic around that ISP. Effects in the Internet can be local, such as the effect of CDN deployments within ISPs. If our observed ISP does not explicitly cooperate with a particular CDN X, X's impacts may not be observed. Moreover, if a service is not delivered into the ISP, the global impact (or even the existence) of that service is easily missed. Netflix provides an example of this: the video-on-demand service is not offered for most networks in central Europe. Thus, a German or Austrian ISP will hardly see any Netflix traffic unless it provides transit. In reality, however, Netflix is a large traffic contributor in other parts of the world (US, UK) and has recently become a main contributor of network traffic volume in some central European traffic exchanges (Chapter 8).

Another limitation of examining ISP traffic is the legal situation. We must not keep traffic for an arbitrary period of time, and we must anonymize, or at least pseudomize, network traffic before processing it. This means that we can keep data only in an aggregated form, which makes it difficult to make long-term measurements. The particular methodology in a long-term observation cannot be modified because the impact cannot be inferred – except by running both analyses live on the same data.

3.3.8 Scope of the ISP Vantage Points

a The ISP vantage points with access to full packets on line speed are ideal for investigating problems and issues of ISP operation. They provide information from inside the ISP regarding user demands, trends, and problems. However, results are rarely generalizable. The view implicitly provides some information on the state of networks outside the ISP. Based on timings and hop-counts, the state of a path from/to a service outside the ISP can be observed, although the cause of that state remains unknown. Regarding user behavior, for an ISP it is crucial to know how its users demand to use the Internet in order to meet those demands, both quantitatively and qualitatively. Although technically and legally limited, and ISP has ways of observing the type and amount of devices that its users operate to access the Internet. This helps to improve localization of traffic within the network, and as such can significantly impact network performance and user experience.

4

Monitoring Internet Vantage Points

"Operating an Internet Vantage Point" can vary from deploying systems as small as a single embedded device, such as tiny RIPE measurement probes [38] to deployments as large as multiple distributed setups that operate management networks and specialized recording and computation equipment, as was shown in Section 3.3. Operation and management of such equipment cannot be done without automation in order to keep it manageable and thus secure. The level of automation, in our experience, largely depends on the measurement type, task, and variety. On the one end of the spectrum, single-purpose systems and repeated measurements are relatively easy to automate and handle. On the other end, there are multi-purpose systems and unique tasks, e. g., research or trouble-shooting, where little or no measurement automation is possible. During this study, we had the chance to plan, deploy, operate, maintain, and use several vantage point types and instances. In this chapter, we describe the monitoring points we used and how we planned, deployed, and operated them, and provide an overview of our best practices. We provide insight on the administrative, legal, and technical levels. The aim is to enable the reader to benefit from our work for the purpose of operating vantage point deployments in an efficient manner.

4.1 General Aspects of Building and Operating a Network Vantage Point

Before examining the details of operating our various vantage points and measurement, we start by providing a general description of common aspects that we faced in all setups. This is followed by a description of the operation of each vantage point covered in this thesis.

4.1.1 Cooperation and Data

For processing network data, particularly packet traces and flow traces, we must be aware of and sensitive towards the interests of other parties involved in the measurements, directly or indirectly. Particular challenges concern *privacy*, *legal restrictions*, *cooperation contracts*, and *data ownership*.

Privacy Protection

Network measurement data most likely contain information on individuals who use the network. Only in a rare case, namely machine-to-machine communication of a known type and of known communication partners, is nobody's privacy impacted by examining the traffic.

Besides legal restrictions and regulations, we are also committed to research ethics when dealing with potentially sensitive data regarding privacy, and we respect to this valuable asset.

The rule of thumb is to record and process data only with the explicit and verifiable agreement of all users whose network traffic is potentially affected. This requirement is achievable only in a few cases, e. g., when deploying a measurement to one's own household, provided that all users of the measured infrastructure are in fact informed that they are being monitored and agree to it. As soon as any other person (for instance, a neighbor) also uses the network, she also has to agree or the measurement needs to stop. This scenario highlights that the permission to conduct measurements is not a one-time issue, but has to be revisited regularly. In general, it is possible to grant access to a monitored infrastructure only if the user signs a contract that at least includes the clearance of conducting measurements, but again, every single contractor has to actively opt-in.

However, there are cases in which those users are simply not known, or there are too many to have all of them agree to a contract. In these scenarios, and depending on the local law (that is, the law under which the measurement is performed), measurements can be taken under different conditions. These are discussed in the following.

Clear and immediate purpose: Measurements that potentially involve the recording of private data must relate to solving a particular and current problem, such as reported performance issues in the network. In contrast, measurements cannot be taken for planning or monitoring network capacity or for understanding network properties.

Anonymization: Data must be anonymized to a level where a person cannot be identified and related to the network traffic. This covers two components: *(a)* the fact of communication itself, and *(b)* the content of the communication.
(a) This component relates to a person's communication patterns. It must not be possible to determine which persons communicate with each other. This includes all kinds of data that relate to a person, such as names, logins, and passwords. In a single instance of communication, this might require the anonymization of the client IP while keeping the server IP in the clear. In practice, the problem is identifying servers and making the distinction between servers and users. For example, the communication peer can be an HTTPS server, such as a Google service that is not related to a single person, or it can be a Bittorrent peer, which can clearly be related to a person.
(b) This component refers to content that may identify or relate to a person. For example, consider email between two email servers that often enough do not encrypt traffic. Although the IPs of both servers do not relate to the persons communicating, the content of the email – including headers, text, and attachments – clearly point to a set of communicating persons. Both cases, (a) and (b), must be handled in such a way that no information about a person is stored as a result of the measurement. In both cases, anonymization is required.

Cooperation Contracts

Besides the users whose network traffic is measured, there is also the aspect of confidentiality among contractors and owners of network infrastructure. In situations in which we do not operate the networks ourselves, we rely on information and resources from external cooperators. Examples of this are infrastructure access, measurement interfaces, or data sets. This cooperation is often subject to a non-disclosure contract. Part of such contracts, often enough, specify that not even the fact of the cooperation and the contract can be disclosed, let alone any details. While it is feasible to live up to such a contract in a single publication, for instance, it is more difficult to maintain the confidentiality over time, when information is aggregated and combined. Whenever we have had to deal with such contracts, we have found it to be good practice to inform the cooperator about the publication in advance and obtain an explicit permit for each one. As a helpful side effect, the contractors can indicate problems in the data or help to clarify particular statements.

Legal obligations

Legal situations are difficult to handle properly for a researcher who is not an expert in legal affairs. For cooperations and publications, we rely on assistance strictly from appointed persons and institutes. This can be the cooperation partner's legal department, or that of our university.

Ownership

Communication data are always owned by the communication partners unless stated otherwise in a legal contract. Such contracts are part of the terms of use of several services on the Internet, the most popular among them being Facebook[1], stating that "[..] *you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook* [..]".
The fact of ownership does not change over time and applies regardless of who stores the data.

Control

In all cooperations, our partners have a veto in publishing the data, and are in the position to grant or deny access to data and machinery at any point in time. For possibly legal situations, our partners are in full control of equipment and data. Moreover, as described later in Chapters 5 and 6, we always include hops in the access path to the measurement systems, which exclusively and entirely controlled by the cooperation partner, e. g., stepping stones, firewalls, and complete network segments.

[1] `http://www.facebook.com/legal/terms`

4.1.2 Ground Truth

We are cautious with the term "ground truth" here and refer to it exclusively as the information that is given to us by the responsible authorities. In particular, we never use the term for the results of our measurements. For instance, we only refer to the number of members of an IXP as "ground truth" if the number comes from the authority (accounting unit at the IXP) and if it is sufficiently supported by documentation. We do not refer to the number as ground truth if we measured the number of members ourselves. For a measurement, ground truth information can help to verify results. However, measurements can also validate ground truth information. The more ground truth we have, the greater our confidence in our measurement results becomes. However, this kind of information can be of a high level, time bounded, and transported via unreliable channels. In the process of measurement data profiling, we always aim to compare data against the latest ground truth, and in doubt we have discussed data discrepancies with the person responsible.

4.1.3 Access

For reasons of ethics and contract compliance, namely privacy protection and confidentiality, we will briefly describe how we protected data from unauthorized view and copies. In particular, we focus on three aspects of preventing unauthorized access, which will be discussed later in the particular setups.

Physical Access (Zutritt)

Physical access means entering areas or spaces where IT infrastructure is operated or kept safe. Such areas are usually protected by access control, e. g., doors, keys, guards, and biometric systems. Depending on the type of facility, physical access control can have quasi-military standards, such as in big data centers, or it may be as weak as unsecured hot-spots mounted on a wall. The term "physical access" always relates to a person.

Logical Access (Zugang)

Logical access describes the fact and possible ways of using an IT infrastructure, i. e., being the user of a system. This also involves the different ways of reaching a system. Logical Access is usually granted to specific persons rather than to groups, and can be protected via secrets or biometric features, for example.

Administrative Access (Zugriff)

Administrative access refers to the application of permissions for information and IT infrastructure by a person. This involves permissions to access files, databases, and other kinds of stored data, as well as the permission to access computational resources. Special attention is to be paid to privileged administrative access, since it usually comes with the ability to extend permissions (administrative and logical, but also physical) of other entities or roles.

Access Techniques

We differentiate between two techniques of access, “immediate” and “remote”. Logical access describes the fact and possible ways of using an IT infrastructure, i. e., being the user of a system. This also involves the different ways of reaching a system. Logical Access is usually granted to specific persons rather than to groups, and can be protected via secrets or biometric features, for example. Immediate access requires physical access. In our use cases, this was inconvenient because measurement deployments are generally hosted in the partners’ domain and not located at our work places. The easiest method of access would be the work desk (if applicable) or the lab, which is located in the same building. In cases in which our cooperation partner granted limited physical access, e. g., the ISP deployment, physical access required us to travel up to several hundred kilometers. In other setups, such as the IXPs, we exclusively relied on remote access because equipment is set up in data centers with access restriction.

Immediate Access

Physical access to a remote site requires the approval of the cooperation partner. It means that only a well-known person can enter the facilities with manual and automatic access control. When a person is unavailable or staff changes, additional procedures are necessary in order to re-enable access. Due to non-disclosure contracts, we will not elaborate on the particular methods, but they involve personal contact as well as electronic and physical tokens. Inside facilities, moving around is limited, and work has to be completed within a limited time span, and in several cases even under supervision. Bringing in additional equipment or exchanging equipment was subject to advance notification.

In cases in which we have the option of even limited immediate access, we use the following:

- keyboard, video, mouse (KVM) switches
- serial console direct links
- laptops
- tokens stored with the cooperation partners
- tokens brought along for accessing disks, servers, racks, and facilities

Even in situations with immediate access, we use techniques of remote access for convenience and security. The main reason for using immediate access typically is maintenance. A particular difficulty is the lack of communication channels once one has entered the locations. Cellphones do not work inside most data centers and network communication to the outside (e. g., SSH, HTTP, and messengers) is usually disabled. This drastically increases the need for careful planning and makes immediate access even less attractive. It should only be used as a last resort.

Remote Access

The regular way of accessing devices and setups is remote access. We differentiate between operational access and maintenance. It is good practice to separate the roles here in order to prevent the confusion of permissions and tasks. Working remotely and in maintenance mode is highly sensitive to misconfiguration and even typing mistakes, e. g., accidental shutdowns of interfaces or machines.

Operational remote access is the way of granting users logical access and administrative access to measurement data. Whatever a user does in this mode, he cannot accidentally damage data or equipment, since permissions are granted on a personal or group level. The measures to protect data from damage include no-access policies, read-only access policies, and automated backups of the most important assets. However, data management (e. g., archiving, backup, and restore) is not subject to operational access.

Maintenance remote access is the part of work that deals with configuration management, data management, and user management. Only specific roles are allowed to perform these tasks, which are granted to trained staff only. Highly sensitive changes in maintenance require careful planning, including disaster recovery planning and the attention of four or more eyes. For example, in setups like the SamKnows [39] distributed measurement, several thousand devices can be affected by a single management operation. As a consequence, one configuration problem can disable a large subset of measurement devices, resulting in long-term damage to the project. This may require shipments, replacement, or physical manipulation of a large amount of devices, which is most likely infeasible.

Our most used remote access techniques are:

- Secure Shell (SSH) console and X-window remote access
- Intelligent Platform Management Interface (IPMI)
- Access via stepping stones

We rely on role separation and security mechanisms for the administrative remote access, such as different user names for a person in more than one role, standard access techniques, e. g., POSIX permissions, and local/network file systems that support it. In addition, we use tools (e. g., mollyguard[2]) for protection against accidental shutdowns and reboots. The most frequently used operational and maintenance remote access technique in all of our deployments is command line remote login via SSH. Only in cases that allow remote graphical devices in the cooperation contract may graphical front-ends for computational software (e. g., R and S-Plus) be exported. Disallowing graphical remote operation is mostly a result of security considerations.

[2] http://linuxg.net/what-is-molly-guard-and-how-to-install-it visited 10/2013

4.1.4 Technical Parameters

In order to establish a measurement deployment, the involved systems and their assembly have to meet several technical requirements. We differentiate between hard- and soft-limiting properties. Hard limiting factors are those that prevent the measurement in the first place, i. e., accessing or recording data in real time. Conversely, properties are soft-limiting if, when undersized, they impact the measurement performance but do not prevent data from being gathered.

Computing power

The three assets we have found to be the most critical are CPU, RAM, and BUS.

CPU: can be both a hard and/or a soft-limiting factor. In data acquisition systems, we always classify it as a hard limiting factor. The necessary amount of speed, cache, cores, and threads in a recording system significantly depends on the peripheral hardware that is used. As a rule of thumb, the more dedicated the peripheral hardware is, the less CPU power is required. Our measurement systems, scaling from a plug-computer to a cluster of multiple 10 Gb/s real-time network recording devices, turned out to scale differently, largely depending on CPU performance. While CPU was the limiting factor on embedded and unsophisticated devices, eventually preventing deployment, it was of no concern for the 10 Gb/s recording cluster because of specialized Endace high-speed network measurement components. When we verified Myricom and Intel 10 Gb/s line cards in order to replace Endace hardware, it was beneficial that we had not undersized CPU capabilities. Those Myricom and Intel line cards are less sophisticated and need more CPU operations for capturing. In some cases, lack of available CPU power could be solved by employing other specialized hardware, as was demonstrated by using GPUs in [84].

RAM: is the most crucial component for analysis. In the common case in which data must not be offloaded (or are too expensive to transfer) to other locations, they must be processed on-site. The statistical tools and other data-analyzing components can require a considerable amount of RAM depending on the data type and task. Capturing devices, at least in our scenarios, do not store much intermediate information for online processing.

BUS: Of the three elements discussed here, BUS is the most underestimated component when it comes to real-time measurement. To capture a 2x10 Gb/s link, a data rate of 2.5 GB/s is required, calling for at least a PCI-E 1.x X16 (4000 MB/s) or PCI-E 2.x X8 (4000 MB/s) or PCI3.x X4 (3938 MB/s). A note should be made here with regard to PCI-E: the length of a board slot to put in a PCI-E card indicates the maximum speed (X lanes), but in practice we have encountered cases in which this assumption was misleading and an X8-sized slot only supported an X4 data rate.

Storage

Measurement systems need storage capacity with capabilities tailored to the use case. In our work, the most important properties are a) *speed* and b) *scalability*. Speed refers to the net read and write operations that have to be supported exclusively or simultaneously and under the consideration of parallel operations (i/o threads), whereas scalability refers to the property of scaling out without performance penalties.

Internal Storage: The most important properties of internal storage that we have encountered are read and write speeds. The most critical application for disk speed is real-time recording. Especially when dealing with sophisticated high-speed packet-level trace measurements, no external storage has ever been able to match the write speed to the speed of capturing a uni-directional 10 Gb/s link, for which a disk write rate of more than $1 * 10/8\ Gb/s = 1.25\ GB/s$. With commodity systems and internal storage and using internal 3Ware-9690 RAID adapters on a PCI-E x8 port, we achieve 1.1 GB/s permanent write speed, which is not enough for a permanently fully utilized 10 Gb/s simplex transfer. In fact, for speed purposes we even use two identical hardware RAID controllers. Each of the controllers runs a single RAID6 configuration and both adapters are software-bundled as RAID0, resulting in a hard/soft-mixed RAID 60, which measures double the speed of each single RAID6. Although this performance is theoretically not sufficient to record a 10 Gb/s link, the operators of 10 Gb/s infrastructure do not utilize the infrastructure's full capacity because of the loss behavior at high utilization on such links. This means that our setup is sufficient. In order to understand whether we may run into recording speed problems, it is possible to compare the recorded data rate with the maximum measured write speed. Regarding full utilization, there are two approaches to further increase write speed of the RAIDs: (a) extending the RAID60 with one more RAID6 hardware controller, and integrating it into the RAID0 software RAID. Because the PCI-E x8 speed does not limit us here, the RAID can perform 150% of the current setup. The other possibility (b) is to extend the RAID configuration with an SSD for "hybrid mode". Although this cannot increase the permanent peak write rate, it suffices to buffer on the few occasions when the network speed exceeds the RAID60 recording speed.

However, hidden details are critical, such as scheduled verification runs at the RAID controllers, which hurt and slow down the RAID performance by 55%. Battery-assisted write back does help to increase performance on a single hardware RAID as well as the combined RAID60. Batteries, however, are vulnerable to heat and must be closely monitored in order not to turn the write speed advantage into an availability disadvantage.

Regarding RAID levels, in our experience hardware-supported RAID6 is more reliable than RAID5 (even with a spare disk) because disks actually *do* fail at the same time. The term "same time" also covers the time span that a raid-rebuild takes for recovering from a failed disk.

Between SAS and S-ATA, we chose S-ATA because of the lower price per volume and the fact that in such RAID configurations, throughput depends more on the bus speed (PCI-E) and the RAID controller's capability rather than on the native disk speed and features. The full-duplex advantage of SAS would only be seen if the disk system were used as a cache rather than storage, which we have avoided doing. The better alternative for caching storage is using dedicated SSDs that either run alone or support the RAID.

In practice, we aim to purchase fully equipped internal storage capacity with identical hard disk models. This makes it easy to adapt the file system parameters to the storage from the beginning because the number of disks does not change later, and in terms of hardware the disks behave identically. When exchanging disks, we always try to replace them with the same brand and

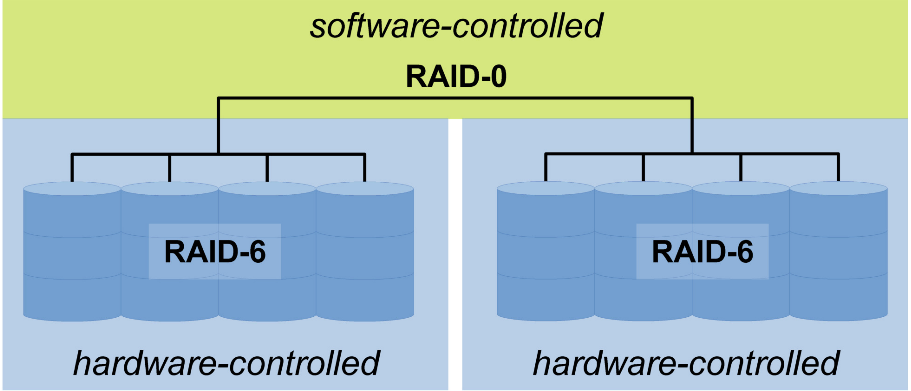


Figure 4.1: Mixed hardware-software RAID60: this level allows for both, higher rates than RAID6 as well as more safety against disk failure.

model. If this is not possible, we use a better (faster) performing model because the slowest disk always limits the speed of an RAID configuration.

In terms of speed, internal storage is clearly preferable over external storage. In terms of scalability, on the other hand, internal storage is limited to the number of disks that can be attached to a single computing device – in our case, the server. This makes the internal disk applicable for live recording and for computing data with the need for high-speed I/O.

Compression on internal disks has proven to be counter productive, unless it comes without a considerable data rate penalty. In addition, our data of full packet traces have a gzip compression rate of 1:1.5 with a compression speed 15MB/s, and a lzip compression rate of 1:1.6 and compression speed of 1.5MB/s, neither of which meets the capturing speed on our equipment, not even with parallel implementations. For gzip, the CPU-bound speed penalty towards 1.1GB/s write speed is of factor 88; for lzip, it is 82; and for parallel lzip implementation using 12 cores, it is 83.

Other caches, such as SSDs, can temporarily, significantly increase the I/O performance of a disk system, such as RAIDs. They are useful in two general cases: (a) if the data to process can be cached or prefetched for reading; and (b) if the cache can bridge for the spin-up time of the RAID on write. For (b), either the RAID's permanent write speed must match the write speed demand, or the cache has to be large enough to catch the whole volume of write demand at once. In our case of recording network traffic, this practically never happens, and we have not used such configurations in our deployments. For computing the data, on the other hand, caches might still be useful regarding I/O speed depending on the level of inter-operation with the RAID system.

When using internal hardware RAID equipment, it turned out to be an advantage to have battery support (e. g., additional battery packs) on the RAID controllers. However, too much heat kills them within days.

External Storage: In summary, external storage is more scalable, both upwards and outwards, but it can come with significantly less speed (the factor mostly depends on the bus technology and the transport medium). We have gained experience with fiber channel, USB2, USB3, and different

SCSI generations including iSCSI and SAS, and operated both hardware raid and software raid configurations – and even mixed ones. Our conclusion is that none of the technologies is capable of storing a single 10Gb/s traffic link in real time, but all of them exceed the speed of data processing. This property, combined with capacity scalability, is perfect for archiving data.

File Systems: First, we need to point out that there is more to performance than picking a particular file system. In general, we have found that the file system has to be configurable according to the RAID configuration that is used. We use the following two parameters to adapt the file system to the RAID configuration: a) the chunk size has to be equal to the chunk size of the RAID system; and b) the number of data disks has to match the number of net disks[3]. In case of the aforementioned mixed hard/soft RAID60 (Figure 4.1), it is essential that hardware RAID6 are configured identically, that the soft-RAID0 shares the same stripe size, and that the file systems' stripe sizes match. From the point of the file system, the number of data disks is the sum of the net disks of all RAID6s. Regarding file systems, we usually rely on XFS. It provides the required configuration capabilities and has run stably over years. However, one particular weakness of XFS is the limited number of files in a directory. While XFS allows for an arbitrary (more i-nodes are possible than bytes[4] are practically available) i-nodes in even a single directory, the practical usable limitation is about 100k files per directory before losing performance to a grade where disk operations stall. We had to adapt our recording procedure to create subdirectories with limited numbers of files.

Network

We generally differentiate between management networks and measurement networks. Both are isolated from each other to avoid interference, particularly measurement bias and management network overload. The latter easily and effectively locks out operators from control, leading to a scenario in which the site becomes inoperable until physical maintenance is done.

Management Network: Management network: We will mostly focus on the on-site local network in this section; the access network to a remote site is covered in Section 4.1.3. The dimensioning of the management network is less critical and typically requires significantly less attention in our setups. The main point to consider is the use case of copying/moving data between machines, which may eventually lead to an extra channel for data movement and control, such as a 10Gb/s transfer network and a 1Gb/s control network. For example, for large data transfers over network file systems, this leaves enough network resources to a command line interface of remote X session while fully utilizing the data channel. An example is provided in Figure 4.2 (black lines).

Monitoring Network: When a direct mapping of measurement devices to tapped lines is unaffordable or unmanageable, a more complex network can be beneficial to multiplex sources and destinations in the measurement data flow. For example, measuring N independent sources of network traffic does not necessarily require the deployment of the same amount of capturing devices. Instead, a multiplexing device can be used, such as a switch. Those devices have another

[3] For RAID5 the net disk number is $n-1$ of the real disk number and for RAID6 it is $n-2$ (not counting potential spare disks)

[4] that is: $2^{64}-1$ on 64-bit systems, $2^{32}-1$ on 32-bit systems

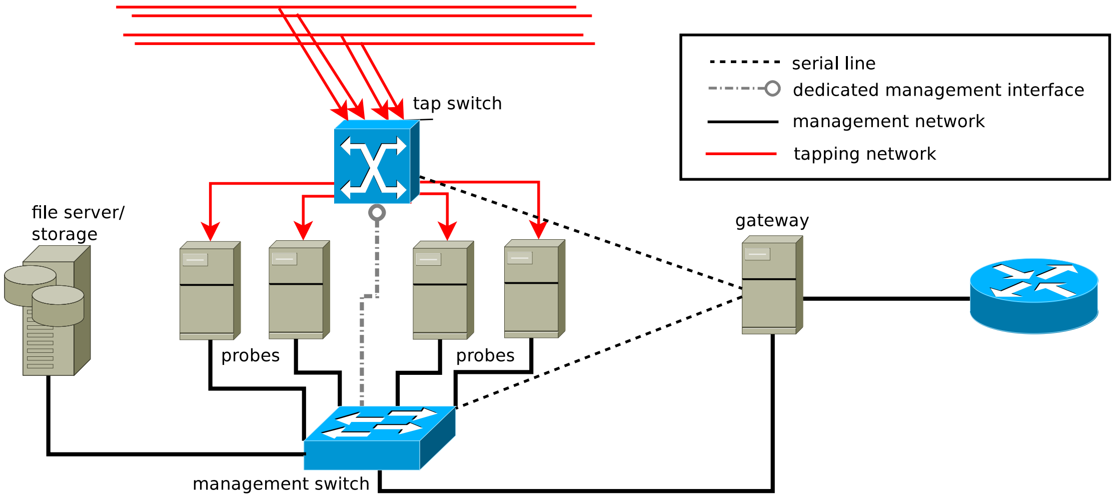


Figure 4.2: Simple example of separated management (black lines) and tapping (red lines) networks.

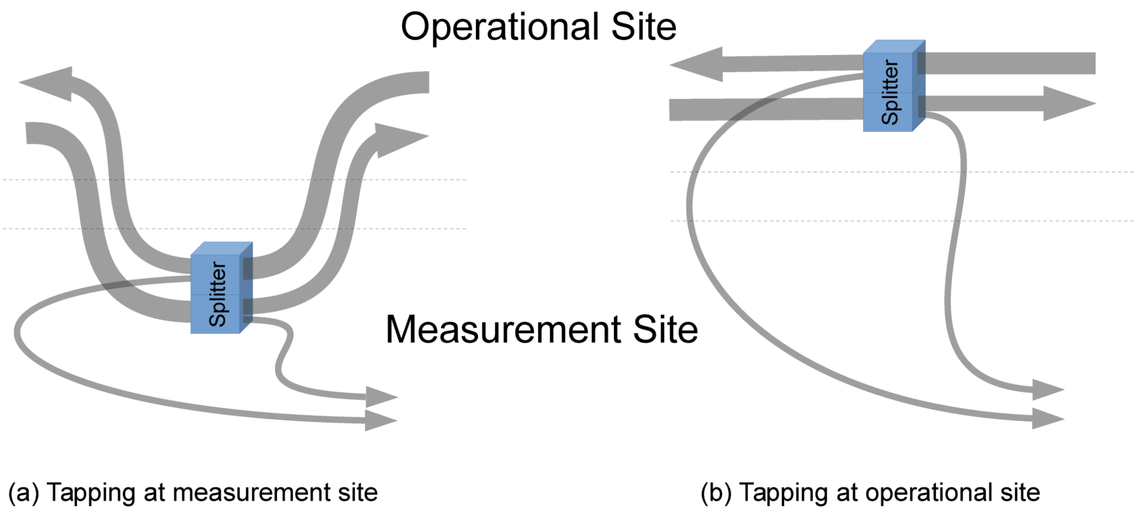
advantage, which is that they can aggregate lines. Thus, for the measuring of up to five channel-bonded 1 Gb/s full-duplex links, a single 10 Gb/s capturing interface is sufficient. However, such equipment has to be tested beforehand in order to obtain information on device-related packet loss, jitter, delay, and reordering. An example setup is given in Figure 4.2 (red lines).

Access Network: Remote vantage points and measurement setups can be protected from unauthorized access in different ways. One common practice among ISPs is to use a dedicated management network for maintenance that is made accessible to partners, e. g., vendors. However, such a network can be complex and would require access via secure private networks, for instance a company network. Therefore, in Figure 4.2 the access network is (invisible) behind a dedicated router.

Specialized Capturing Devices

Capturing “Internet traffic” can have multiple qualitative and quantitative dimensions. The low-end probing or low-speed capturing setups can generally be covered with commodity hardware[5]. On the other hand, high-speed and high-volume capturing use cases call for sophisticated equipment. Normal equipment is not sufficient for the following two reasons:
(a)capturing at maximum interface data rate already leads to packet loss; and
(b)time stamping is inaccurate and added after preprocessing (e. g., forwarding to the operation systems kernel).
The only hardware we have found capable of coping with our needs is Emulex Endace DAG cards, which provided the necessary features of accuracy and reliability.

[5] This is not the case for low power – low space scenarios.



(a) Tapping at measurement site

(b) Tapping at operational site

Figure 4.3: Different ways to tap an optical line; setup (b) is preferable over (a) because it prevents mechanical interference on operational links (*bold*) when touching the monitoring links (*thin*).

Other devices we use in our setup for non-invasive measurements are optical tapping devices and 10G Ethernet switches for collection and distribution of network line traffic. In fact, both devices are minimally invasive, as discussed below.

Optical splitters: Optical splitters: Tapping an optical line (such as any operational network traffic) should be done out-of-band. The advantage of doing this is that in non-invasive setups, a broken measurement setup does not affect the operation, reliability, and performance of the measured object, e. g., a line. Optical splitters forward a fraction of the signal energy (light) on an additional link without a feedback channel. Each fiber of a bidirectional link has to be tapped independently. The operation is illustrated in Figure 4.3. Besides optical splitters, copper-based signal splitters for electrical signal transmission are available, but we do not cover them here. Operational networks are sensitive to interruption, since user traffic is affected immediately and directly. It is therefore good practice therefore to implement the splitter outside of the measurement processing equipment. In this way, the risk of accidentally damaging an operational link during maintenance work is minimized. The setup is illustrated in Figure 4.3 b). However, in the case of optical splitters, the original signal is affected in its strength. Before deploying the (intrusive) optical splitter, the operator must know about the expected signal attenuation. Splitters do not necessarily split the operational and branched signal strengths 50% to 50%; 70% to 30%, 90% to 10%, and 99% to 1% are also common. The signal attenuation prevents us from putting too many of those devices in line without losing the signal in both the operational line path and the branched path. The signal attenuation caused by splitters among a path can be calculated as follows: remaining strength $= \prod_{R_x}$ with "R_x = fraction of energy forwarded along the path at element x"

Monitoring ports: High-speed Ethernet switches are appropriate devices for multiplexing sources and destinations of line tapping. All signals from any switch interface are either incoming-only from optical splitters or outgoing-only to network traffic capturing cards. This makes a switch-port expensive because the r/x modules are used exclusively either as a receiving or trans-mitting device. However, the devices we operate cannot combine both roles with the "*monitor*" feature. A positive side effect of using a switch is signal amplification. A sufficiently strong signal as input will exit the switch at full strength according to the specs of the fiber medium, which in turn allows it to be split by passive optical devices and to be sent to more than one device. In our case, the switch typically does not support multiple monitoring destinations (ports) for a signal. If the equipment allows for more than one monitoring session, we have to test whether the backplane capacity is sufficient to operate without interference due to limited capacity. A switch may introduce measurement bias, namely delay, jitter, reordering, and loss.

Supply

One critical (but often overlooked) aspect when deploying measurement equipment is supply. Before or while planning, the following issues have to be addressed: electrical power, cooling, (rack) space, and physical access.

Electrical power: Measurement devices, computational devices, storage devices, network equipment, and auxiliary (i. e., monitors or KVM switches) need electrical energy. Its consumption needs to be planned so that a sufficient amount of well-dimensioned and well-placed supplies is reserved. With the term "well-dimensioned", we refer to the fact that industry-grade universal power supplies (UPS) may require a 380V input. In fact, a large fraction of devices have more than one power supply input. In those cases, the amount of critical (i. e., storage) power interfaces should be configured in a way that independent sets of power supplies can each saturate the electrical energy consumption. For example, given a server system with two power supply units, each of them must be able to support the system's power consumption alone. There are two reasons for this: (a) a power supply unit can fail; and (b) the network source of one of the power supply units may fail. If feasible, the supplies should have different sources of electrical power, i. e., one directly connected to the power supply, and the other via a UPS device. In case of power failure the UPS takes over, and in case of a failing UPS the regular power line still delivers. In cases of UPS support, the UPS capacity is a limiting factor for connecting devices.

Cooling: Electrical systems produce heat, and even single-system components can overthrow heat management within a device. The consequence is damaged parts (or whole systems) over a short or mid-term period. Parts that are particularly sensitive to heat are batteries, disk drives, and expensive capturing equipment. In our systems, most heat comes from CPU and disk drives. In systems with extensive GPU utilization, this is also a potent heat source. There are three methods of cooling:

1. Transport the heat off the system
2. Transport the heat off the rack; and
3. Eliminate the heat in the location

1) Transporting the heat off the system is a question of systems design. Vendors equip their devices with sufficient cooling for that purpose. If the system is used unmodified, the dimensioning suffices (or is subject to warranty). In cases in which additional components are added to the device, cooling is an item for consideration. We consider that two components are important: the cooling power (measured in Joule) and the fit into the cooling strategy, i. e., air stream or fluid stream. Even a heat-neutral additional item can cause devastating harm to a system if it blocks the heat management (air stream). Not all cooling components are located within a closed system, however. Heat management also relies on sufficiently dimensioned cooling input and output channels that have to be kept unblocked, e. g., by cable trees.

2) Transporting heat off the rack is critical to supply enough cooling to the system. Our racks either have to provide an open design, such as porous containment, or implement a dedicated heat management system that is connected to the data center's cooling system. The latter is common practice among the big data centers. When putting equipment in remote sites, coping with the heat depends on the experience and capability of the hoster.

3) Transporting heat out of the location is the last step of the cooling chain. We do not go into detail regarding sustainable cooling and power reuse here. However, as with the previous point, dealing with the heat management of a room or a data center requires planning, monitoring, and experience from the hoster.

What is always required from any supplier of equipment by a professionally maintained location is a quantification of waste heat. Device vendors usually account for it in the documentation; however, as we pointed out in (1), assembly of additional components can be critical.

In less professional environments, e. g., a user's home, where equipment is deployed, for people's safety heat monitoring and precautions should never be left to the location owner's responsibility. Monitoring and precaution measures (switching off equipment) should be included.

(Rack) space: The physical dimensions of (sets of) devices may easily require more space than the obvious volume (e. g., in a rack) for the following reasons:

- **Cooling:** In order to provide space for air streams, devices may require mounting gaps in some directions.
- **Mount space:** Because of particular mounting systems, devices can require more space in some directions. For example, when using fiber patch panels, they cannot be located arbitrarily close to a wall or door because the fiber strings cannot be bent more than a defined radius.
- **Maintenance space:** Access to devices can be necessary for maintenance. Not all disk systems are front loaders, and not all mounting equipment (e. g., slides) allows the entire device to be completely pulled out of a rack.
- **Auxiliary devices:** Even deployments that are not supposed to evolve over time can require more space over time. Examples we have encountered are additional patch panels, additional switches (to account for more tapped lines), and equipment as simple as a temporary keyboard/monitor deployment.

Good practice in all of those cases is to run the equipment in a test deployment before shipping it to the remote site.

4.1.5 Documentation

Documentation has several dimensions; this section covers those dimensions. We differentiate *between deployment documentation*, *operation documentation*, and *measurement documentation*. The reasons for splitting this information relate to confidentiality and a clear separation of tasks.

Deployment Documentation

Measurements are the result of a process that is usually known to only a small number of people. Even fewer (or completely different) people may know the equipment and components. As a consequence, this kind of information must be documented. It may be documented in templates, since equipment for a site does not change frequently. However, the configuration of equipment can change, e. g., software versions or firmware versions that are adopted for a particular use case.

Physical locations involve geographical coordination of deployments and operation sites, street addresses, building navigation, room labeling, room locations, rack placements, and rack plans.

Logical locations are soft coordinates, such as network address space, interface addresses, file systems, file locations, and service location, e. g., DNS or Lightweight Directory Access Protocol (LDAP) servers.

Equipment is documented with regard to its purpose and required connectivity (e. g., management network components, measurement network components, monitors, keyboards, serial lines, usb devices, and storage devices). Moreover, equipment documentation includes auxiliary devices, vendor manuals, and relevant device configuration details.

Operational Documentation

This level of documentation is dedicated for maintainers of measurement setup who deal with changes to this setup. Operational documentation targets the “admins” and in particular it includes confidential details (if applicable), such as credential management, user management, authentication system, or configuration options. Operational documentation may reference or cover the deployment documentation for clarification regarding specifics of a particular measurement data set.

Configuration Options

The system configuration in the operational documentation includes the following.
Measurement setups describe different modes of operation for conducting different types of measurements.
Device options describe the modes of operation of dedicated devices that are included in the measurement, e. g., in our case, Endace cards, switches, and network bridges.
Default settings are applied when the user does not provide a specific parameter for a setting. Some of those default settings do have an impact on the measurement and device operation. Default settings are common for hardware and software configurations. The documentation must describe default settings and behavior and, if applicable, their impact on the different operation modes.
Software repositories have to be documented in terms of location and usage, as far as they relate to the operation of the systems. Our repositories even cover different versions of software for particular types of measurements.

Measurement Documentation

This level is best characterized as users' documentation. It applies to the level of daily operation and is accessible to all users, e. g., people working with the measurement data. The key is a data dictionary that holds information regarding all measured and stored data sets, e. g., meta-information about expiration dates and history of data processing. It may refer to deployment documentation for clarification on the setting of a particular measurement data set.

Documenting all equipment is not sufficient. Even in simple setups, a single piece of equipment or any combination of components can bias the measurement results. For example, consider a scenario in where there are two optical links of the same line but in opposite directions. A switch can monitor the aggregate of the links and as such replay the network traffic of both directions on a single interface where the measurement equipment is connected. Depending on the hardware and software of the switch, packets might be reordered from the switch. Thus, there is bias that can only be taken into account if the details of the setup are known.

Related Traces

Network traces can consist of more than a single chunk. Examples include:

- multiple lines that are part of a trace, including multiple directions of traffic, e. g., for optical lines
- data from multiple, possibly related locations in timely or locally distributed measurements

First, it has to be clear which traces are part of a measurement. For instance, port trunking[6] makes it possible to split traffic on layer 2 over multiple physical links between switches. To measure effects on the physical paths (different delay or physical signal loss), it is inevitable

[6] also referred to as "link aggregation" or "channel bonding"

to measure and analyze all involved parts and directions in the first place. In order to reliably correlate those multiple parts, we need the information on whether (and how) synchronization is done. For analyses involving both directions of a network link, it must be known how much jitter is introduced by the measurement and what the timing offsets are. Using dedicated measurement equipment, e. g., Endace cards, is only one part to the story. The other part is knowing how fast clocks skew and how frequently they are re-synchronized. Using off-the-shelf equipment and NTP time synchronization can render a trace useless for timing analysis. Therefore, this synchronization information is critical.

Situation (Occasion)

Measurements happen either non-recurrently, randomly, or on a regular basis. However, in many cases, “real world” events trigger a measurement in order to observe network conditions, such as IPTV broadcasts of major sports events or the launch of an entertainment system. The occasion is not always clearly known in advance. It has become common practice to measure constantly and keep traces only on the basis of reported network events. Such events include network attacks, denial of service, and service degradation, and they are reported by users or detected by technical systems. By the time a problem is reported, the traffic will already be lost for a measurement and analysis. Documenting a reason for keeping a particular trace helps other investigators to understand what is in the trace data and how to analyze it.

Technical Parameters

Capturing involves many parameters that cannot reliably be detected from the trace itself. The two most common techniques are filtering and sampling.

Filtering can happen on a low level and on the basis of the traffic properties, i. e., only considering specific protocols (e. g., IP/TCP/HTTP) or communication entities (e. g., IP and MAC-address). However, traffic can also be filtered on a higher level, which requires preprocessing. Examples are “traffic of the first percentile regarding flow length”, “connections using HTTP 1.1 pipelining”, and “over the top (OTT) traffic”. Moreover, the criteria may not be known at the time of trace recording, but may instead be specified after some processing. One such more complex rule to apply filtering could be “only traffic of the 5% observed OTTs that are using HTTP pipelining and an inter-packet arrival time of less then 1 second at maximum”. However, any filtering is possible, and the result may mislead observers who not know about those filters.

Sampling can be property preserving or non-preserving, but it always comes at the cost of information loss. If sampling is applied, the central problem is deciding which kind of information loss is acceptable. This question relates not only to the data, but first and foremost to the intended analyses. Thus, by choosing a sampling methodology, we limit and pre-select the kind of analysis that can be applied to the trace.

For example, in a packet based network:

- Sampling on a low-level packet selection may disallow connection analysis of the trace.

- Sampling on a flow basis prevents us from obtaining inter-packet arrival time.
- Sampling by collecting only headers prevents payload inspection.

Typically, the trace data themselves do not include the sampling methods and parameters that have been applied, and this information is not deductible from the trace. Documenting the sampling methodology prevents us from drawing wrong conclusions when analyzing the trace. Popular techniques of sampling are packet sub-sampling, in which only a subset of packets are recorded, and header trace sampling, in which the recorded volume of the data is restricted. Header traces can be tailored by a fixed snap length, or by restricting the recorded volume length to a predefined set of protocol headers, which is more complicated and more useful.

Responsible Personnel

Our measurements have technical components regarding which the person who executed the trace can explain details or decisions that are not documented. However, there is also a legal component, for instance regarding who has possibly violated regulations, e. g., anonymization. For legal reasons and technical reasons, it is necessary, or at least advisable, to know who recorded network data.

Original purpose

Although the original purpose is not necessarily important for future use, it usually provides information about the parameters that are used for capturing, and puts a measurement into context. It also links the measurement to events that are not necessarily obvious, such as “Opening and online broadcast of the Olympic games”. Such a description can help to compare traces.

Text

We recommend giving as much context as possible. In our study, completing a static template would limit us in providing useful information regarding a trace or other measurement. Free text helps to document irregular but particularly interesting information. It supports us in understanding aspects such as the following:

- the reasons for the choice of a particular of preprocessing;
- additional legal requirements;
- the project context;
- the topics targeted; and
- the tools used or modified.

Expiration Date

For technical reasons (e. g., disk space) as well as legal reasons, traces have an expiration date. For proper data management, expired traces must be deleted for legal reasons, or should be archived for data management reasons. Keeping track of this information allows for proper data management. For example, we can prioritize analyses based on the expiration date of the trace and the expected complexity of the applied analysis.

Modifications

As previously discussed, traces can be modified before or after they are stored in different ways. All modifications are to be documented alongside with the information regarding when, how, and why a trace was altered, and from which trace it originates.

Keywords

A list of keywords can help to automatically index and categorize large amounts of data sets over time and make knowledge accessible to subsequent projects.

Documentation Formats

In general, it is useful to have a format that is both human-and machine-readable in order to manage and understand the data. However, this depends on the audience and the documentation type. Moreover, such formats (e. g., XML) are not suitable for all of the described aspects, and as such we recommend a structured format for information on *date/time*, *duration*, *expiration date*, *person*, and *keywords*. For the other parameters described above, the most obvious technique should be used in a consistent manner (i. e., a consistent data format and structure across traces). Documentation comes in various formats, such as structured and unstructured text, diagrams, schematics, configuration snippets, programming code, vendor information, user manuals, databases, and photos.

Templates

At first glance, documentation on the level of single measurements seems to require huge overhead. Since our measurement setups are mostly static, good practice is to provide a template with the information that can be used for all traces from a deployment. Relying on a static description that is part of a global documentation is insufficient. Even if this information is kept up to date and changed over time, it will be invalid for retrograde measurements. Consider different machines for measurements on the same site for the same medium. In this example, scripts help to gather information on the capturing system for the following parameters: machine name, machine equipment (measurement cards), equipment IDs, and firmware versions of specialized equipment, e. g., in order to mark traces that have been captured with invalid hardware or configuration.

Meta Information Retrieval

Several types of information can be gathered automatically at the time of measurement. Examples are CPU and memory consumption of the involved processes and the whole system or I/O load of disks and interrupt statistics of the system in order to detect problem sources that are not visible from the trace itself, i. e., packet loss due to an overly high load of a critical component of the measurement system. This information can be collected locally, i. e., on measurement servers, but sometimes it has to be pulled from other measurement components, (e. g., tap switches via SNMP) or other control mechanisms that do not interfere with the equipment operation. For instance, pulling a switch configuration and statistics during the time of a measurement increases the confidence in the measurement data.

As extensive as measurement documentation may already appear, it is not necessarily finished once a measurement is completed and validated. In order to avoid repetition of a particular analysis, analysis information can be linked to the trace information.

In terms of volume, raw packet traces are highly expensive compared to other data representation: the volume of full packet traces is larger than the traffic volume itself. When storing packet traces, additional space is required due to the storage format. At the ISP, an additional 2.1% of space is used for full packet traces with the ERF format, and 2.6% is used with the PCAP format. Both trace recording formats contain supplementary information, e. g., timestamps. The numbers vary with recording snap length and packet size distribution.

4.1.6 Monitoring and Alerting

Monitoring the state and operation of equipment is relevant for operational, management, and measurement issues. For operational and management tasks, monitoring serves the purpose of safety, equipment protection, reliability, and security, while for measurements the confidence in data is the major aspect. Alerting to a system on operational problems can result in automated action to change the state of a system if the operators predicted that state.

Monitoring Parameters

Sensors

System state is monitored by using its sensors, which can be either physical sensors in hardware or logical sensors in software

Hardware sensors are common for the following use cases, and are mostly used to prevent physical damage:

- temperature of board, CPU, disk, and chassis
- power supply states
- power indicators (device is switched on/off)
- voltage of different components

- fan speeds
- S.M.A.R.T parameters of disks
- state indicators of RAID controllers
- chassis state indicators

While most of these parameters are only relevant in a live system, the power indicator is an exception. Indeed, most monitoring systems implement a hierarchical or sequential checking of parameters. The electrical power indicator is obviously a parameter with which to start. While this particular information can indirectly be derived from responsiveness of other components, e. g., accessibility, it is better if queried directly via IPMI and similar remote management facilities.

Software sensors are powerful for monitoring system utilization rather than monitoring hardware or environment. They are more flexible than hardware sensors due to their programmability. The most important parameters to monitor are:

- log files of the system and user land services
- CPU usage over time
- memory utilization for detecting resource shortage
- process accounting
- availability of local and remote services
- access events of users and processes
- network reachability of devices in the same or other network segments
- network utilization
- results of automatic procedures, such as cron jobs, etc.

Usage

Usage monitoring mostly relies on software sensors, and helps to understand how the resources of a system are utilized and for what purpose. Monitoring the usage of a system mostly has administrative motivations. For real-time measurements, the system usage provides hints regarding the root cause of bottlenecks during measurements, such as CPU shortage, network capacity shortcomings, or storage speed insufficiencies. This monitoring also helps to point out interference of processes, and can help to schedule tasks properly.

Security

Security monitoring works similarly to usage monitoring. Because our network measurement systems work on data that legally require privacy protection, it is the operator's responsibility to protect those data from unauthorized access. Knowledge of break-ins, take-overs, and other improper use of systems is essential for operators. Security monitoring is not restricted to attacks and unauthorized attempts from the outside, but also requires the monitoring of local user behavior and local processes to report malicious activities and attempts thereof. Security monitoring cannot be executed on the monitored system. Once the system is compromised, local security information is compromised too. A dedicated write-only log facility is one practical way to record and process sensitive log data.

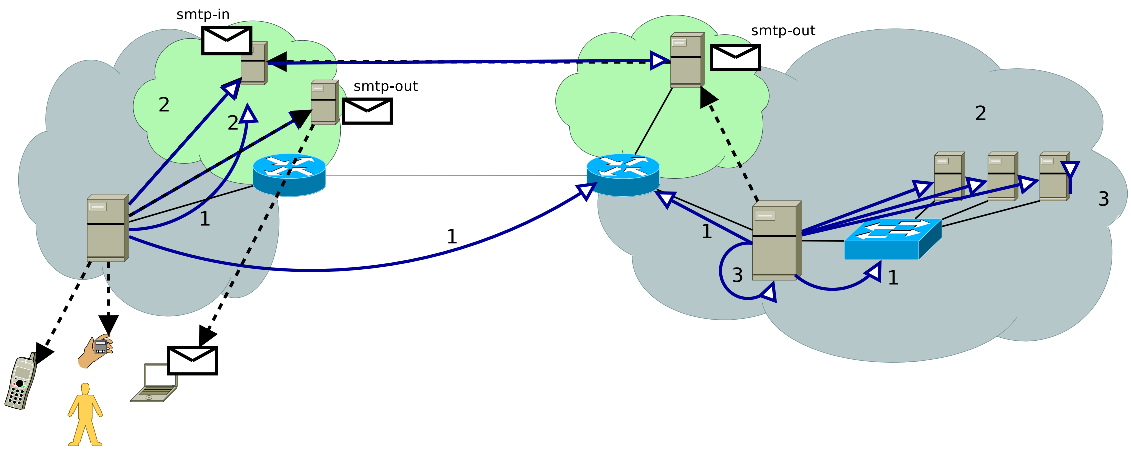


Figure 4.4: Schematics of remote monitoring with restricted signaling channels

Monitoring and Alerting Techniques

Automated monitoring deployment of infrastructure is not particularly difficult. Nevertheless, in our deployments we have encountered some special situations. We will elaborate on these situations in the following.

What is special about the setups? The security concept of our ISP and IXP deployments does not allow (or enable) active signaling from the measurement site to any external entity. This effectively disables any autonomous active alerting. For this reason, we rely on aggregation of information and on frequently pulling the monitoring state. There are still possibilities of signaling out that require additional cooperation from our partners, however. We describe them in Chapters 5 and 6.

Open source monitoring systems, such as Nagios [97] or Icinga, are extensible via plugins for arbitrary passive and active checks, and they provide mechanisms for iterative problem checking and reporting. For instance, if a network of N monitored machines is unreachable, it makes no sense to check (or report problems) for unreachable network services on those machines. The mechanism builds on the idea of dependency trees that are implemented in a rule-set configuration. As a result, information to operators is generated when actions are required according to the rule set. There are general limitations to machines monitoring themselves: if the information aggregator does not work, no message can be generated. There are two solutions to this issue. The first is to rely on a dead-man switch: the operator is notified if the system does not report a status at all. This requires a local monitor in addition to remote site monitors. A possible setup is given in Figure 4.4. The other solution is to implement a mutual monitoring, i. e., two or more systems monitor each other's parameters.

If the remote site is not directly reachable, we implement work-arounds. It must be noted that this has to be approved by the operators of the local and the remote site, and requires their active cooperation.

Pulling information with a dedicated account can be effective if the operators of intermediate systems agree to the existence of such non-personalized accounts. The account must not have any permissions beyond automatically connecting to the next hop. This can be achieved with static command lines per SSH-key on each of the stepping stones, which automatically attempt to establish a connection to the next hop. In this case, if the non-personalized account is compromised, no harm is done to the intermediate systems. The last step connects to the remote monitoring aggregator. It works with the same techniques, only the SSH command pulls status data from the monitor. This technique relies on frequent probing, however, which is generally not appreciated.

Signaling out over SMTP is an option if the remote site operator allows sending mail via local mail servers. The remote monitor is either provided with an SMTP account or white-listed. The remote site's mail system forwards the monitor's messages to a set of email accounts. Another monitor then collects those messages and decides whether the absence or content of a message is a reason for alerting.

Signaling out over one-way channels can be implemented, for instance by sending an SMS via dedicated devices in the remote site. Information can be sent out but not received on behalf of the monitor. Since most data centers are electrically shielded, however, this option is rarely applicable in practice.

Reporting to a remote monitor applies if the remote monitoring is restricted to forwarding information within its own network. A person then has to pick up the status from that network and signal out via telephone. This option is the least scalable because false alarms and testing always require attention from several persons.

Finally, **signaling out to a responsible person** may follow rules. Those rules may depend on the criticality of a monitored parameter as well as on time of day. Examples of systems to signal out are email notification, short messages (SMS), messaging protocols (XMPP), and smart phone monitoring apps. Convenient but much less reliable are customized login messages with status reports. They still provide an additional channel through which to pass information.

4.1.7 Synchronization

Services and devices need to be synchronized with regard to their configuration, but also with regard to time. Time synchronization matters for system operation and maintenance as well as for measurements. System configurations can have either a "hard" or "soft" synchronization. With "hard synchronization", we refer to items that are identical among sets of systems, such as LDAP configuration, network file system mount points, and login rules. They can be synchronized with rsync or git if the configuration mechanism is based on files. Soft synchronization, on the other hand, refers to cases in which the presence of similar configurations among machines is required, but some configuration settings are different for each machine. Examples are Kerberos [123], or SSH machine credentials and X.509 [76] cryptographic certificates. When configurations are different across machines but do not change over time, we do not refer to synchronization but to instantiation, e. g., in the case of host names.

Organizational Synchronization

The easiest, and by experience most unreliable, method is synchronization through personal communication among a group of people. This requires a case-by-case coordination of activities. The advantage of this is that it can work for planning activities in advance, e. g., for coordinating resource reservations. On the other hand, we have experienced that this requires a limited number of reasonably reliable people and does not scale, which is a disadvantage. Over time, teams change, and personal communication as well as team communication usually change too. Situations in which this method may be recommended (mostly because of simplicity) include:

- Situations in which the number of involved people (participants) is less than five.
- Situations in which participants do also have social contact.
- Situations in which the number of events to synchronize is small, and those events do not occur on a regular basis.
- Situations in which interests of all participants are aligned.

Technical Synchronization

The alternative to personal synchronization is the implementation of enforcing mechanisms and preventing interference via technical means. Examples of this are:

- Implementing locks during maintenance (updates, reconfiguration)
- Implementing locks during measurements (resource allocation of subsystems)

This synchronization requires a lock to be enforced and monitored. The advantage of this method is that it involves a clear and documented rule set, whereas the disadvantage is overhead. Technical enforcement can be used only for the most common and simple cases, while interdependencies are typically difficult to overlook and implement.

4.1.8 Planning and Future Scalability

Successful deployments of measurement systems tend to grow in physical and longitudinal dimension once their results are beneficial. When planning a deployment, it is helpful to develop strategies for project extension and termination. In particular, the latter may happen not only due to insufficient results, but also due to legal changes and changes in an organization.

Proof of Concept

In the first phase, the system needs to pass tests regarding the right choice of technology and the validation of operational decisions. This depends on the type of measurement and equipment, but the general principle follows these steps:

1. **Lab tests:** All equipment is tested in a lab or other controlled environment. It undergoes stress tests and load scenarios for typical and extreme loads in order to test primary and secondary operational parameters. Primary parameters ensure the proper operation for the deployments' use cases, i. e., validity of measurement data and robustness in high-load scenarios. The secondary parameters are operational prerequisites, such as temperature thresholds, component stability, power consumption, and physical robustness. A verification of required certifications for operation (e. g., TÜV-GS or FCC compliance) is part of this step.
2. **Test deployment:** This is the isolated deployment in the targeted environment. It includes testing local and remote access and environment verification, e. g., electrical power and cooling capacities. In this step, the deployment should already be operated by the employees responsible for operating it on a regular basis. This is usually not the same staff members who configured and tested the system. Therefore, joint deployment at this early stage can help to verify deployment manuals.
3. **Multi-instance test deployment:** When relying on multiple measurement devices for the same purpose, validity and comparability of devices are tested in this step. Having multiple identical devices performing the same network measurement on the same data and at the same time will show consistency in real-world deployment. Cases of overloading and the effects of unexpected events can be detected at this point. In addition, maintenance procedures are tested here.
4. **Distributed test deployment:** For distributed measurements, this is the time to test the deployment. In particular, this helps to verify deployment instructions and includes last-minute management and reliability tests. As different problems will occur at different deployment instances later on, staff members have to be trained for trouble-shooting and user dialogue.

Planning for Success

The primary goals are reliability and scalability along technical and organizational dimensions. Success is followed by demand: when deployments grow, the original hardware (disk space, tapping points) might not suffice anymore. Individual components may have to be upgraded (scaling up), and additional components and instances may have to be deployed (scaling out). For scaling up, the existing systems must have reasonable expansion capacity, e. g., card slots and memory slots, while for scaling out, more space, energy, and connectivity are required. When scaling up the user base, equipment has to be extended in places where physical access is limited, e. g., work spaces. Scaling out, on the other hand, may require new methods and capacity planning for maintenance, troubleshooting, and operation. Not only primary systems, such as data collectors, but also secondary systems, such as data processors and communication channels, have to be considered.

Extension Planning

The dimensions of growth we have experienced during our study include the following.

Physical space covers deployment space of devices, access space, and auxiliary equipment. It can also require storage for archive data, such as tapes, disks, printed documentation, manuals, parts,

and tools.
Logical space mostly covers mostly space for data storage on disk or archive media.
Computational resources cover computational requirements (e. g., CPU, RAM, and storage capacity).
Automation is required as setups grow and scale out. This includes deployments, operation, and maintenance.
User number and demands have increased with participants in the project. Our users prefer different tools and resources in order to work efficiently. Changes to software and physical work environments were the most common demands in our experience.

We conclude that it is beneficial to avoid complicated and manual setup, maintenance, and operation of the systems. Particularly if external partners are involved, we make them aware of resource limitations and scalability issues. Thus, they may reserve space and other resources in advance for future use.

Major Extensions

The largest extension we experienced was the cloning of a complex measurement site to an additional physical location. The setup involved hardware planning, procurement, access issues, and everything that we had already done for the first location. Having a complete documentation on how the site works enabled us to offload work and also helped to avoid known problems. Cloning complex and large setups is certainly rare and time-consuming. Still, we estimate that we reduced the setup time from 1.5 months for the initial setup of the measurement to about 7 days for the clone.

Another type of extension is the growth of an existing setup by including more machines and additional network resources. The extension can for instance be initiated from outside by a partner who changes access network design and access technologies. New types of machines have to be deployed, and other systems need to be scaled up, scaled out, or moved. In all cases, the effort is considerable and cannot be planned and accounted for in a regular maintenance schedule.

Minor Extensions

Unlike the large effort that is required for major extensions, minor extensions can be resolved within regular maintenance intervals. Examples are components, such as hard disk and RAM upgrades, new software, and updates. The extension of tapping points is another case that does not take much time or effort. The convenience of adding tapping point capability on demand motivated us to change the design of our setup. Of course, the minor extensions could not involve touching any single point of failure.

Planning for Failure

Planning for failure is critical to avoid wasting resources. Particularly with highly sophisticated equipment, which is expensive and difficult to obtain, project failure should not lead to a loss of resources and investment. The main goal here is to enable re-dedication. In cases in which re-dedication seems impossible (e. g., because hardware has to be tailored), awareness prevents spending on the wrong equipment and resources. It also leads to a more cautious planning. While over-sophistication is one end of the scale, under-dimensioning can be another reason for preventing re-dedication. Server machines, for instance, can be re-dedicated only if they are sufficiently powerful for general purpose tasks, e. g., file servers and Web servers.

4.1.9 Measurement Operation

In this section, we cover the methods and tools that we have used to capture and process different types of network data. We keep the descriptions close to our use deployments of ISP, IXP, and active measurements. The specifics of those cases are described in dedicated sections below, but we first focus on general and common aspects and techniques.

Computing Machinery

Concurrent and conflicting resource demands on a measurement system can bias the measurement in unforeseen, non-reproducible, and non-detectable ways. To minimize interference of activities on measurement systems, we recommend splitting the role of systems into capturing and processing devices. The separation can be done in the following ways.

Single-purpose hardware dedication means that hardware is exclusively dedicated to a particular role or task. This mostly applies to equipment that is specialized in collecting traces, managing storage capacity, or processing data.

Time slicing is the easiest technical method to dedicate resources to isolated tasks.

Configuration changes can range from enabling or disabling features by software or OS reconfiguration, to image-based booting. In the first case, it is sufficient to change or replace configuration files and toggle kernel feature support. This is limited to single machines and requires configuration management.

The latter case of image-based role management helps to quickly re-configure larger parts of the setup to serve particular requirements spanning multiple systems. For example, one can reuse a PC platform with dedicated measurement hardware but rare measurement occasions for computing in a distributed computing environment. One boot image would start the device with a kernel that supports measurement hardware; the measurement can then be executed and the data saved to external storage. On a reboot with a different system image, the machine would join the cluster of machines to process this data. We have had excellent experiences with such systems in our "Routerlab" [121].

Capturing Machinery

When we use sophisticated hardware for dedicated tasks, e. g., high-speed network packet capturing, the most important parameter is stability of the operational parameters throughout each measurement as well as throughout consecutive measurements. This is one particular reason why virtualized measurements are unreliable and strongly discouraged, especially when one considers time as a measurement dimension. The following rules of thumb help to design and operate a stable and reliable capturing system:

- Use native systems, i. e., do not virtualize.
- Use real-time capable systems, i. e., systems that are dimensioned to take loads higher than the expected peak utilization.
- Force independence from external and auxiliary hardware, e. g., avoid shared storage for live recording.
- Avoid background jobs, e. g., cron, RAID checks, and other services.
- Prepare for failure, i. e., account for remote system and data recovery management.

For data acquisition, we differentiate between the following three modes.
Plain Dumping writes the output from a capturing process to a storage medium for later analysis. The processing overhead is low, but the sensitivity of the data may not allow for this mode.
Pre-computation manipulates the captured data regarding pseudonymization, anonymization, or aggregation before storing it. If possible, the processing should be executed on specialized equipment that can compute at line speed. Data manipulation should be executed in a consistent way, with techniques such as prefix preserving pseudonymization.
Inline computation processes the data as they come from the capturing source and only stores the processed results. This is possible for lightweight operations, such as simple counting. In our work, the computational complexity often prevented us from using this mode.

Depending on the type of measurement, parallelization is possible. In our passive measurements, we identified only a few opportunities to take advantage of multiple CPU cores. It was mostly the system bus or the type of measurement (stream capturing) that prevented us from using parallelization. In active measurements, multiple CPU cores can speed up the process if the measurement method adapts to it, for instance by synchronized target selection.

Processing Machinery

In non-real-time processing systems, restrictions are more relaxed than in real-time capturing systems. This allows for virtualization, instantiation, clustering, parallel computing, and resource pooling. One restriction is to avoid interactions with the capturing machinery, such as external storage sharing. The usage of external resources, such as cloud computing hosters, is strictly forbidden for data processing machinery due to legal, confidentiality, and privacy reasons.

Single devices for storage and processing are the most predictable, straight forward, and easy-to-setup systems for processing. However, once the machinery fails, the raw data and results may be lost.
Clusters for storage and processing can be fixed or flexible in terms of membership. They

require more resource management planning in the first place, but usually pay off as equipment fails, since they come with integrated fail-over and fault tolerance mechanisms.
Parallel processing solutions, such as “private clouds”, are the most scalable way of assigning computation resources, but also require additional attention to reassigning instances to different tasks because of privacy and confidentiality issues. In addition, such environments require trained users for parallel and distributed data processing.

Measurement Capabilities

In our measurements, we mostly rely on the following types of data acquisition across different vantage points:

- **Link dumps** to record full or size-limited fractions of frames or packets from a medium.
- **Flow data** containing aggregated or highly sampled information on a link basis.
- **Feeds of formatted data** containing fragments of data streams or data collections.
- **Snapshots** containing the data constituting the state of a system.

Most of the tools we used are described in the background section of this thesis.

Measurement Techniques – best practices

Full Packet Trace Analysis

We collect full network packet traces in two fairly challenging environments:

1. 10 Gb/s high-speed capture on server-grade equipment with specialized monitoring hardware
2. 1 Gb/s packet capture on limited hardware (Asus WL500g)

In both cases, we have found that it is infeasible to perform complex inline data analysis during packet capturing. This shifts the problem to lightweight pre-computing and storing data to disk for later analysis. The pre-computation in the high-speed setup is implemented as a shared library for pseudonymization of IP addresses. Other features, such as limiting the snap length or demultiplexing the network stream over multiple channels, are implemented in the driver of the Endace DAG cards. In the scenario of limited hardware resources, we cannot afford to capture 1 Gb/s network streams. Not even simple dumping is feasible due to limits in write speed capabilities to any storage medium. In this case, the pre-computation has to be lightweight and must reduce data volume significantly. In our use case, we had to compare network line data on two links, which could be tapped with a single device. We reduced the data to an amount (and level of completeness) that was sufficient yet relevant for analysis. In this case, we only stored duplicates or uniquely seen traffic based on hashing, and used volume reduction by limiting or re-representing parts of the traffic.

Because of the infeasible complexity of inline analysis, we distinguish between capturing tools on the one hand and processing tools on the other hand.

Capturing tools

- The *fishing-gear* tools and their library are useful for Endace DAG-based high-speed network traffic capturing. They not only enable plain and pre-processed dumping to disk, but also provide tools for handling the ERF file format, which is the native output of Endace DAG devices, including conversion to the more common pcap format.
- The *diffdump* library and tools are suitable for use on embedded devices and provide capturing and data reduction preprocessing methods.
- *SFLOW* information arrives in the form of messages, each containing a set of samples. For SFLOW, unreliable UDP transport is used, but the data include sequence numbers and device id. This enables one to determine data completeness. Instead of processing the data on arrival, we store it in pcap format for multiple reasons: a) less processing overhead, and thus less resource competition with the dumping; b) we do not want to rely on the included timestamps entirely, but want to be able to verify this information with system timestamps that come with pcap; and c) we want to conduct a computationally expensive in-depth analysis later on. This process consumes disk space at rapid speed. To enable efficient resource use, we synchronously rotate and asynchronously compress the pcap data.

To date, the most versatile tool we have found for looking at full packet network traffic traces is the Bro-IDS [129]. A description of Bro is given in Section 2.1.

Packet Header Trace Analysis

In contrast to full packet traces, headers are sufficient in situations in which flow traffic analysis is sufficient and protocol statistics do not require deep packet inspection. The information loss at limited snap length, e. g., 128 bytes, reduces the data volume by a factor of 6 in case of medium packet sizes of 700 bytes. A limit of 128 bytes is SFLOW's default sampling size, and it suffices for capturing all protocol headers, including long header sizes, such as IPv6. The factor may not seem large, but it accounts for the difference between a day and a week of data recording.

Long-Term Analysis

Once traces are analyzed and do not have to (or must not) be stored, we store aggregated results. This typically requires much less storage space than the original full packet or header traces. As a consequence, we can only compare current measurements to those aggregated data. This makes it necessary to document information about of the original capturing procedure and processing methodology.

Synchronization

When one process on a single machine is not sufficient for the measurement, processes must be synchronized throughout systems or subsystems. Some hardware already provides high-precision synchronization features, e. g., the Endace DAG cards that support nano-second precision time synchronization among devices. Synchronization is used for parallelization of activities, such as simultaneous measurement using multiple devices, but also for serialized coordination of activity. One example of this is the resource sharing of disk space throughout recording systems. When resources in a measurement system run short during a data recording session, other systems may take over. In this case, the reference metric (e. g., time) does not only have to be synchronized, but the hand-over procedure also needs synchronization. We use different types of signaling, as will be discussed below.

Network packets: Broadcast, multicast, and unicast network packet events are a flexible way of signaling to a set of connected systems. One requirement on the receiver side is a process that interprets the packet and triggers an appropriate action. Since network traffic processing is a fast and nearly real-time operation in current Unix/Linux operation systems, the resolution and accuracy are better than network time. In addition, this works asynchronously and on an event basis. It enables event-driven signaling, e. g., "disk full" messages, and accordingly automatic fail-over.

Timers: are a tool for recurring actions, such as file rotation or system checks. If systems that do not share a communication channel require synchronization, system timers become convenient. However, they depend on local time information, and thus can be quite inaccurate. In the case of a large deployment of home measurement devices (which cannot be signaled actively), however, timers are our only option for synchronization.

Parallel command line tools (e. g., pssh) can manually synchronize activity among systems. This can be due to either timing motivation or convenience reasons. The convenience lays in the flexibility: it is even possible to maintain systems with this technique by deploying identical configuration files or installing software.

Combining different synchronization methods is possible, for instance by using the network time protocol (NTP).

Pipelining

During our work, we have come across situations in which the recording data throughput has been small enough to store on limited capacity local storage, but at the same time too big to offload to larger external storage. In such a case, synchronization and pipelining help to store a large trace in the external storage. The key is for systems to synchronize the measurement in a timely manner. While one machine at a time captures the date, the others offload them to the online storage or compute them. The following description focuses on storage, but the idea also applies to computing instead of storing.

Under the condition that the bundled transfer data rate at least equals the recording rate, a number of pipelining devices solves the problem that we encountered. Let R be the recording rate, B the back-end data transfer rate of a single channel, and C the capacity of data transfer to be at least as large as R. Then the number N systems that enable the setting (provided the time for role change is negligible) is $N = 1 + R/B$.

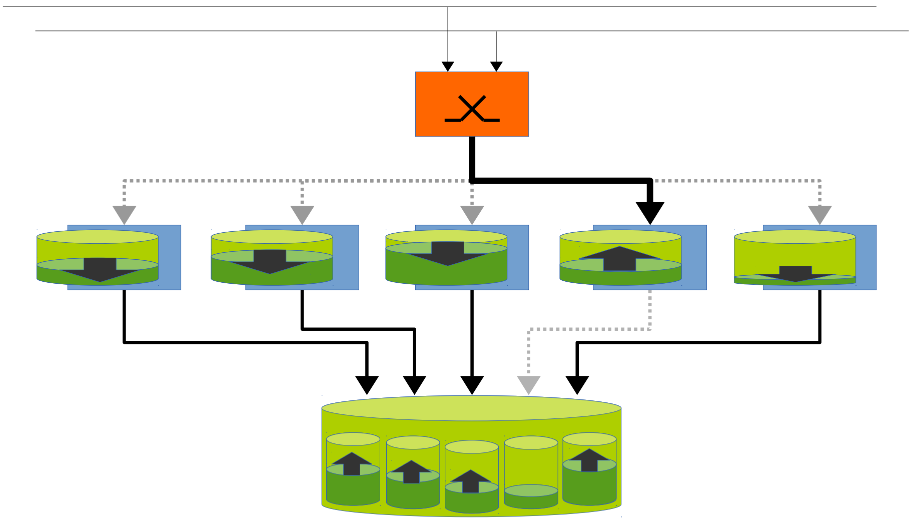


Figure 4.5: Rotating measurement: the disk-writing external storage link capacity as bottle-neck; one node collects data, and the other nodes offload their data into the external storage.

4.1.10 Maintenance Operation

Next, we provide a high-level overview of how we have maintained our deployments in general and how some of the best practices have paid off. More details are given in the sections on monitoring different vantage points below.

User management

User management concerns two principal aspects: security and resource management. It must be synchronized and provide fall-backs in case the default authentication system components fail. Synchronization is important because of consistency. In our experience, it helps users to find the same mechanisms on similar systems. In addition, it helps to provide different environments and mechanisms for different roles and stages of authentication and authorization. For example, a person should not authenticate with identical credentials, e. g., passwords or private keys, for different roles on a single system. If there are stages of authentication along a path that traverse different networks and authentication domains, users often need multiple credentials before they even reach the system on which to work. This setup helps security in the sense that different types and instantiations of authentication systems rarely share the same problems and weaknesses at the same time. Yet, it is inconvenient for users and they will find a more convenient and insecure way to overcome the restriction. Our rules of thumb regarding user management and authentication are as follows:

- Use different authentication methods across different domains: company network, access networks, and deployment site.

- Require different authentication credentials or even authentication techniques for the different roles in a system (e. g. user and super-user).
- Use centralized user and permission management in each domain.
- Monitor the access.

Privilege Management

All aspects of user management also apply to privilege management. We have observed that simplicity is the key for users to work efficiently. In systems, we mostly rely on basic POSIX permission control and only use services that support it. In addition, advanced POSIX capabilities and system hardening techniques help to enforce permissions effectively.

Separation of Management and Measurement Networks

We have already discussed this concern. Confusing management and measurement networks floods the management network with large volumes of measurement data at a high rate and disables system operation and management via the network.

Network Interfaces

The following are a few rules help to effectively separate the networks:

- Measurement should not interfere with the working environment.
- Disable all traffic forwarding on measurement systems.
- Use orthogonal hardware for management and measurement, network components, and system components.
- Use representative naming of devices.
- Monitor data rates in the management network constantly.
- Use redundant channels for management, i. e., serial lines, IPMI console, and SSH access.

Failure Handling

Technical systems, and particularly remotely operated systems, are usually designed for handling failure not as an exception but as a natural occurrence. If physical access to the facilities is not completely impossible, we have the option of walking in and fixing the systems. However, this involves considerable effort (travel) and the risk of touching things that are better left untouched, such as fibers and power cables. The goal of handling failures is always to automate it from detection, to confirmation, and finally to resolution.

Typical Failures

- **Disk failure:** Although vendors attribute the physical (HDD) and solid state (SSD) disks with a high mean time between failure (MTBF), their life time is limited, and 24/7 operation makes failures unavoidable. In a few cases, complete sets of independent hard disks have failed in a short period of time after operating well for a reasonable period of time.
- **Misconfiguration and lock-out:** Changing configurations on a remote site involves highly sensitive parts, e. g., configuring network devices, interfaces, VLANs, IP address configuration, firewalls, routing, and forwarding. Having a backup channel for accessing networked systems, such as IPMI and serial lines, helps to restore a working configuration.
- **Update problems:** Unattended upgrades are a source of trouble for components and systems. It is beneficial to disable unattended updates and execute them manually. This enables testing on a more accessible reference system before deploying updates.
- **Boot failure:** Reboots of any kind bear the risk of triggering problems, e. g., hardware failure or configuration failure. If the system is hosted remotely, IPMI is useful, at least for powering devices on and off, for watching boot messages, and for interacting with the system at boot time. In addition, all default mechanisms that require interaction at boot time are better turned off. A common example is the âĂIJpress key to continueâĂİ prompt, configurable in the BIOS settings.
- **File system failure:** Unresolved file system problems can make a system inoperable and lead to data loss/data corruption. If such problems are detected at boot time, the startup procedure may stop. To minimize the effects of damaged file systems, the boot process should include as few automatic file system mounts as possible, particularly regarding external file systems. Most of them can be checked and mounted automatically once the system is up. In addition, regularly running the auto-check feature for file systems, backups, and snapshots helps to prevent data loss.
- **Timing:** In systems where time synchronization is critical for system operation, e. g., Kerberos-based authentication, we advise maintaining at least one independent authentication mechanism for one single privileged account. In addition, we recommend redundant time synchronization options.

Problems can be avoided by considering failure as a regular aspect of operation. In particular, we avoid physical presence with three measures: 1) *redundancy and fail-over techniques*, 2) *a local technical contact*, and 3) *virtualization of critical services*.

Critical Failures

Our criterion for the category of critical failures is unplanned downtime of systems operation that disrupts current measurement or prevents us from using the resources. Reasons for this category of failures are:

- Hardware failures require replacement of parts.
- Firmware problems may require replacement of parts.
- Attacks and intrusions require offline examination.

All critical failures result in the requirement of physical access to hardware and cannot be solved remotely. Moreover, not all of those failures are preventable by redundancy alone. This holds true particularly for intrusion incidents, although in the course of our work we never experienced such events due to multiple security measures.

Remote Problem-Solving

Independent maintenance channels are a major enabler for remote problem-solving. We used IPMI and related technologies, serial line access, and network access. In a few cases, we were able to receive remote-hands support from the data center operator, mostly for physical capacity extension.

Data Management

The amount of data on sites with large volumes of measurement data requires an effective data management strategy. This includes an overview of the data that are collected and analyzed, and provides information on validity of data sets (e. g., expiration date) for planning and monitoring disk space and archive space. The stored amount of data over time typically increases, although not all data can be kept – both technically and legally. The methodology of data management has already been covered in Section 4.1.4 and includes aggregation and meta-data maintenance. We have had excellent experiences with distributed version control systems for developing and maintaining our tools for data analysis and management. This enabled us to reconstruct measurement procedures and to compare fresh and historical data sets.

Synchronization among Personnel and Machinery

For cost effectiveness, measurement setups may be deployed and operated for multiple purposes. The setup, maintenance, and operation of such deployments are time–consuming, expensive, and require considerable amounts of manpower. The problems with reusing deployed resources relate to *purpose*, *people*, and *time*.

Purpose

A single measurement entity can be of interest to multiple parties, e. g., network traces during major real–world events. For example, the interests in analyzing network traffic during world soccer championship can include:

- General network behavior (traffic pattern change, data volume)
- Web-browsing behavior (i. e., "Does the behavior change?")
- IPTV content popularity (i. e., "Do we see more/less IPTV traffic?")
- Technical QoS and QoE observation (packet inter–arrival times, parallel streams, retransmission detection, session resets)

- User behavior on quality insufficiencies (abort, retry, or bit–rate changes)
- Identification of devices and applications used for watching the event

Time

We regularly capture data over a longer period of time in order to investigate the longitudinal properties of network traffic. The frequency and continuity of measurements significantly depend on the volume of data to be monitored and on the methodology of evaluation. If data can be collected and processed online, then we can easily afford continuous monitoring over months given the feasibility of storing results. In cases of high-speed and high-volume network measurements, however, this is infeasible. Indeed, in only very few cases can data be processed to a reasonable level of detail. For example, it is possible to count packets in real-time inline, but it already becomes problematic when we aim to collect per-flow statistics, because for each packet some computation (e. g., hashing the tuples) is executed and state is required per flow. For instance, at a single 10 Gb/s Ethernet link, we count up to 1.7 million frames per second when the average observed frame size is 733[7].

Instances

In our setting, multiple researchers and technicians use the same vantage point because vantage points are rare due to the rare opportunity of such measurements and the cost of operation. Traces can be shared among the participants only if the trace is stored for later analysis. In addition, online analysis is too complex for piping the data through complex analyzers at capturing time. Moreover, there are conflicts between the properties of the traffic. If not synchronized, concurrent measurements compete for resources on the machinery, can easily interfere with each other, and lead to bias, i. e., the lack of processing capacity will cause packet loss. Thus, in some cases recording multiple instances of data with the same physical resources is infeasible. In the case of Endace DAG cards, only one process can initiate capturing on the hardware at a time. In other cases, i. e., using commodity hardware and tcpdump, it is not a problem to copy multiple instances of a network stream to different processes. Detecting race conditions requires manual or automatic checking of the measurement system state. This can turn out to be non-trivial because the race condition can involve multiple subsystems, e g. switches or routers.

Mostly, multiple and resource-competing interests can be resolved by storing the trace and making it available for later analysis. This in turn requires additional processing, such as anonymization, which requires allocation of additional resources during recording. It also requires full packet traces, and as such makes storage limitations likely – at least in our setups.

[7] This number was measured on a trace in the access network of a major European ISP as by April 2013.

5 Measuring the network of a large European ISP

In this chapter, we focus on the technical challenges and solutions of network monitoring that are not specific to the business of a particular ISP. Legal challenges are business specific, whereas others are specific to the country where the business is located.

5.1 Cooperation and Data

Contract

The contract partners and details of agreements are subject to non-disclosure, but we can sketch some important settings and operational practices.

Data

The network packet header and full packet traces we record and analyze in cooperation with our partner are stored anonymized and for limited time only. After a specified period of time, we will only keep aggregated analyzed data – still anonymized. The capturing happens for the purpose of troubleshooting and can be conducted by only a few people who have signed up for the conditions with our partner. At all times, all data belong to the cooperation partner and access is permitted on a per-case basis. Data remain in the measurement systems only. All analyzed data have to be acknowledged for export (e. g., for the purpose of publication). In addition, the publication itself is subject to approval.
Although we co-maintain and co-operate the systems, the sites are under full administrative and technical control of the cooperation partner, including the possibility to lock us out if circumstances should require it, i. e., due to legal reasons or corporate policy changes. For the legal situation, the Telekommunikationsgesetz (TKG) applies. The operator's data protection officials approve and revisit the project and contract situation on a regular basis.

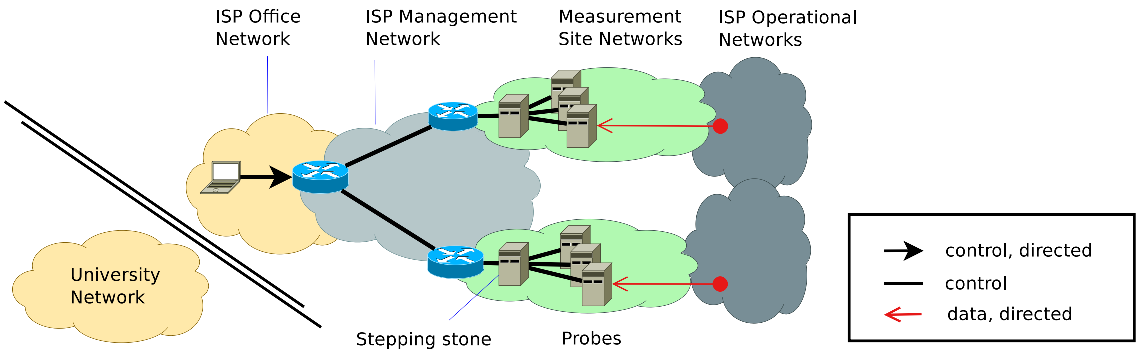


Figure 5.1: Accessing an ISP measurement site: "directed links" are accessible in one direction only, enforced by firewalls. Control path and data path are clearly distinct.

5.2 Ground Truth

There are only few "ground truth" details of interest to the measurement, including

- purpose of each link and connected devices;
- link aggregation information, if applicable; and
- number of customers on a link, if applicable

5.3 Access

Immediate Access

Immediate access is highly limited and rare, but since we co-deploy the equipment and occasionally need to fix the systems physically, we have the possibility to enter the facility under supervision. Personal approval is required for every single case and is personalized to researchers who are well known to the staff operating the facility. All operations at the site require approval beforehand and the purpose is either to repair or extend equipment. In general, physical access is granted exclusively for maintenance operation.

Remote Access

We mainly operate the sites via remote access. Since the sites are owned by and located within the cooperation partner, we are subject to the corporate policy of remote access. That is, we are given devices that are both physically and cryptographically registered for remotely accessing the portion of the corporate network that gives access to the site. Cryptographically, a two-factor authentication is used: personalized tokens and personalized secrets. An overview of the network access scheme is shown in Figure 5.1.

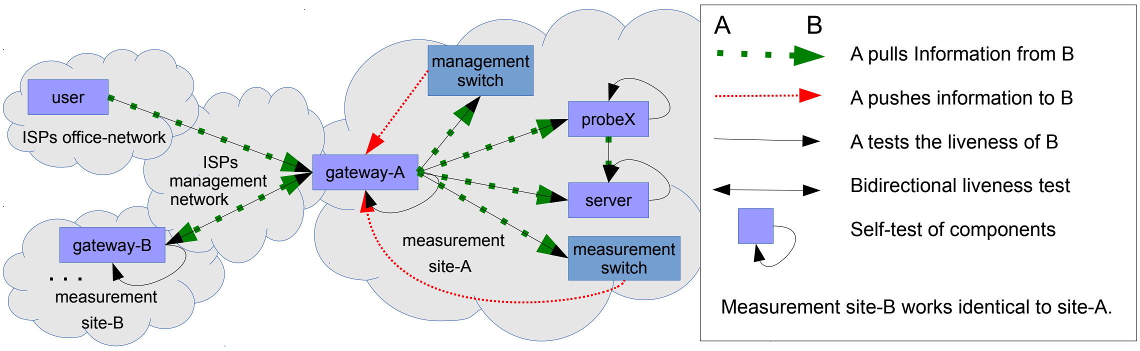


Figure 5.2: Monitoring the sites requires a user to actively pull for information. It returns the current state the monitoring system has computed at the time of the request

.

It is only permitted and possible to initiate communication towards the vantage point deployment sites; the other way of communication, e. g., breaking out, is prohibited and technically impossible.

5.4 Documentation

We use two approaches for documentation. One is required by our partners and involves the corporate intranet for project management, including documentation and file sharing features. All information that is related and interesting to the cooperation partner is maintained there. However, since access to the portal is inconvenient due to VPN and network separation policies, in addition we prefer to host some information directly on the measurement systems. For this, it became convenient to use version control systems for documentation and software maintenance. This approach, is better accessible for users who work with measurement systems. By contract, no confidential information is stored in those repositories. In addition, the documentation, confidential or not, is never publicly accessible.

5.5 Monitoring and Alerting

Monitoring and alerting are difficult because active alerting is suppressed by a strict “no breakout” policy. Currently, we no not run a sophisticated monitoring, but in the following we discuss how it can be done.

The setup is separated into two physical vantage point locations (see Figure 5.1 in Section 5.3). With the ISP’s management network, the sites can exchange traffic, but not at high data rates. Thus, monitoring is hierarchical within each site. We recommend monitoring reachability and service state between the sites. Since breaking out is impossible due to policy, we cannot rely on active monitoring in the sense that problems are reported actively up to the point where a person receives information. Within the ISP's network we can monitor actively, but passing the information over to a real person cannot be initiated from within the ISP's management network.

Thus, we have to pull information actively in the same way that we reach the systems. A scalable way would be to pull information periodically with a local monitoring system. However, we cannot use this mechanism due to the access policy. All accounts are assigned to persons coming with personalized tokens and credentials. They can never be used for automated logins, and therefore pulling state information has to be done manually. We use a chain of hops to connect to the measurement site, mostly via SSH, and then rely on SSH configuration capabilities to pull monitoring information. Concretely, this is a monitoring SSH account on the monitor machine that can be queried via SSH and does not receive command line prompts but instead delivers monitoring information. As a consequence, the monitoring relies on a manual and periodic action of the operator. This is a weak design but the only way to comply with the ISP's access policy.

As mentioned, within the vantage points and among them we are free to use any technology as long as it does not involve external channels, e. g., mobile networks. The monitoring therefore relies on network connectivity, service reachability, and system monitoring. The probing machines, which capture the traffic, are monitored for disk status, temperature, utilization, and user access. In addition, activity and health status of measurement equipment is monitored, e. g., Endace cards report link status, mode of operation, firmware version, and counters. The management switch, which connects a vantage point's setup in a physical location, actively reports connected and disconnected ports via SMTP traps in order to detect physical access attempts. Gateways to the deployments are the obvious place for mutually monitoring each other's availability.

5.6 Synchronization

Our ISP cooperations involve regular weekly telephone conferences with all project members on a private conference system. In addition we have irregular conference calls on a project basis. Email communication works over the ISP's systems and accounts, which is required by confidentiality policies. The ISP operates an intranet application for sharing files and project data (e. g., documentation), which we access with their equipment and the ISP's corporate private network. Thus, information regarding the activities and details on running the vantage point never leave the ISP's systems or the ISP's sphere of influence. Moreover, we attend regular project steering committee meetings with the ISP, where we report past and current activities and where we sketch ideas and plans for future work.

Because of the small number of people who are involved in these activities, we mostly rely on soft synchronization regarding administrative issues and maintenance. Hard synchronization that involves the locking of resources is necessary during packet capture sessions. For system updates, the operating system enforces the resource locking automatically. For switch configuration, we rely on serial lines that cannot be used concurrently. Regarding capturing, our software for high-speed network packet capturing locks the card as soon as a recording is initiated. This conflicts with automatic measurement invocations. For example, if a capturing process starts unattended (i. e., via a cron job) and attempts to access blocked equipment, it fails. The reason for denying the capture is often another active capturing session. We consider automatic measurements as having a higher priority as they are part of a sequence of measurements, which is more difficult to start than a single trace. The mechanism of resolving the conflict would be as follows.

1. Automatic and manually invoked capture sessions start with different priorities and commands. The invocation of an automatic trace recording enforces the acquisition of all required resources, e. g., by gracefully terminating processes that block resources and signaling this to the user who owned the terminated process. The process also requires tap-switch reconfiguration, which is achieved by sending commands to its management unit. In this way, each new automatically invoked measurement wins the resources
2. Races among automatically invoked capturing sessions remain unsolved. A schedule may help to avoid obvious races in the first place. Documentation and management of automatic processes keep track of information on the owner, intended start time, duration, and resource requirements for a capturing session. This information can be managed via XML forms or other structured formats. The process of modifying the automatic recording schedule involves collision detection and reporting.

While the first suggestion was easy to realize, we have never implemented the latter because our schedule is relatively sparse, and thus soft-synchronizing among people works well.

5.7 Planning and Future Scalability

When planning and deploying the sites, we were aware of the general problem of scalability. In the first iteration, we had one deployment location and started with a basic setup. We used a gateway connected to a switched network with one compute server and two measurement probes. The probes were connected directly to one optical splitter each, to account for both directions of the link. The access to the gateway was different back then and changed over time. The probes were each equipped with a 1-port 10 Gb/s Endace 5.2X network measurement card, and used the Endace time synchronization for capturing time stamps between the two cards. Each of the probes had a hardware-accelerated local raid and both connected via Ethernet and network file system to the compute server. The server connected to an external hardware-accelerated large RAID device.

With time, new links required a probe machine, equipped with a 4-port 1 Gb/s card that covered the additional links. Then the success of this project made more links available, and we decided to use a tapping switch, equipped with 1 Gb/s and 10 Gb/s port modules. This afforded us remote link and selection as well as selecting collectors for each measurement. At some point, we mounted additional probe machines equipped with Endace measurement cards. Connecting them to all measurement links simply required that they be connected to the tapping switch. Additional monitoring links followed. Thus, the site grew as the demand for measurements and analysis grew.

For our second physical vantage point location within the same ISP, we cloned the setup. With the availability of better remote management modules, all applicable systems were upgraded. The setups differ in terms of storage and computational power, due to the availability of components at the time of purchase. The old probe machines are still useful, since the computational power for capturing was sufficient from the beginning and over time we replaced the internal raid hard drives by larger ones.

With the extension to the initial deployment, we decided to automate. In the first place, we simply synchronized the login information with a semi-automatic mechanism of database-supported user management. This still required copying information to the systems and running configuration tools. At later stages, however, LDAP replaced this mechanism and we deployed an internal DNS service. With the second deployment, we synchronized LDAP and DNS between the two sites.

5.8 Measurement Operation

For legal reasons, each measurement requires a use case and justification. Measurements are synchronized between the project participants for the sake of legal and organizational issues. The following parameters of capturing are determined in advance, since they may involve resources (e. g., storage and/or switch monitoring sessions):

1. **Set of links:** Monitoring more than a single link can be done for two reasons: (a) to measure a device in the path between the two links, and (b) to measure link aggregation setups.
2. **Date, time, and duration:** This is due to synchronization among measurements. We cannot simply measure an arbitrary number or sets of links at the same time. The duration is bound by the available space on the storage systems.
3. **Snap length of packets:** This parameter is important in the context of privacy protection as well as in the context of storage management. Full packet traces are expensive to store and require additional care to respect privacy. They cannot be offloaded to external storage in real time, not even for 1 Gb/s tapping link speed. Limited snap-length traces require less storage capacity, are less problematic with regard to privacy protection, and in some cases can be analyzed or offloaded during capturing time to external storage over the management network. This allows for longer tracing duration on the one hand, but on the other hand snap-length limited traces are also of limited use compared to full packet traces.
4. **Privacy protection measures:** When capturing a trace, it is a legal requirement that the purpose and the analysis have to be known. This allows us to determine the legally required precautions. It also allows for anonymization that preserves privacy without limiting the analysis. For example, timing and link-specific measurements are agnostic of source, destination, and content of network traffic.
5. **Filters and sampling parameters:** Occasionally, the complete traffic on a link is not required for the measurement. Filtering and sampling allow for drastic data volume reduction and may enable inline computation or data offloading.
6. **Hardware Resources:** This specifies the hardware resources and configurations involved in the measurement. Moreover, it includes checking for synchronization among the equipment, health, and resource reservation. While some equipment can only be used exclusively at one time, other equipment needs to be shared between measurements, e. g., the tapping switch.

7. **Storage resources:** We specify limits on the size but also on the physical location for storing the measurement data.
8. **Person responsible:** A single person is in charge of supervising a particular trace collection. This person is responsible for meeting all legal and technical requirements. In addition, this person maintains the documentation on the trace.
9. **Trace permission and location:** In cases in which data from a trace need to be stored in a non-aggregated format, the trace needs to be protected from unauthorized access. Protection can be enforced via access permission and additional encryption.

5.9 Maintenance Operation

The challenge of maintaining the ISPs network measurement deployment is in the limited opportunity for physical access. We receive almost no remote-hands support from the ISP and we face tough restrictions regarding monitoring. There is no permanent technical communication channel, e. g., a tunnel. We approach these challenges mainly with a systems design for high failure tolerance for *hardware* and *services*.

Hardware failure tolerance

Hardware fails regularly, since systems and components have a limited lifetime. Examples of failing components in our setting include disks, main-boards, raid controllers, batteries, and measurement devices. However, we have also experienced problems with KVM switches, network switches, and serial lines. In the case of hard disk arrays, we switched from RAID level 5 with spare disk to RAID 6, and we operate each system with a RAID 1 (i. e., mirror mode) system disk configuration. The performance penalty is negligible for our use cases. In some cases with larger (24) numbers of disks, we switched to RAID60, as described in Section 4.1.4. Still, a single point of failure is hardware raid controllers. Since hardware raid controller failure can lead to complete and irreversible data loss, we recommend two measures:

- Use raid controllers that do not store configuration on the controller. Otherwise a controller failure results in data loss.
- Use identical or similar but tested raid controllers across machines. Data can easily be restored by connecting a similar controller that is borrowed from another system.

The same applies for specialized hardware, such as measurement cards. In the case of Endace measurement hardware, we have regularly experienced longer than half-year repair times, occasionally even without success. This causes downtime of parts of the system. If more systems use the same hardware, it becomes more affordable to keep spare parts of critical components.

For some hardware components, failure is difficult to detect remotely; for instance, cables can only be checked via physical access. In cases in which redundancy mechanisms are not built into components, we provide an alternative technology to work around failing parts. This includes electrical power resets, network connectivity for remote access, and data recovery. IPMI technology and electrical power switches are alternatives to each other for emergency electrical power control. IPMI and serial line connectivity can serve as a backup console communication channel for network access. For data recovery, backups and data mirroring across systems work well.

A particularly problematic case of hardware failure is on-site equipment for on-site access, e. g., Keyboard, Video, Mouse (KMVs) or displays. The problem is that this kind of failure requires returning to the location with a replacement. In the case of several 100 km of travel distance to the site, this kind of failure is not tolerable. The best work-around and backup technology is network connectivity for site-local access, so that a person on-site can connect a mobile system to the deployment (e. g., a laptop). This is not only a matter of configuration but first and foremost a matter of security.

Service Failure Tolerance

Service failure can be caused by human interaction, software failure, and configuration errors. Particularly services that are required for remotely accessing a system or those dealing with authentication and authorization information are failure critical. Services with which we have experienced the most failure are SSH, LDAP, and the DNS.

To meet this challenge, we use the same approach as for hardware failure: redundancy and backup channels. The DNS and LDAP are configured for master-slave data and service duplication. When server certificates are involved, the master and the slave server certificates have different expiration times. Thus, in cases of certificate expiration, the service is still available. DNS and LDAP are essential for authentication and SSH is required for remote access. Thus, if any of those services fail, we need a channel into the machines, independent from network-based authentication systems. Every single system has a local authentication mechanism that has been tested without network connectivity. Mostly, we rely on IPMI remote consoles, serial lines, and a local authentication that overrides the network-based authentication system.

6 Measuring the Network of a Large European IXP

6.1 Cooperation and Data

Contract

The contract partners and bodies are subject to confidentiality agreements and non-disclosure. We cannot elaborate on the details and identities, but can again sketch the general setting, operation, and some of the important conditions and procedures for cooperation. As already pointed out, such contracts are often the only possibility to gain access to a specific type of vantage point and to its data. The huge drawback is clearly the non-verifiability of data and results. However, peers in other locations can validate the results if the methodology is documented.

The contract explicitly allows for publication only with the agreement of the partners. In practice, we send them figures, drafts, and final versions of intended publication and receive permission for publication with a short delay. We cannot publish operational details of the IXPs or any of their members if the information was obtained from measurements at this deployment location. Moreover, we cannot elaborate on specific members' data or usage of the IXP links. However, anonymized information can be published, e. g., pointing to the members' role (i. e., "large European ISP", "minor content delivery network", or "local data center"). Information that we publish is aggregated and we take care not to interfere with the operation and the business of the IXP and its members.

6.2 Ground Truth

We obtain important ground truth from the IXP, which is either not measurable or helps to validate our measurements. Such data are often publicly available on the IXP's website. They can include member-specific information and internal networks. Some IXPs (i. e., LINX) even publish highly detailed information here[1], such as per-member and per-device organization name and ASN, IP addresses, MAC addresses, physical location, service speed, switch-ID and switch-port, VLANs, and membership type. However, there are no per-member statistics that can reveal information

[1] see: `http://www.linx.net/pubtools/member-techlist.html`

about the business of members. Yet, the information helps to assemble a picture of the market and to correlate it with routing tables and other data sets.

Other ground truth information includes:

- **Operational details** cover the operational model of the IXP and shed light on the IXP's business decisions. For example, some IXPs allow remote peering. The information relates, for example, to capacity planning, when there are multiple facilities and instances that the IXP operates.
- **Configuration**, e. g., network structure or service configuration, is important. For example, it is indispensable to know how data collection mechanisms are set up in order to draw the correct conclusions from measurements. Parameters of the configuration include sampling rates, filtering mechanisms, and filter settings.
- **Market insight** regarding the different IXPs provides information about the motivation of members, re-sellers, and partners to join that IXP. Because the IXP business has so far been understudied, we cannot rely only on publicly available information. Instead, we collect pieces of the puzzle from the best available sources: the market participants.
- **Service catalogues** are one of the major distinctions between IXPs. They can help in understanding which kinds of players in the Internet ecosystem are attracted to a specific IXP, and sometimes it is enlightening to see which members use which subset of services or – more interestingly – refuse to use a subset of free-of-charge services.

6.3 Access

In general, we cannot physically access the IXPs for different reasons. First, the IXPs operate at locations that we can only reach by long distance traveling. Second, they operate in highly secure environments, such as data centers. In some cases, even the location of a data center is confidential. In addition, the IXPs have strict requirements for external staff to enter their operation facilities. Finally, the third and most convincing argument is the existence of complete (and excellent) technical support from IXPs. After all, the main business of IXPs is to connect members to their locations. For that reason, IXP members place routers and other equipment with the IXP. However, IXPs do not let members' technicians enter the facilities at their convenience; therefore, they run a well-established service of remote hands and equipment deployment that we gladly take advantage of.

In general, we cannot physically access the IXPs for different reasons. First, the IXPs operate at locations that we can only reach by long distance traveling. Second, they operate in highly secure environments, such as data centers. In some cases, even the location of a data center is confidential. In addition, the IXPs have strict requirements for external staff to enter their operation facilities. Finally, the third and most convincing argument is the existence of complete (and excellent) technical support from IXPs. After all, the main business of IXPs is to connect members to their locations. For that reason, IXP members place routers and other equipment with the IXP. However, IXPs do not let members' technicians enter the facilities at their convenience;

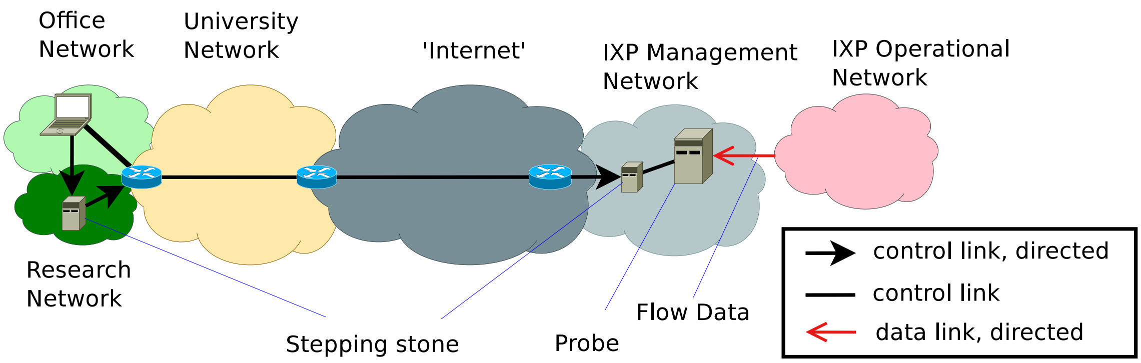


Figure 6.1: Accessing an IXP measurement site: "directed links" are accessible in one direction only, enforced by firewalls. Control path and data path are clearly distinct.

therefore, they run a well-established service of remote hands and equipment deployment that we gladly take advantage of.

1. We can access the remote site only from specific machines within our own network that are unreachable from outside our network.
2. Access to our machines that connect to the IXPs is carefully secured and includes multi-factor authentication. At the same time, these machines are restricted to connectivity to a well-defined set of IP addresses and only use state-of-the-art cryptographic secure protocols.
3. The remote network is reachable via "stepping stone" only. That stepping stone is under control of the cooperation partner (IXP) and also uses state-of-the-art security mechanisms to prevent the most popular attacks (e. g., password guessing). Access is granted using personalized tokens only.
4. The stepping stone connects to a well-defined set of machines and is physically separated from the IXP's operational network. Moreover, access is monitored. Signaling out from the IXP's network is blocked.
5. Our machines within the network are accessible via the IXP's own network only, including the stepping stone. The stepping stone uses a different authentication system than the IXP's. The IXPs IT staff has unlimited access to our equipment

This enables the cooperation partner to have all components in its network under full control, allowing for interruption and access limitation to the measurement deployment.

6.4 Technical Parameters

We show the state-of-the-art deployment layouts in Figures 6.2 and 6.3. The setup in Figure 6.2 depicts the current deployment, and Figure6.3 shows the deployment under construction. Presently there is no separation of the collector and analyzer. Thus, a problem with the analyzer may result in the interruption of the recording process. In the future, it will be possible to separate both functions with additional hardware deployment. Our motivation for separating the collector and

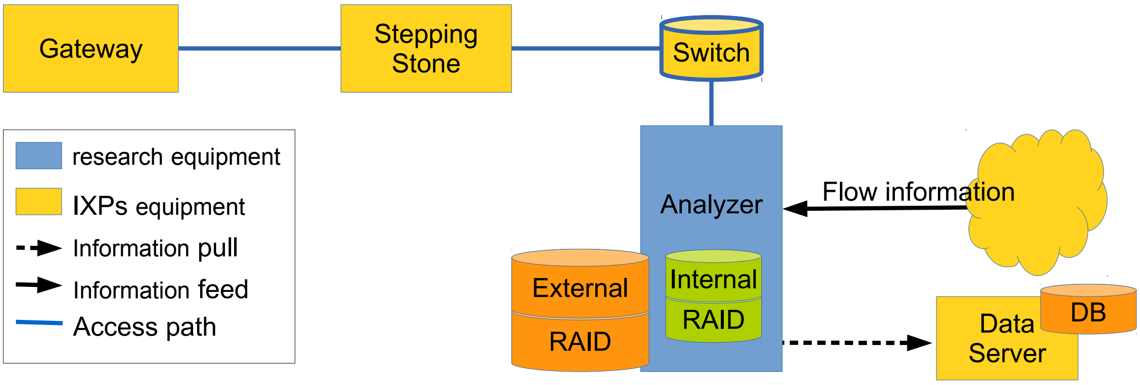


Figure 6.2: IXP measurement site: Collecting flow information at an IXP

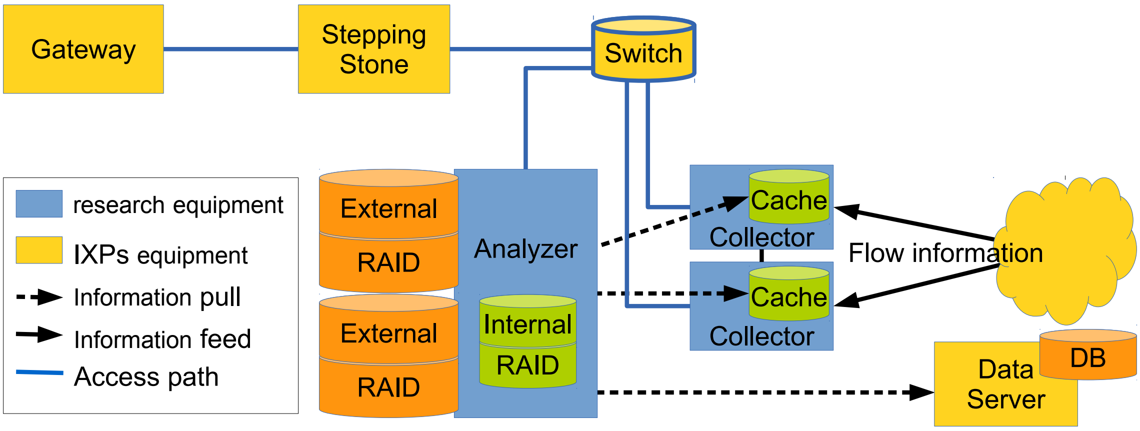


Figure 6.3: IXP measurement site (improved): Collecting flow information at an IXP with additional reliability

analyzer is the possibility of resource demand interference regarding storage, CPU, RAM, and OS. At this point, we restrict analysis tasks to a low priority compared to capturing, preprocessing, and recording. However, if system resources are over-utilized, the collection process suffers. The split allows us to take down the analyzer for maintenance without interrupting the measurement process. An even better redundancy can be achieved by two independent collectors, both capable of and enabled for recording by default. This requires a setup where both collectors are fed the same data simultaneously. The analyzer and storage machine can then pull cached data from the collectors at their own convenience. In addition, we separate collecting the flow data in one machine from processing it in another. This improves the reliability of our setup. A synchronization mechanism ensures that data can be cached in both collectors redundantly. Once the analyzer copies and removes data from one collector, the same data are automatically removed from the other collector as well. Consistency can be achieved by remote mounting of the cached data on the analyzer via a network file system (e. g., NFS) and by using special copy and move operators that take care of the synchronization between both collectors.

6.5 Monitoring and Alerting

Again, on-site monitoring is possible by using a state-of-the-art monitoring system, such as Nagios, which can monitor state of services and devices. However, active signaling out is not an option, so we propose an email-based system that involves the IXP's email system. However, simply relying on sending email if a problem occurs is not sufficient, because a failure of the email mechanism will go undetected. Therefore, we use a "dead-man switch", and send (machine-readable) status reports via email to our own monitoring system. The emails are end-to-end encrypted to ensure confidentiality. Our own monitor then checks for the presence as well as the status of the report.

6.6 Synchronization

At this point in the process of measurement, we need to synchronize the activities on the measurement deployments. The motivation for this is that we currently have only one (powerful) physical machine deployed, which does both data collection and data analysis. In this configuration, a system overload (e. g., fully utilizing CPU or RAM, or imposing a high read/write load) impacts the reliability of the current data collection process.

Organizational Synchronization: To prevent situations of overloaded systems and measurement bias during the capturing, we rely on an asynchronous information flow between the involved researchers. This means that information is exchanged on demand and on a per-case basis only. For the synchronization between our researchers and the IXP staff, we use the IXP's systems (i. e., trouble ticket system) and rely on email and phone calls.

Technical Synchronization: We rely on prioritizing the capturing process above the level that a regular user has. One of the shortcomings of this method is the fact that a large number of parallel processes can still slow down even the most highly prioritized. The same observation is valid for using disk space and disk performance. Thus, even the most highly prioritized process cannot write if no space is available to write to. In addition, a high number or concurrent reads/writes impacts the performance of each single instance. In the next iteration of upgrading the deployment, a physical separation of collector and analyzer tasks will solve the problem. The setup, which is in the phase of deployment, is described in Figure 6.2. The challenge in terms of synchronization is then the data flow coordination between machines, but this is a much less critical task and does not impact the collection itself.

6.7 Planning and Future Scalability

As explained in Section 4.1.8, we started out with a simple deployment until we recognized the success of the measurements. However, we have reserved enough physical space and the commitment of our partners to cooperate when demand rises. The following steps are either in planning or in implementation.

External storage: While raw data can be aggregated for most methods of analysis, it is beneficial to keep full data slices over a period of time to have the opportunity to run analyses on historic data and cover the longitudinal dimension. It also pays off to keep a recent window of recorded data, because it happens that we learn about Internet incidents from the news. To analyze the effects at our measurement points, the data need to have been recorded at the time of the incident. Another reason is the availability of other types of data over time. For example, new data types are added to what we already have, but types of information in the data change. As an example, one of the major European IXPs just switched from SFLOW monitoring to IPFIX-based monitoring. This requires more storage and provides different kinds of data.

Over time, recording capacity demand does not grow per se, but our demand for reliability first increases. One first step to adding reliability and stability is to physically separate recording and analysis. As an additional measure, it is possible to deploy redundant recorders, so that maintenance or problems on one recorder do(es) not affect the availability of recording itself. The recorders need enough internal storage to account for a reasonable time span of caching data before loading off to the storage that is connected to the analyzers.

Analyzing capacity: Depending on the resources that we use at the vantage points (physical or virtualized infrastructure), the computation demand can be scaled. In a virtualized environment, where the equipment completely belongs to the partner who provides the resources, it is feasible to scale up capacity quickly and on a demand basis. In cases in which we run our own equipment inside the IXP, conversely, we have to account for the systems' cost and time of deployment. While at the moment we parallelize per task, we still rely on more traditional ways of data analysis that will change with availability of easy-to-use massive-parallel computation approaches, such as Stratosphere [33]. In case such a system turns out to be applicable for our purpose, i. e., accounting for non-repeating and ad-hoc data analysis, we will consider adding new machines for data analysis.

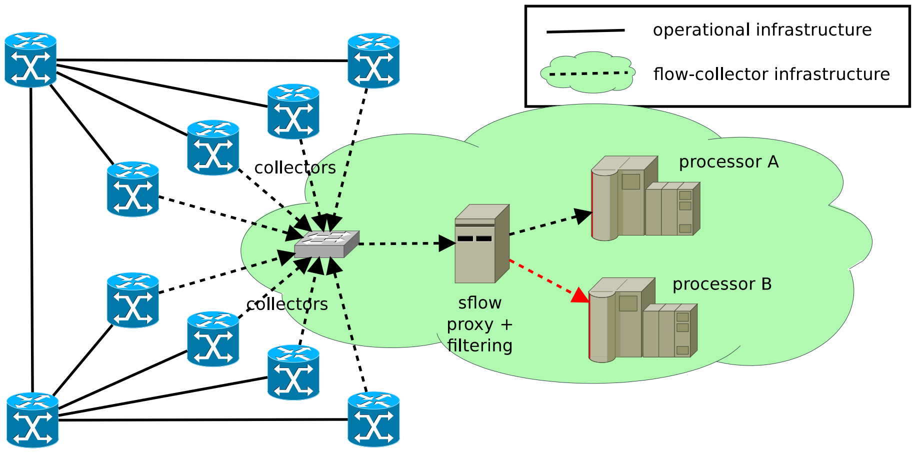


Figure 6.4: Flow collection setup: The operational and the management network are separated and the SFLOW-proxy can apply different filters per flow-processor, i. e., sampling rate or input set.

6.8 Measurement Operation

Our measurement and recording operation involves the following components: aggregation, filtering, sampling, and anonymization via pseudonymization. While the first three are typical measurement approaches for reducing data volume, the latter is a legal requirement.

Typically, measurements based on SFLOW data are performed continuously, and filtering and sampling are done by the router or aggregation infrastructure. With regard to filtering, we mostly keep the “interesting” parts of data, e. g., data from when something remarkable happened in the worldwide Internet ecosystem. In addition, “regular” network traffic data are stored as well to have data against which to compare the “irregular” behavior. Sampling is used as a mechanism and as a result we keep traffic slices, i. e., a week per month over a longer period of time.

We collect a variety of data types and formats at each IXP, the most important format being SFLOW. The SFLOW collection mechanism (see Figure 6.4) is based on UDP data streams that are sent to our equipment. Not picking up the data flow is uncritical for the sending machines, and in addition we suppress the corresponding ICMP messages back to the sender if our equipment cannot accept the arriving data.

The flow information can either be collected with SFLOW-collector software, or simply be recorded via tcpdump. The advantage of the latter methodology is the timestamps per frame, which provide information on the time synchronization between the different SFLOW sources.

Other sources of information that we receive from the IXPs are route-server information and BGP traffic from the route-server. This information is not provided as a stream; instead, it is collected and aggregated on a timely basis (i. e., hourly) and available for download from within the IXP’s management network.

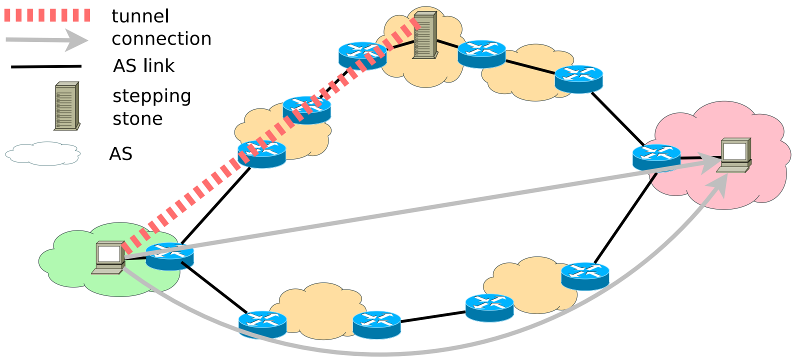


Figure 6.5: Different access paths: in case the default router (lower one) breaks, once a tunnel is established the backup path can be used for accessing the destination.

For our work, we cannot only rely on data that we collect at the IXP. There are cases in which related information is required that must actively be fetched, e. g., DNS or X.509 certificate information. This cannot be done with our deployment at the IXP. Instead, we conduct additional active measurements based on the IXP's data. For this, we use our own machines within the university network.

6.9 Maintenance Operation

Regarding maintenance, most of the ISP-related aspects discussed in Section 5.9 apply here as well. We have established backup channels for accessing our infrastructure. In addition, we have found it to be useful to have more than one path to the IXP's networks, since we face failure of our own research network (DFN) towards the partners' networks on a regular basis. Thus, we have deployed a stepping stone in some other AS that is not involved in the AS path between our partner and ourselves, and our cooperation partner white-listed that stepping stone for accessing our deployment through that channel. Figure 6.5 sketches this setup.

7 Measuring from an Active Vantage Point at a University

For our active measurements, we run different labs within our partition of the TU Berlin campus network. The labs use different network environments with restrictions and limitations regarding connectivity and access. In addition, parts of our measurement infrastructure are logically located outside of our assigned subnet in order to meet requirements of the IT department and for our own security considerations.

7.1 Cooperation and Data

In terms of cooperation, connectivity-wise we rely on services of the IT department since we run our own subnet within TU Berlin. Our subnet is partitioned and we separate lab and desktop environments. Desk space refers to the subnet as well as to physical space where workstations are operated for office and research work. The lab space is partitioned into different subnets that serve different purposes and have different access restrictions. Regarding network connectivity, the labs are within reach from the office subnet, but are restricted regarding network traffic and access. Labs are all separated from the office for the following reasons.

Lab equipment is mostly rack-mounted, while workstations should be physically accessible by users. Physical separation prevents unauthorized bridging networks by conventional means. Different access policies are much easier to handle on a per-network and a per-location basis. We completely isolate the traffic of some networks and restrict the traffic of others, because measurement traffic must not interfere with office work. While break-outs from the office network are necessary and permitted for daily office work, the policy for lab traffic is much more restrictive in this regard and allows outgoing connections on a per-machine and per-experiment basis. In addition, the network separation allows for better separation of administration. Thus, the rules for the different networks can be configured by those operators who are most knowledgeable in each lab environment. Moreover, we run different services in the labs and the office network, which are specific to network purpose. In the following, we describe some of ours labs.

1. **Isolated research network:** The *Routerlab*[1] is a private network and consists of a number of routers, switches, load generators, and other equipment. We use it for research and teach-

[1] Routerlab: `http://www.inet.tu-berlin.de/?id=routerlab`

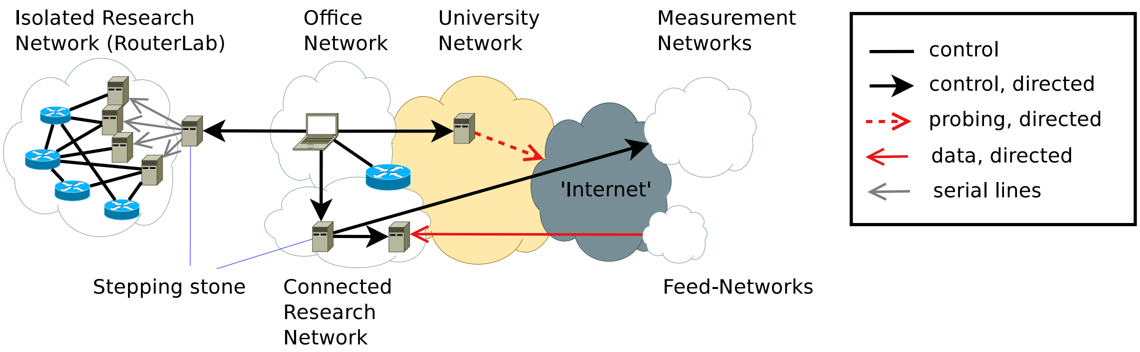


Figure 7.1: Running an active measurement setup: directed links are accessible in one direction only, enforced by firewalls and filters. Particularly the connectivity to partners are encoded in the rules, for both control paths (black) and data paths (red).

ing, and for configuring, running, and measuring arbitrary network setups with real-world hardware. A much larger and more widely known but similar deployment is Wisconsin Advanced Internet Lab (WAIL[2]) at the University of Wisconsin – Madison. For security reasons, Routerlab devices other than the stepping stone have no physical or routed network connectivity to any other network. The stepping stone itself is reachable exclusively for authorized users inside our own network. For remote access, we use the stepping stone that connects all other equipment via serial line technology, i. e., either directly via serial lines or via serial line switches. Management of components and a resource reservation system is implemented with our dedicated "LabTool" software [109].

2. **Connected research network:** The *Lionden* network consists of machines that are either accessible from the Internet for receiving active data feeds or require us to access services on the Internet for active measurements and passive data feeds. Active feeds are push services that we receive from external sources. One example is a spam feed from Abusix3 that arrives as email traffic. For that purpose, the feed data are received on SMTP port 25 (TCP). Lionden is additionally protected with a firewall, which takes care of enforcing the policies for network traffic towards and outbound from Lionden. This implies some limitations and drawbacks regarding active measurements. The firewall considers network state, which by its capacity practically limits the number of parallel measurements effectively. Moreover, the firewall also serves for the office network segment, and we try not to interfere with the networks. Machines that require connectivity to the Internet are either exclusively reachable from outside or they can connect out, but never both. Our active measurements are mostly of a qualitative nature; we do not measure speed or other time- and volume-based metrics. This allows for a virtualized environment, which in turn allows us to restrict access to machines via "pseudo serial lines" of the virtualization system and prevents breaking out from such machines in the direction of our other networks. On the firewall, we block by default all network traffic originating from our measurement networks towards the campus address space (including our partition).

[2] WAIL: http://wail.cs.wisc.edu/

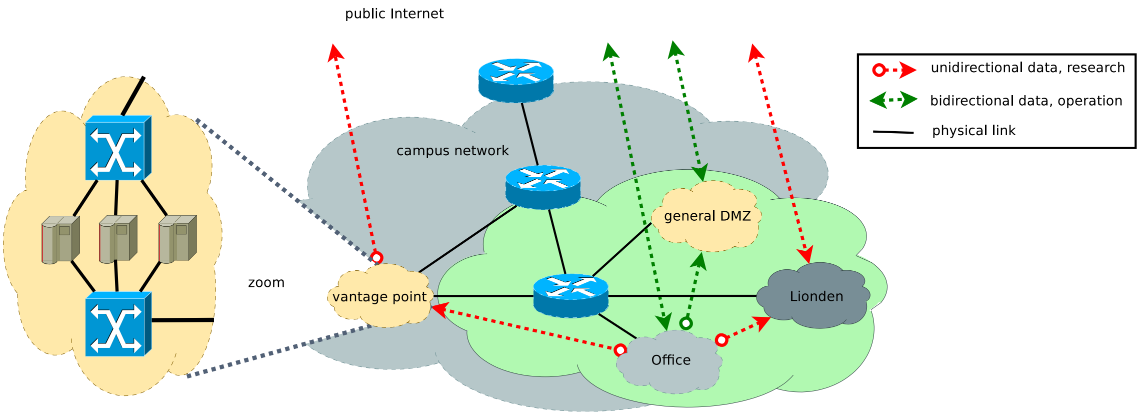


Figure 7.2: Architectural Overview of the active measurement vantage point at our Labs at TU Berlin

3. **Vantage Point:** In addition to operating a DMZ for day-to-day public service, we also operate a research DMZ that is located outside the central firewall and hosts machines that are used for massive-parallel and high-frequency active measurements. Those measurements cannot pass a stateful firewall because of too complex for our regular firewalls and gateways. This brings various restrictions to the systems that we operate in this network segment. First, they do not offer any network services to public networks, nor do they accept connections from the campus network or the Internet. Those restrictions are monitored. The measurement machines connect to networks with at least two interfaces: one for management and operation, and the other for measuring network traffic. Management and measurement networks must not be connected via routing, proxying or tunneling. The connectivity scheme in this part of our network is shown in the left zoomed part of Figure 7.2.

In figure 7.2 we sketch the part of our network. Unidirectional data flows describe connections that can only be initiated in one direction.

This cooperation has been highly successful in the past, but the central IT department has a number of restrictions that require some consideration when it comes to active measurements. We would like to point out explicitly, that we do not trace the campus network, nor does the IT department provide us with any network traffic trace data.

7.2 Access

Access to the measurement setup is not as critical and sensitive as access to the ISP and IXP measurement setups. Nevertheless, we implemented a network policy to prevent unauthorized access. Just as the passive measurement sites, the setups are only accessible from our internal network.

Isolated research network

Since the traffic in our Routerlab is experimental, we cannot allow it to interfere with real-world networks. Therefore, this network is shielded with a stepping stone, which only connects to the Routerlab and does not leak traffic. As a consequence, the network access is one-way only and limited to control traffic to disable artificial traffic from experiments from leaving the isolated environment.

Connected research network

The Lionden network is operated for the purpose of having an additional barrier to accessing our external measurements. We run two roles of machines in this network: stepping stones, and active and passive collectors.

Stepping stones are machines with a narrow user base that do not offer any services besides putting authenticated users through. The addresses of those machines are given to our partners to implement network access control for their infrastructure.

Active and passive collectors are machines that collect data from external sources. This explicitly excludes probing traffic. Services that we run here are (s)FTP-based data collection of publicly available services and SMTP-based spam feed collection from an external partner. The rules for the permitted network traffic are deployed in the firewall, allowing traffic only from and to specific addresses and services from and to our dedicated machines. Moreover, we restrict data rates on the link that connects Lionden to the Internet.

Control access (e. g., SSH) to this research network must originate from our office network. The authentication system is distinct from the office environment, and allow-lists are specified per machine.

Active Probes

Active probe machines are connected to at least two networks via separate physical network interfaces. One interface connects to the research and office networks, and it allows for incoming[3] encrypted control traffic (SSH) only. No measurement network traffic is allowed towards our own networks. The other interfaces connect directly to the campus network. The reason is to avoid stateful firewall inspection of measurement traffic, which would put a significant burden on the packet filters in the central firewall. Network rules block all traffic that is unrelated to the measurements, including all traffic that is not initiated by the probe itself. The network rules are more relaxed than those of the connected research network since the measurements are not long-term and the data are also less sensitive. Authentication and access lists are implemented per machine and meet up-do-date standards regarding security.

[3] incoming to the probes, originating from our local networks

7.3 Planning and Future Scalability

The active measurement equipment is located in physical reach of its operators. It is common to deploy new equipment and to reuse platforms. In short, the important aspects of planning are power consumption, space, and re-usability.

7.4 Summary

From eight years of experience planning, deploying, and operating passive and active network measurement infrastructure with highly different requirements and with different partners, we have shared insight into the multiple dimensions of different challenges. This includes operating networks in different remote operator locations and home measurement deployments.

8

The Pulse of Today's Internet at an IXP

In this chapter, we aim for enriching the knowledge on understanding the Internet and current trends by looking at the data plane traffic of a single major IXP, running the "European business model". We show that the AS level on the one hand still is an important entity for understanding the Internet but on the other hand, the usage of address space has changed in a way that cannot be described on AS granularity. For our study we use not only data from measurements at this major European IXP, but we enrich this data with active measurements, taken from an academic network vantage point.

8.1 IXP as Rich Data Source

In this section, we describe the IXP measurements that are at our disposal for this study and sketch and illustrate the basic methodology we use to identify the traffic components relevant for our work. We also list and comment on the different IXP-external datasets that we rely on throughout this chapter to show what we can and cannot discern from the IXP-internal measurements alone.[1].

8.1.1 Available IXP-internal Datasets

The work is based on traffic measurements collected between August 27 (beginning of week 35) and December 23 (end of week 51) of 2012 at one of the largest IXPs in Europe. At the beginning of the measurement period in week 35, this IXP had 443 member ASs that exchanged on average some 11.9PB of traffic per day over the IXP's public peering infrastructure (i. e., a layer-2 switching fabric distributed over a number of data centers within the city where the IXP is located). During the measurement period, the IXP added 1 or 2 members per week. Specifically, the measurements we rely on consist of 17 consecutive weeks of uninterrupted anonymized SFLOW records that contain Ethernet frame samples that were collected using a random sampling of 1 out of $16K$. SFLOW captures the first 128 bytes of each sampled frame. This implies that in the case of IPv4 packets the available information consists of the full IP and transport layer headers and

[1] For an overview of the importance of IXPs for today's Internet, we refer to [66] and section 3.2

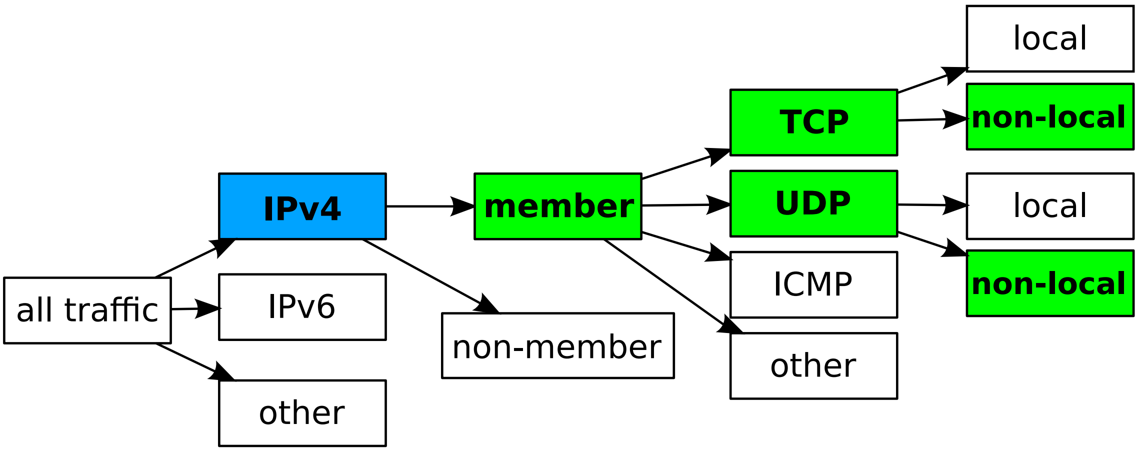


Figure 8.1: Traffic filtering steps

74 and 86 bytes of TCP and UDP payload, respectively. For further details about the IXP infrastructure itself as well as the collected SFLOW measurements (e. g., absence of sampling bias), we refer to [43]. In the following, we use our week 45 data to illustrate our method. The other weekly snapshots produce very similar results and are discussed in more detail in Section 8.3.

8.1.2 Methods for Dissecting the IXP's Traffic

Peering Traffic

Figure 8.1 details the filtering steps that we applied to the raw SFLOW records collected at this IXP to obtain what we refer to as the "peering traffic" component of the overall traffic. As shown in Figure 8.1, after removing from the overall traffic, in succession, all non-IPv4 traffic (i. e., native IPv6 and other protocols; roughly 0.4 % of the total traffic, most of which is native IPv6), all traffic that is either not member-to-member or stays local (e. g., IXP management traffic; about 0.6 %), all member-to-member IPv4 traffic that is not TCP or UDP (i. e., ICMP and other transport protocols; less than 0.5 %), this peering traffic makes up more than 98.5 % of the total traffic. As an interesting by-product, we observe that 82 % of the peering traffic is TCP and 18 % is UDP.

Web-Server Related Traffic

We next identify the portion of the peering traffic that can be unambiguously identified as Web server-related traffic. Our motivation is that Web servers are generally considered to be the engines of e-commerce, which in turn argues that Web server-related traffic is, in general, a good proxy for the commercial portion of Internet traffic. Accordingly, we focus on HTTP and HTTPS and describe the filtering steps for extracting their traffic.

To identify HTTP traffic, we rely primarily on commonly-used string-matching techniques applied to the content of the 128 bytes of each sampled frame. We use two different patterns. The first pattern matches the initial line of request and response packets and looks for HTTP method words (e. g., GET, HEAD, POST) and the words HTTP/1.{0,1}. The second pattern applies to header lines in any packet of a connection and relies on commonly used HTTP header field words as documented in the relevant RFCs and W3C specifications (e. g., Host, Server, Access-Control-Allow-Methods). Using these techniques enables us to identify which of the IP endpoints act as servers and which ones act as clients. When applied to our week 45 data, we identify about $1.3M$ server IPs together with roughly $40M$ client IPs. Checking the port numbers, we verify that more than 80 % of the server IPs use the expected TCP ports, i. e., 80 and 8080. Some 5 % of them also use ports 1935 (RTMP) as well as 443 (HTTPS). Note that by relying on string-matching, we miss those servers for which our SFLOW records do not contain sufficient information; we also might mis-classify as clients some of those servers that "talk" with other servers and for which only their client-related activity is captured in our data.

With respect to HTTPS traffic, since we cannot use pattern matching directly due to encryption, we use a mixed passive and active measurement approach. In a first step, we use traffic on TCP port 443 to identify a candidate set of IPs of HTTPS servers. Here, we clearly miss HTTPS servers that do not use port 443, but we consider them not to be commercially relevant. However, given that TCP port 443 is commonly used to circumvent firewalls and proxy rules for other kinds of traffic (e. g., SSH servers or VPNs running TCP port 443), in a second step we rule out non-HTTPS related use by relying on active measurements. For this purpose, we crawl each IP in our candidate set for an X.509 certificate chain and check the validity of the returned X.509 certificates. For those IPs that pass the checks of the certificate, we extract the names for which the X.509 certificate is valid and the purpose for which it was issued. In particular, we check the following properties in each retrieved X.509 certificate: *(a) certificate subject*, *(b) alternative names*, *(c) key usage* (purpose), *(d) certificate chain*, *(e) validity time*, and *(f) stability over time*. If a certificate does not pass any of the tests, we do not consider it in the analysis.

We keep only the IPs that have a certificate subject and alternative names with valid domains and also valid country-code second-level domains (ccSLD) according to the definition in [81]. Next, we check if the key usage explicitly indicates a Web server role. In the certificate chain we check if the delivered certificates do really refer to each other in the right order they are listed up to the root certificate, which must be contained in the current Linux/Ubuntu white-list. Next, we verify the validity time of each certificate in the chain by comparing it to the timestamp the certificate fetching was performed. Lastly, we perform the active measurements several times and check for changes because IPs in cloud deployments can change their role very quickly and frequently. Ignoring validity time, we require that all the certificates fetched from a single IP have the same properties. In the case of our week 45 data, starting with a candidate set of approximately $1.5M$ IPs, some $500K$ respond to repeated active measurements, of which $250K$ are in the end identified as HTTPS server IPs.

When combined, these filtering steps yield approximately $1.5M$ different Web server IPs (including the $250K$ HTTPS server IPs). In total, these HTTP and HTTPS server IPs are responsible for or "see" more than 70 % of the peering traffic portion of the total traffic. Some $350K$ of these IP addresses appear in both sets and are examples of multi-purpose servers; that is, servers with one IP address that see activity on multiple ports. Multi-purpose servers are popular with commercial Internet players (e. g., Akamai which uses TCP port 80 (HTTP) and TCP port 1935 (RTMP)), and their presence in our data partially explains why we see a larger percentage of Web server-related traffic than what is typically reported [85, 113], but is often based on a strictly port-based traffic classification [68, 105].

Among the identified HTTP and HTTPS server IPs, we find some $200K$ IPs that act both as servers and as clients. These are responsible for some 10 % of the server-related traffic. Upon closer inspection of the top contributors in this category, we find that they typically belong to major CDNs (e. g., EdgeCast, Limelight) or network operators (e. g., Eweka). Thus, the large traffic share of these servers is not surprising and reflects typical machine-to-machine traffic associated with operating, for example, a CDN. Another class of IPs in this category are proxies or clients that are managed via a server interface (or vice versa). In the context of this chapter, it is important to clarify the notion of a server IP. Throughout this chapter, a server IP is defined as a publicly routed IP address of a server. As such, it can represent many different real-world scenarios, including a single (multi-purpose) server, a rack of multiple servers, or a front-end server acting as a gateway to possibly thousands of back-end servers (e. g., an entire data center). In fact, Figure 8.2 shows the traffic share of each server IP seen in the week 45 data. It highlights the presence of individual server IPs that are responsible for more than 0.5 % of all server–related traffic! Indeed, the top 34 server IPs are responsible for more than 6 % of the overall server traffic. These server IPs cannot be single machines. Upon closer examination, they are identified as belonging to a cast of Internet players that includes CDNs, large content providers, streamers, virtual backbone providers, and resellers, and thus represent front-end servers to large data centers and/or anycast services. Henceforth, we use the term server to refer to a server IP as defined above.

8.1.3 Available IXP-external Datasets

When appropriate and feasible, we augment our IXP-based findings with active and passive measurements that do not involve the IXP in any form or shape and are all collected in parallel to our IXP data collection. Such complementary information allows us to verify, check, or refine the IXP-based findings.

One example of a complementary IXP-external dataset is a proprietary dataset from a large European tier-1 ISP consisting of packet-level traffic traces.[2] With the help of the network intrusion detection system Bro [129] we produce the HTTP and DNS logs, extract the Web server-related traffic and the corresponding server IPs from the logs, and rely on the resulting data in Section 8.2.

For another example, we use the list of the top $280K$ DNS recursive resolvers—as seen by one of the largest commercial CDNs—as a starting set to find a highly distributed set of DNS resolvers that are available for active measurements such as doing reverse DNS lookups or performing

[2] For this trace we anonymized the client information before applying the analysis with the network intrusion detection system Bro. We always use a prefix preserving function when anonymizing IPs.

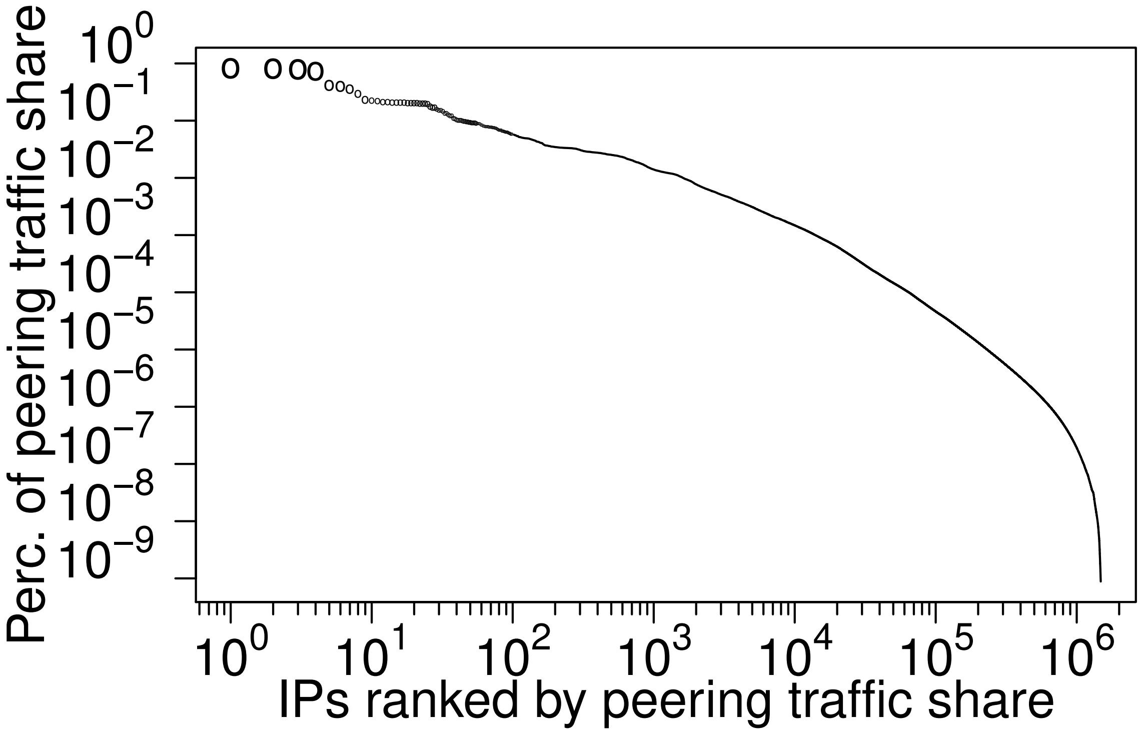


Figure 8.2: Traffic per server IP sorted by traffic share

active DNS queries. From this initial list of DNS servers, we eliminate those that cannot be used for active measurements (i.e., those that are not open, delegate DNS resolutions to other resolvers, or provide incorrect answers) and end up with a final list of about $25K$ DNS resolvers in some $12K$ ASs that are used for active measurements in Section 8.2.3.

Other examples of IXP-external data we use in this work include the publicly available lists of the top-1M or top-1K popular Web sites that can be downloaded from `www.alexa.com`. We obtained these lists for each of the weeks for which we have IXP data. We also utilized blogs and technical information found on the official Web sites of the various technology companies and Internet players. In addition, we make extensive use of publicly available BGP-based data that is collected on an ongoing basis by RouteViews [139], RIPE RIS [31], Team Cymru [34], etc.

8.1.4 IP Server Meta-data

Our efforts in Section 8.4 rely on meta-data that we collect for server IPs and that is obtained from DNS information, URIs, and X.509 certificates handed out by HTTPS servers.

Regarding DNS information, obvious meta-information is the hostname(s) of a server IP. This information is useful because large organizations [88] often follow industrial standards in their naming schema's for servers that they operate or host in their own networks. Another useful piece of meta-data is the Start of Authority (SOA) resource record which relates to the administrative authority and can be resolved iteratively. This way one can often find a common root for organizations that do not use a unified naming schema. Note that the SOA record is often present, even when there is no hostname record available or an ARPA address is returned in the reverse lookup of a server IP.

Next, the URI as well as the authority associated with the hostname give us hints regarding the organization that is responsible for the content. For example, for the URI `youtube.com`, one finds the SOA resource record `google.com` and thus can associate Youtube with Google.

Lastly, the X.509 certificates reveal several useful pieces of meta-data. First, they list the base set of URIs that can be served by the corresponding server IP. Second, some server IPs have certificates with multiple names that can be used to find additional URIs. This is typically the case for hosting companies that host multiple sites on a single server IP. In addition, it is used by CDNs that serve multiple different domains with the same physical infrastructure. Moreover, the names found in the certificates can be mapped to SOA resource records as well.

Overall, we are able to extract DNS information for $71.7\,\%$, at least one URI for $23.8\,\%$, and X.509 certificate information for $17.7\,\%$ of the $1.5M$ server IPs that we see in our week 45 data. For $81.9\,\%$ of all the server IPs, we have at least one of the three pieces of information. For example, for streamers, one typically has no assigned URI, but information from DNS. Before using this rich meta-data in Section 8.4, we clean it by removing non-valid URIs, SOA resource records of the Regional Internet Registries (RIRs) such as `ripe.net`, etc. This cleaning effort reduces the pool of server IPs by less than $3\,\%$.

8.2 Local yet Global

The main purpose of this section is to show that our IXP represents an intriguing vantage point, with excellent visibility into the Internet as a whole. This finding of the IXP's important global role complements earlier observations that have focused on the important local role that this large European IXP plays for the greater geographic region where it is located [43], and we further elaborate here on its dual role as a local and as a global player. Importantly, we also discuss what we can and cannot discern about the Internet as a whole or its individual constituents based on measurements taken at this vantage point. The reported numbers are for the week 45 measurements when the IXP had 452 members that exchanged some 14PB of traffic per day and are complemented by a longitudinal analysis in Section 8.3.

8.2.1 On the Global Role of the IXP

By providing a well-defined set of steps and requirements for establishing peering links between member networks, IXPs clearly satisfy the main reason for why they exist in the first place – keeping local traffic local. To assess the visibility into the global Internet that comes with using a large

		week 45	educated guessesof ground truth
Peering Traffic	IPs	$232,460,635$	unknown $<2^{32}$
	#ASs	$42,825$	approx. $43K$
	Subnets	$445,051$	$450K^{+}$
	countries	242	250
Server Traffic	IPs	$1,488,286$	unknown
	#ASs	$19,824$	unknown
	Subnets	$75,841$	unknown
	Countries	200	250

Table 8.1: IXP summary statistics—week 45

European IXP as a vantage point, we focus on the peering traffic component (see Section 8.1.2) and summarize in Table 8.1 the pertinent results.

First, in this single geographically well-localized facility, we observe during a one-week period approximately a quarter billion unique IPv4 addresses (recall that the portion of native IPv6 traffic seen at this IXP is negligible). While the total number of publicly routed IPv4 addresses in the Internet in any given week is unknown, it is some portion of the approximately three and a half billion allocated IPv4 addresses, which suggests that this IXP "sees" a significant fraction of the ground truth. The global role of this IXP is further illuminated by geo-locating all $230M^{+}$ IP addresses at the country-level granularity [133] and observing that this IXP "sees" traffic from every country of the world, except for places such as Western Sahara, Christmas Islands, or Cocos (Keeling) Islands. This ability to see the global Internet from this single vantage point is visualized in Figure 8.3, where the different countries' shades of gray indicate which percentage of IPs a given country contributes to the IPs seen at this IXP.

Second, when mapping the encountered $230M^{+}$ IP addresses to more network-specific entities such as subnets or prefixes and ASs, we confirm the IXP's ability to "see" the Internet. More precisely, in terms of subnets/prefixes, this IXP "sees" traffic from $445K$ subnets; that is, from essentially all actively routed prefixes. Determining the precise number of actively routed prefixes in the Internet in any given week remains an imprecise science as it depends on the publicly available BGP data that are traditionally used in this context (e. g., RouteViews, RIPE). The reported numbers vary between $450K$-$500K$ and are only slightly larger than the $445K$ subnets we see in this one week. With respect to ASs, the results are very similar – the IXP "sees" traffic from some $42.8K$ actively routed ASs, where the ground truth for the number of actively routed ASs in the Internet in any given week is around $43K$ [9] and varies slightly with the used BGP dataset.

Lastly, to examine the visibility that this IXP has into the more commercial-oriented Internet, we next use the Web server-related component of the IXP's peering traffic (see Section 8.1.2). Table 8.1 shows that this IXP "sees" server-related traffic from some $1.5M$ IPs that can be unambiguously identified as Web server IPs. Unfortunately, we are not aware of any numbers that can

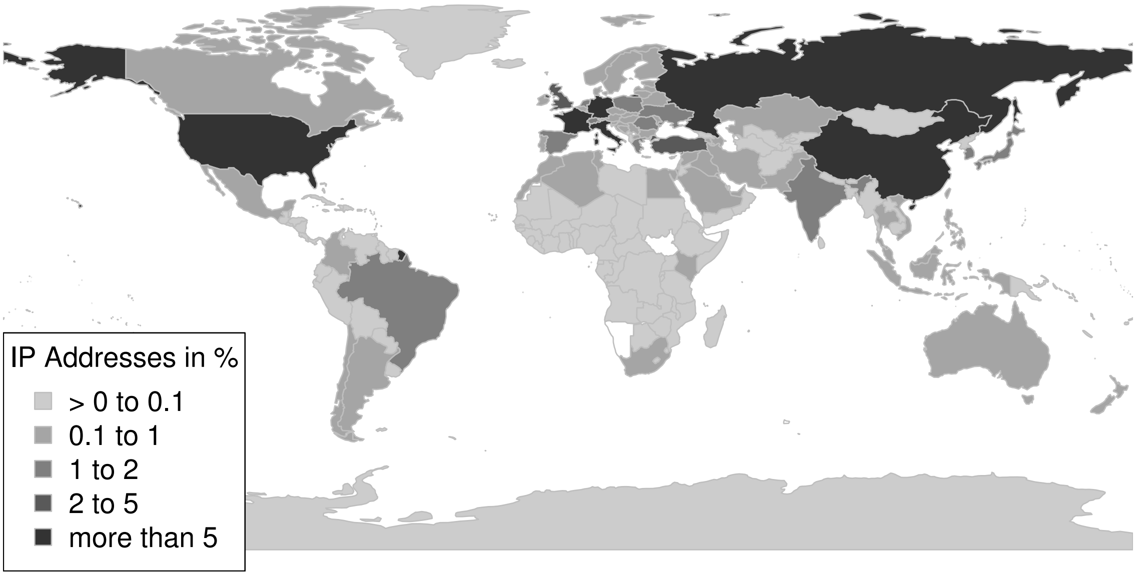


Figure 8.3: Percentage of IPs per country—week 45

be reliably considered as ground truth of all server IPs in the Internet in any given week. Even worse, available white papers or reports that purportedly provide this information are typically very cavalier about their definition of what they consider as "Web server" and hence cannot be taken at face value [100, 141].

To indirectly assess how the roughly $1.5M$ Web server IPs seen at this IXP stack up against the unknown number of Web server IPs Internet-wide, we use an essentially orthogonal dataset, namely the HTTP and DNS logs from a large European tier-1 ISP that does not exchange traffic over the public switching infrastructure of our IXP. Applying the methods as described in Section 8.1, we extract the Web server IPs from this ISP dataset and find that of the total number of server IPs that are "seen" by this ISP, only some $45K$ are not seen at the IXP. Importantly, for the server IPs seen both at the IXP and the ISP, those we identified as server IPs using the IXP-internal data are confirmed to be indeed server IPs when relying on the more detailed ISP dataset. In any case, mapping the $1.5M$ server IPs from the IXP to prefixes, ASs, and countries shows that this IXP "sees" server-traffic from some 17 % of all actively routed prefixes, from about 50 % of all actively routed ASs, and from about 80 % of all the countries in the world.

8.2.2 On the IXP's Dual Role

Visuals such as Figure 8.3 illustrate that by using this IXP as a vantage point, we are able to see peering traffic from every country and corner of the world or from almost every AS and prefix that is publicly routed. However, such figures do not show whether or not certain countries or corners and ASs or prefixes are better visible than others in the sense that they are responsible for

		All IPs	Server IPs	All IPs	Server IPs
	rank	Country	Country	Network	Network
IPs	1	US	DE	Chinanet	Akamai
	2	DE	US	Vodafone/DE	1&1
	3	CN	RU	Free SAS	OVH
	4	RU	FR	Turk Telekom	Softlayer
	5	IT	GB	Telecom Italia	ThePlanet
	6	FR	CN	Liberty Global	Chinanet
	7	GB	NL	Vodafone/IT	HostEurope
	8	TR	CZ	Comnet	Strato
	9	UA	IT	Virgin Media	Webazilla
	10	JP	UA	Telefonica/DE	Plusserver
Traffic	1	DE	US	Akamai	Akamai
	2	US	DE	Google	Google
	3	RU	NL	Hetzner	Hetzner
	4	FR	RU	OVH	VKontakte
	5	GB	GB	VKontakte	Leaseweb
	6	CN	EU	Kabel Deu.	Limelight
	7	NL	FR	Leaseweb	OVH
	8	CZ	RO	Vodafone/DE	EdgeCast
	9	IT	UA	Unitymedia	Link11
	10	UA	CZ	Kyivstar	Kartina

Table 8.2: Top 10 contributors—week 45

more traffic that is exchanged over the public switching fabric of the IXP. In particular, we would like to know whether the importance of the local role that this IXP plays for the larger geographic region within which it is situated is more or less recovered when considering the peering or server-related traffic that the IPs or server IPs are responsible for, respectively. To this end, we show in Table 8.2 the top-10 countries in terms of percentage of IP addresses (and associated traffic) and percentage of server IPs (and associated traffic). In addition, we show the top-10 networks. While the role of the IXP for the European region becomes more dominant when we change from peering to server-related traffic, there are still prominent signs of the IXP's global role, even with respect to the commercial Internet, and they reflect the relative importance of this IXP for countries such as USA, Russia, and China or ASs such as AS20940 (Akamai), AS15169 (Google), and AS47541 (VKontakte).

		Member AS	Distance 1	Distance > 1
		$A(L)$	$A(M)$	$A(G)$
Peering Traffic	IPs	42.3 %	45.0 %	12.7 %
	Prefixes	10.1 %	34.1 %	55.8 %
	ASs	1.0 %	48.9 %	50.1 %
	Traffic	67.3 %	28.4 %	4.3 %
Server Traffic	IPs	52.9 %	41.2 %	5.9 %
	Prefixes	17.2 %	61.9 %	20.9 %
	ASs	2.2 %	61.5 %	36.3 %
	Traffic	82.6 %	17.35 %	0.05 %

Table 8.3: IXP as local yet global player—week 45

For a somewhat simplified illustration of the IXP's dual role as a local as well as global player, we divide the set of all actively routed ASs into three disjoint sets, $A(L)$, $A(M)$, and $A(G)$. $A(L)$ consists of the member ASs of the IXP; $A(M)$ consists of all ASs that are distance 1 (measured in AS-hops) from a member AS; and $A(G)$ is the complement of $A(L) \cup A(M)$ and contains those ASs that are distance 2 or more from the member ASs. Intuitively, the set $A(L)$ captures the importance of the local role of the IXP, whereas the set $A(G)$ is more a reflection of the IXP's global role, with A(M) covering some middle ground. Table 8.3 shows the breakdown of the IPs, prefixes, and ASs for peering traffic and Web server-related traffic, respectively, for the three sets. It basically confirms our above observation that there is a general trend towards the set $A(L)$ as we move from IPs and the peering traffic they are responsible for to server IPs and their traffic. Note, while the relative importance of the IXP's local role over its global role with respect to the commercial Internet (i.e., server-related traffic) makes economic sense and is well-captured by this cartoon picture, in reality, there is potentially significant overlap between the sets $A(L)$, $A(M)$, and $A(G)$, e. g., due to remote peerings, IXP resellers, and non-European networks joining the IXP for purely economic reasons. But this is unlikely to invalidate our basic findings concerning the IXP's dual role.

8.2.3 On the IXP's "Blind Spots"

While the IXP "sees" traffic from much of the Internet, taking measurements exclusively at this single vantage point can tell us only so much about the network as a whole or its individual constituents. Hence, knowing what we can discern about the network with what sort of accuracy is as important as understanding what we cannot discern about it, and why.

We show in Section 8.2.1 how the use of an essentially orthogonal IXP-external dataset (i. e., the HTTP and DNS logs from the large European tier-1 ISP) enables us to indirectly assess how the approximately $1.5M$ server IPs seen at the IXP in a given week compare to the unknown number of server IPs network-wide. In the following, we discuss additional examples where the use of IXP-external data, either in the form publicly available measurements, active or passive measurements, or proprietary information, enables us to check, validate, or refine what we can say with certainty when relying solely on IXP measurements.

To examine in more detail how the approximately $1.5M$ server IPs seen at the IXP in a given week compare to all server IPs in the Internet, we now use a more extensive combination of IXP-external measurements. To start, using the list of the top-1M Web sites available from www.alexa.com and based on the URIs retrieved from the limited payload part of the sampled frames at the IXP, we recover about 20 % of all the second-level domains on Alexa's top-1M list of sites; this percentage increases to 63 % if we consider only the top-10K list and to 80 % for the top-1K. Note that many hostnames on these lists are dynamic and/or ephemeral. Next, to assess how many additional server IPs we can identify using the approximately 80 % of domains we cannot recover using the URIs seen at the IXP, we rely on active measurements in the form of DNS queries to those uncovered domains using our set of $25K$ DNS resolvers across $12K$ ASs. From this pool of resolvers, we assign 100 randomly-selected resolvers to each URI. This results in approximately $600K$ server IPs, of which more then $360K$ are already seen at the IXP and identified as servers.

To provide insight into the remaining $240K$ server IPs that are not seen as a server at the IXP, we classify them into four distinct categories. First, there are servers of CDNs that are hosted inside an AS and serve exclusively clients in that AS ("private clusters"). These servers reply only to resolvers of that AS for content that is delivered by the global footprint of those CDNs. Traffic to these servers should not be observable at the IXP as it should stay internal to the AS. Second, there are servers of CDNs or cloud/hosting providers that are located geographically far away from the IXP. If these networks have a global footprint and distribute content in a region-aware manner, it is unlikely that these server IPs are seen at the IXP. The third group includes servers that some ASs operate for the sole purpose of handling invalid URIs. Finally, the last category contains those servers of small organizations and/or universities in parts of the world that are geographically far away from the IXP. These IPs are typically not visible at the IXP. In terms of importance, the first two categories account for more than 40 % of the $240K$ servers not seen at the IXP.

For a concrete example for illustrating "what we know we don't know", we consider Akamai. In our week-long IXP dataset, we observe some $28K$ server IPs for Akamai in 278 ASs (for details, see Section 8.4). However, Akamai publicly states that it operates some $100K$ servers in more than $1K$ ASs [47]. The reasons why we cannot see this ground truth relying solely on our IXP-internal data are twofold and mentioned above. First Akamai is known to operate "private clusters" in many third-party networks which are generally not visible outside those ASs and therefore cannot be detected at the IXP. Second, we cannot expect to uncover Akamai's footprint in regions that are geographically far away from the IXP, mainly because Akamai uses sophisticated mechanisms to localize traffic [108, 124]. Akamai's large footprint makes discovering all of its servers difficult, but by performing our own diligently chosen IXP-external active measurements [93] that utilize the URIs collected in the IXP and the open resolvers discussed in Section 8.1.3, we were able to discover about $100K$ servers in 700 ASs. Thus, even for a challenging case like Akamai, knowing what our IXP-internal data can and cannot tell us about its many servers and understanding the underlying reasons is feasible.

Regarding our assumption that server-related traffic is a good proxy for the commercial portion of Internet traffic, there are clearly components of this commercial traffic that are not visible at the IXP. For example, the recently introduced hybrid CDNs (e. g., Akamai's NetSession [42]) serve content by servers as well as by end users that have already downloaded part of the content. Since the connections between users are not based on a HTTP/HTTPS server-client architecture but are P2P-based, we may not see them at the IXP. However, while the traffic of these hybrid CDNs is increasing (e. g., the service is mainly used for large files such as software downloads), the overall volume is still very low [74].

Lastly, by the very definition of an IXP, any traffic that does not pass through the IXP via its public-facing switching infrastructure remains entirely invisible to us. For example, this includes all traffic that traverses the IXP over private peering links. IXPs keep the private peering infrastructure separate from its public peering platform, and we are not aware of any kind of estimates of the amount of private peering traffic handled by the IXPs.

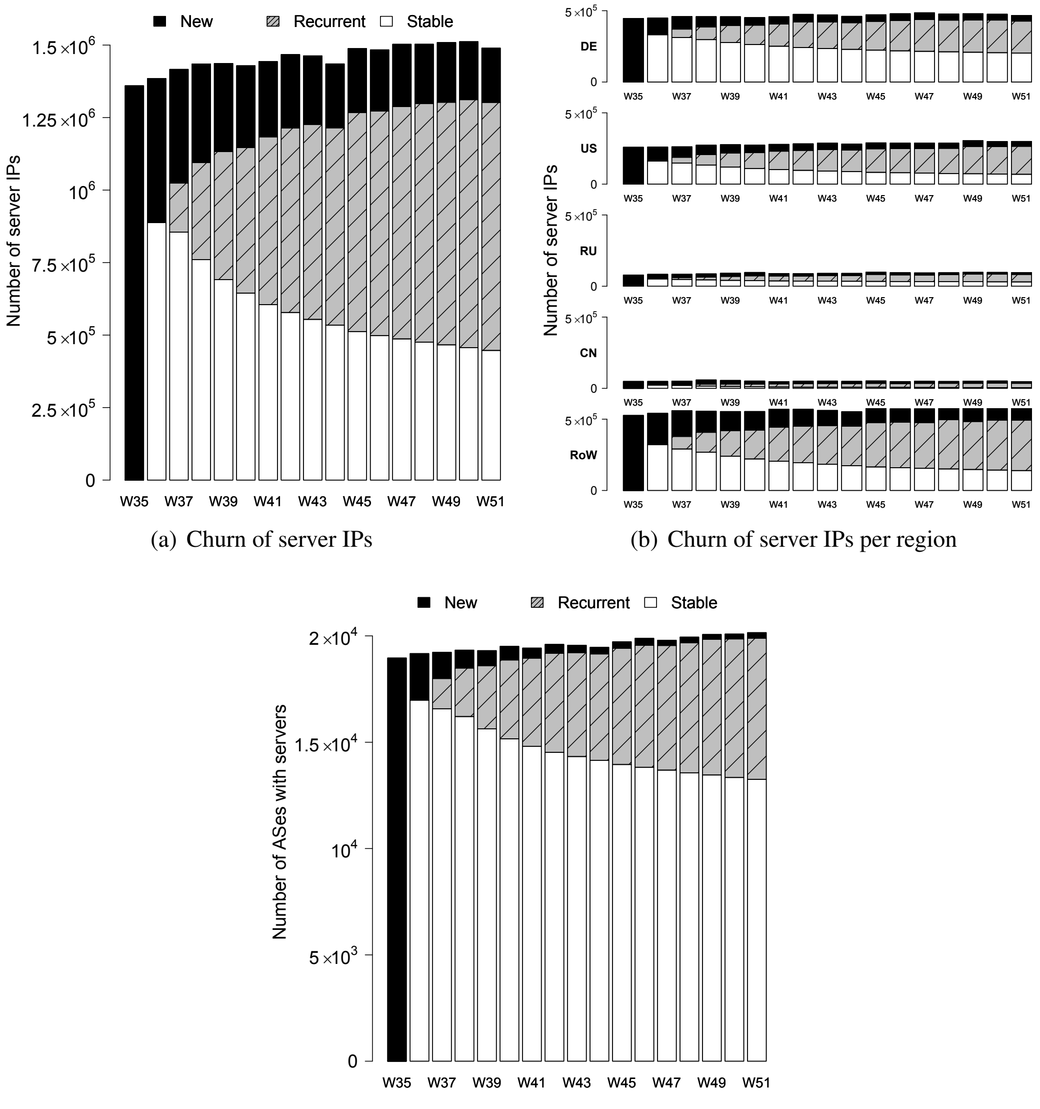
8.3 Stable yet Changing

In this section, we report on a longitudinal analysis that covers 17 consecutive weeks and describes what using our large IXP as a vantage point through time enables us to say about the network as whole, about some of its constituents, and about the traffic that these constituents are responsible for.

8.3.1 Stability in the Face of Constant Growth

Publicly available data shows that during 2012, this IXP has experienced significant growth, increasing the number of member ASs by 75 and seeing the average daily traffic volume grow by $0.1\,\%$. In terms of absolute numbers, we see in week 35 a total of 443 IXP member ASs sending an average daily traffic volume of 11.9PB over the IXP's public-facing switching infrastructure. By week 51, the member count stood at 457, and the average traffic volume went up to 14.5PB per day. For what follows, it is important to note that these newly added member ASs are typically regional and local ISPs or organizations and small companies outside of central Europe for which membership at this IXP makes economic sense. To contrast, all the major content providers, CDNs, Web hosting companies, eye-ball ASs, and tier-1 ISPs have been members at this IXP for some time, but may have seen upgrades to higher port speeds since the time they joined.

Given our interest in the commercial Internet and knowing (see Section 8.1) that the server-related traffic is more than $70\,\%$ of the peering traffic seen at the IXP, we focus in the rest of this chapter on the server-related portion of the IXP traffic. The initial set of findings from our longitudinal analysis paints an intriguingly stable picture of the commercial Internet as seen from our vantage point. In particular, analyzing in detail each of the 17 weekly snapshots shows that during every week, we see server-related traffic at this IXP from about $20K$ (i. e., about half of all) actively



(a) Churn of server IPs

(b) Churn of server IPs per region

(c) Churn of ASs with servers

Figure 8.4: Churn of server IPs and ASs that host servers—weeks $35 - 51$

routed ASs, some $75K$ or approximately 15 % of all actively routed prefixes, and from a pool of server IPs whose absolute size changes only so slightly but tends to increase in the long term.

This last property is illustrated in Figure 8.4(a) when focusing only on the absolute heights of the different bars that represent the total number of server IPs seen in a given week. When considering the full version of this figure, including the within-bar details, Figure 8.4(a) visualizes the weekly churn that is inherent in the server IPs seen at the IXPs. To explain, the first bar in Figure 8.4(a) shows the approximately $1.4M$ unique server IPs that we see in week 35. The next bar shows that same quantity for week 36, but splits it into two pieces. While the lower (white) piece reflects the portion of all week 36 server IPs that were already seen during week 35, the upper (black) piece represents the set of server IPs that were seen for the first time during week % 36. Starting with week 37, we show for each week $n \in \{37,38,\ldots,51\}$ snapshot a bar that has three pieces stacked on top of one another. While the first (bottom, white) piece represents the server IPs that were seen at the IXP in each one of the week k snapshot $(k = 35,36,\ldots,n)$, the second (grey-shaded) piece shows the server IPs that were seen at the IXP in at least one previous week k snapshot $(k = 35,36,\ldots,n-1)$, but not in all; the third (top black) piece represents all server IPs that were seen at the IXP for the first time in week n.

A key take-away from Figure 8.4(a) is that there is a sizable pool of server IPs that is seen at the IXP during each and every week throughout the 17-week long measurement period. In fact, this stable portion of server IPs that is seen at the IXP week-in and week-out is about 30 % as can be seen by looking at the bottom (white) portion of the week 51 bar. Instead of requiring for a server IP to be seen in each and every week, we also consider a more relaxed notion of stability called recurrence. This recurrent pool of server IPs consists of all server IPs that, by week 51, have been seen at the IXP during at least one previous week (but not in each and every previous week), is represented by the grey-shaded portion of the week 51 bar, and consists of about 60 % of all server IPs seen in week 51. Note that the number of server IPs seen for the first time in week n (top black portion) decreases over time and makes up just about 10 % of all server IPs seen in week 51.

To look in more detail at the stable and recurrent pools of server IPs and examine their churn or evolution during the 17-week long measurement period, we rely on the GeoLite Country database [120] to geo-locate the server IPs to the country level and group them by geographic "region" as follows: DE, US, RU, CN, RoW (rest of world). Figure 8.4(b) is similar to Figure 8.4(a), but shows for each week the portions of IPs for each of these five regions and visualizes the make-up of these server IPs in the same way as we did in Figure 8.4(a). Note that the shown region-specific stable portions in week 51 add up to the 30 % number observed in Figure 8.4(a), and similarly for the region-specific recurrent portions in week 51 (their sum adds up to the roughly 60 % portion of the recurrent pool shown in Figure 8.4(a)). Interestingly, while the stable pool for DE is consistently about half of the overall stable pool of server IPs seen at the IXP, that pool is vanishing small for CN, slightly larger for RU. This is yet another indication of the important role that this IXP plays for the European part of the Internet.

An even more intriguing aspect of stability is seen when we consider the server-related traffic that the server IPs that we see at the IXP are responsible for. For one, we find that the stable pool of server IPs is consistently contributing more than 60 % of the server-related traffic. That is, of the server IPs that this IXP "sees" on a weekly basis, more than 30 % of them are not only seen week

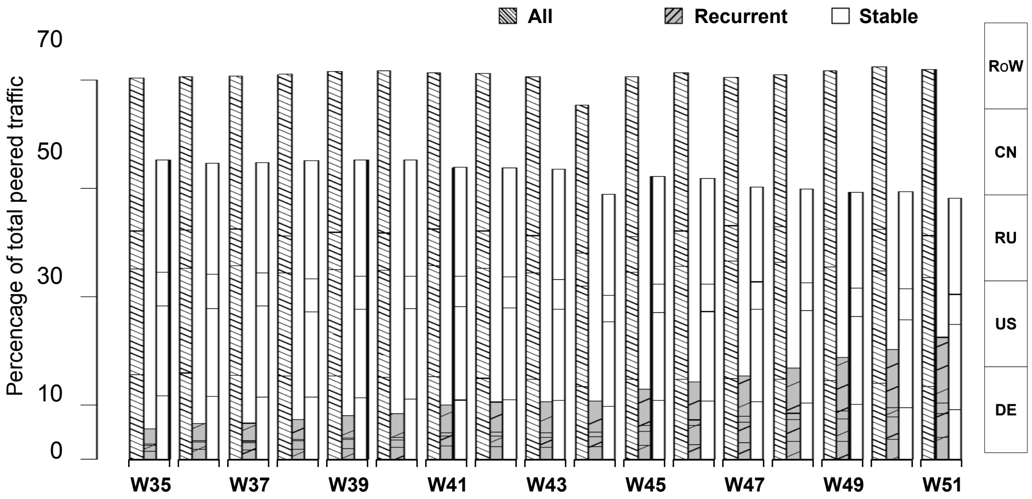


Figure 8.5: Churn of server traffic by region—weeks $35 - 51$

after week, but they are also responsible for most of the server-related traffic that traverses the IXP each week. When considering the weekly recurrent pools of server IP (grey-shaded segments in Figure 8.4(a)), their traffic portions keep increasing, but only to less than 30 % of all server traffic. To examine the make-up of the server-related traffic attributed to the stable and recurrent pools of server IPs, respectively, Figure 8.5 shows for each week n three bars, each with five segments corresponding to the five regions considered earlier. The first bar is for the server-related traffic portion of all peering traffic that all server IPs see at the IXP in week n; the second bar reflects the server-related traffic portion in week n attributed to the recurrent pool of server IPs in that week, while the third bar shows the server-related traffic portion in week n that the stable pool of server IPs is responsible for. From Figure 8.5, we see that while the stable and recurrent pools of server IPs from China are basically invisible at the IXP in terms of their traffic, both US and Russia have the property that the stable pool of server IPs is responsible for much all the server-related traffic seen from those regions at the IXP.

In addition to examining the churn and evolution of the server IPs seen at the IXP, it is also instructive to study the temporal behavior of the subnets and ASs that the encountered server IPs map into. To illustrate, we only consider the ASs and show in Figure 8.4(c) for ASs what we depicted in Figure 8.4(b) for server IPs. The key difference between server IPs and ASs is that the stable pool of ASs represented by the white portion of the week 51 bar is about 70 % compared to the 30 % for the stable pool of server IPs. Thus, a majority of ASs with server IPs is seen at the IXP during each and every week, and the number of ASs that are seen for the first time becomes miniscule over time. In summary, the stable pool of server IPs (about 1/3 of all server IPs seen at the IXP) gives rise to a stable pool of ASs (about 2/3 of all ASs seen at the IXP and have server IPs) and is responsible for much of the server-related traffic seen at the IXP.

8.3.2 Changes in Face of Significant Stability

One benefit of observing a significant amount of stability with respect to the server-related portion of the overall peering traffic seen at the IXP is that any weekly snapshot provides more or less the same information. At the same time, subsequent weekly snapshots that differ noticeably may be an indication of some change. Next, we briefly discuss a few examples of such changes that we can discern about the Internet as a whole and some of its individual constituents when we have the luxury to observe and measure the network at this IXP for a number of consecutive weeks.

The first example is motivated primarily by our ability described in Section 8.1.2 to specifically look for and identify HTTPS server IPs, but also by anecdotal evidence or company blogs [57, 58] that suggest that due to widespread security and privacy concerns, the use of HTTPS is steadily increasing. To examine this purported increase, we extract for each weekly snapshot all HTTPS server IPs and the traffic that they contribute. When comparing for each week the number of HTTPS server IPs relative to all server IPs seen in that week and the weekly traffic associated with HTTPS server IPs relative to all peering traffic, we indeed observe a small, yet steady increase, which confirms that the Internet landscape is gradually changing as far as the use of HTTPS is concerned.

For a different kind of example for using our IXP vantage point, we are interested in tracking the recently announced expansion of Netflix using Amazon's EC2 cloud service [40] into a number of Scandinavian countries [59]. To this end, we relied on publicly available data to obtain Amazon EC2's data center locations [49] and the corresponding IP ranges [50]. We then mined our 17 weeks worth of IXP data and observed for weeks 49, 50, and 51 a pronounced increase in the number of server IPs at Amazon EC2's Ireland location, the only data center of Amazon EC2 in Europe. This was accompanied by a significant (but still small in absolute terms) increase in Amazon EC2's traffic. All this suggests that it may be interesting to watch this traffic in the future, especially if the observed changes are in any way related to Netflix becoming available in Northern Europe towards the end of 2012. Yet another example concerns the detection of regional or national events at this IXP. For example, considering in more detail week 44, which shows up as a clear dip in, say, Figure 8.4(a), we notice that this week coincides with Hurricane Sandy that had a major impact on the US East Coast region. To examine its impact, we use the IXP vantage point to discern this natural disaster from traffic that we see at the IXP from a particular Internet constituent, a major cloud provider. Using publicly available information about the cloud platform's data centers and corresponding IP ranges, we look in our data for the corresponding server IPs and find a total of about $14K$. A detailed breakdown by data center location for weeks $43 - 45$ shows a drastic reduction in the number of server IPs seen at the IXP from the US East Coast region, indicating that the platform of this major cloud provider faced serious problems in week 44, with traffic dropping close to zero. These problems made the news, and the example highlights how a geographical distant event such as a hurricane can be discerned from traffic measurements taken at this geographically distant IXP.

Lastly, we also mention that an IXP is an ideal location to monitor new players such as "resellers". Resellers are IXP member ASs, and their main business is to provide and facilitate access to the IXP for smaller companies that are typically far away geographically from the IXP. For IXPs, the emergence of resellers is beneficial as they extend the reach of the IXP into geographically distant regions and thereby the potential membership base. For example, for a particular reseller at our IXP, we observed a doubling of the server IPs from $50K$ to $100K$ in four months, suggesting that this reseller has been quite successful in attracting new networks with significant server-based infrastructures as its customers.

8.4 Beyond the AS-level View

To illustrate the benefits and new opportunities for Internet measurements that arise from being able to use our IXP as a vantage point with very good visibility into the Internet, we describe in this section an approach for identifying server-based network infrastructures and classifying them by ownership. In the process, we report on a clear trend towards more heterogeneous networks and network interconnections, provide concrete examples, and discuss why and how this observed network heterogenization requires moving beyond the traditional AS-level view of the Internet.

8.4.1 Alternative Grouping of Server-IPs

To this point, we have followed a very traditional approach for looking at our IXP data in the sense that we measured the IXP's visibility into the Internet in terms of the number of actively routed ASs or subnets seen at the IXP. However, there exist Internet players (e. g., CDN77, a recently launched low-cost no-commitment CDN; Rapidshare, a one-click hosting service; or certain meta-hosters that utilize multiple hosters) that are not ASs in the sense that they do not have an assigned ASN. Thus, as far as the traditional AS-level view of the Internet is concerned, these players are invisible, and the traffic that they are responsible for goes unnoticed, or worse mis-attributed to other Internet players. Yet, being commercial entities, these companies actively advertise their services, and in the process often publish the locations and IP addresses of their servers. This then suggests an alternative approach to assessing the IXP's ability to "see" the Internet as a whole—group servers according to the organization or company that has the administrative control over the servers and is responsible for distributing the content. While this approach is easy and works to perfection for companies like CDN77 that publish all their server IPs, the question is what to do if the server IPs are not known.

Accordingly, our primary goal is to start with the server IPs seen at the IXP and cluster them so that the servers in one and the same cluster are provably under the administrative control of the same organization or company. To this end, we rely in parts on methods described by Plonka et al. [131] for traffic and host profiling, Bermudez et al. [55] for discerning content and services, and Ager et al. [45] for inferring hosting infrastructures from the content they deliver. We also take advantage of different sets of meta-data obtained from assorted active measurement efforts or available by other means as discussed in Section 8.1.4. Recall that this meta-data includes for every server IP seen in the IXP data the corresponding URIs, the DNS information from active measurements, and, where available, the list of X.509 certificates retrieved via active measurements. In the rest of this section the reported numbers are for week 45.

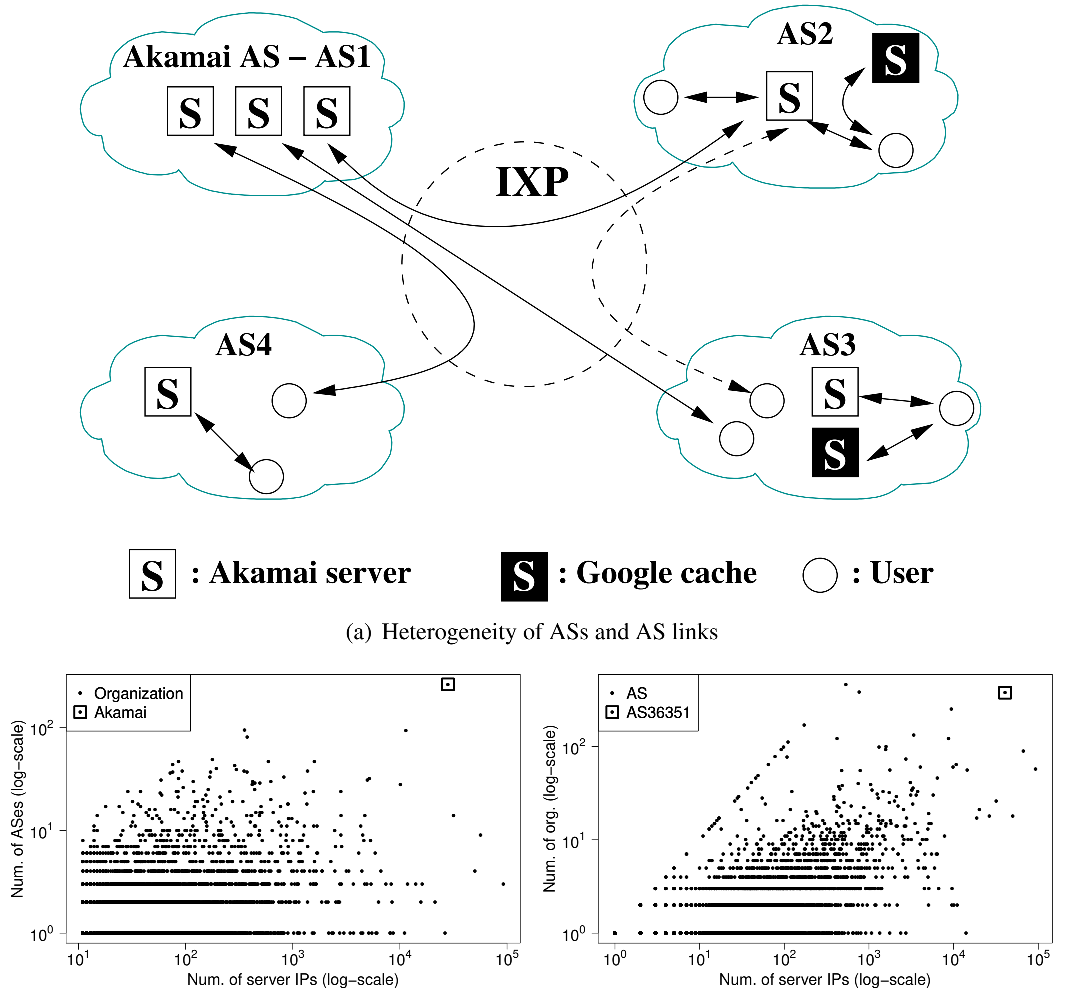
The clustering proceeds in three steps. First, we focus on those server IPs for which we have a SOA resource record and consider a first category of clusters that have the property that all server IPs assigned to a given cluster have the IP and the content managed by the same authority. We identify those clusters by grouping all server IPs where the SOA of the hostname and the authority of the URI lead to the same entry. Prominent organizations that fall into this first category are Amazon and big players like Akamai and Google when they are located in their own ASs or when they are in third-party ASs but have assigned names to their own servers. 78.7 % of all our server IPs are clustered in this first step.

In a second step, we consider clusters with the property that for the server IPs in a given cluster, most of the server IPs and most of the content are managed by the same authority. This can happen if the SOA is outsourced (e. g., to a third-party DNS provider) and is common property among hosters and domains served by virtual servers. In these cases, to group server IPs, we rely on a majority vote among the SOA resource records, where the majority vote is by (i) the number of IPs and (ii) the size of the network footprint. This heuristic enables us to group some server IPs together with organizations inferred in the previous step and also applies to meta-hosters such as Hostica. 17.4 % of all our server IPs are clustered in this second step. Lastly, for the remaining 3.9 % of server IPs that have been seen in our IXP data and have not yet been clustered, we only have partial SOA information. This situation is quite common for parts of the server infrastructure of some large content providers and CDNs such as Akamai that have servers deployed deep inside ISPs. In this case, we apply the same heuristic as in the second step, but only rely on the available subset of information.

To validate our clustering that results from this three-step process, we manually compare the results by (1) checking against the coverage of the public IP ranges that some organizations advertise (see Section 8.3.2), (2) utilizing the information of certificates that point to applications and services, and (3) actively downloading either the front page (e. g., in the case of Google, it is always the search engine front page) or requested content that is delivered by a CDN (e. g., in the case of Akamai, any content is delivered by any of its servers [150]). Our method manages to correctly identify and group the servers of organizations with a small false-positive rate of less than 3 %. Moreover, we observe that the false-positive rate decreases with increasing size of the network footprint. However, there are false-negatives in the sense that our methodology misses some servers due to the “blind spots” discussed in Section 8.2.3.

8.4.2 New Reality: ASs are Heterogeneous

Equipped with an approach for grouping server IPs by organizations, we examine next to what extent this grouping is orthogonal to the prevailing AS-level view of the Internet. The issues are succinctly illustrated in Figure 8.6(a) where we augment the traditional AS-level view (i. e., a number of different ASs exchanging traffic over (public) peering links at an IXP) with new features in the form of AS-internal details (i. e., the third-party servers that the ASs host). Note that while the traditional view that makes a tacit homogeneity assumption by abstracting away any AS-internal details may have been an adequate model for understanding some aspects of the Internet and the traffic it carries at some point in the past, things have changed, and we assert



(a) Heterogeneity of ASs and AS links

(b) Scatter plot of number of server IPs vs. the number of ASs per organization

(c) Scatter plot of number of organizations vs. the number of server IPs for each AS

Figure 8.6: Heterogeneity of organizations and ASs

that the cartoon picture in Figure 8.6(a) captures more accurately the current Internet reality; that is, a trend towards distributed network infrastructures that are deployed and operated by today's commercial Internet players.

To quantify how much closer the cartoon Figure 8.6(a) is to reality than the traditional AS-level view, we apply our clustering approach to the $1.5M$ server IPs seen in our week 45 IXP data and obtain some $21K$ clusters, henceforth referred to organizations or companies. Among them are the well-known big players like Akamai with $28K$ active server IPs, Google with $11.5K$ server IPs, and several large hosters, each with more than $50K$ server IPs (e. g., AS92572 with $90K^+$ server IPs; AS56740 and AS50099, both with more than $50K$ server IPs). Indeed, of the $21K$ identified organizations, a total of 143 organizations are associated with more than 1000 server IPs and more than $6K$ organizations have more than 10 servers IPs. For the latter, Figure 8.6(b) shows a scatter plot of the number of server IPs per organization vs. the number of ASs that they cover. More precisely, every dot in the plot is an organization, and for a given organization, we show the number of its server IPs (x-axis) and the number of ASs that host servers from that organization (y-axis).[3] We observe that operating a highly diverse infrastructure is commonplace in today's Internet and is not limited to only the Internet's biggest players, but reporting on the bewildering array of scenarios we encountered when examining the extent of the different organizations and the networks they partner with is beyond the scope of this chapter.

The realization that many organizations operate a server infrastructure that is spread across many ASs implies that the complementary view must be equally bewildering in terms of diversity or heterogeneity. This view is captured in Figure 8.6(a) by focusing on, say AS1, and examining how many third-party networks host some of their servers inside that AS. Thus, yet another way to quantify how much closer that cartoon figure is to reality than the traditional AS-level view with its implicit homogeneity assumption concerning the administrative authority of servers hosted within an AS is shown in Figure 8.6(c). Each dot in this figure represents an AS, and the number of organizations a given AS hosts is given on the y-axis while the number of identified server IPs is shown on the x-axis. As before, the figure only shows organizations with more than 10 servers. We observe that many ASs host a sizable number of server IPs that belong to many organizations; there are more than 500 ASs that host servers from more than five organizations, and more than 200 ASs that support more than 10 organizations.

Indeed, this observation is again fully consistent with public announcements [79, 26] and content providers' efforts [23] to install their own brand of single-purpose CDNs inside various ISPs. The end effect of such developments is a clear trend towards more heterogeneous eye-ball ISP networks by virtue of such ASs hosting more servers from an increasing number of interested third-party networks. In view of similar announcements from key companies such as Google [146, 64, 80], Amazon [4], or Facebook [22], the challenges of studying, leave alone controlling, such increasingly intertwined networks and traffic are quickly becoming daunting. As an example, consider a large Web hosting company (AS36351), for which we identified more than $40K$ server

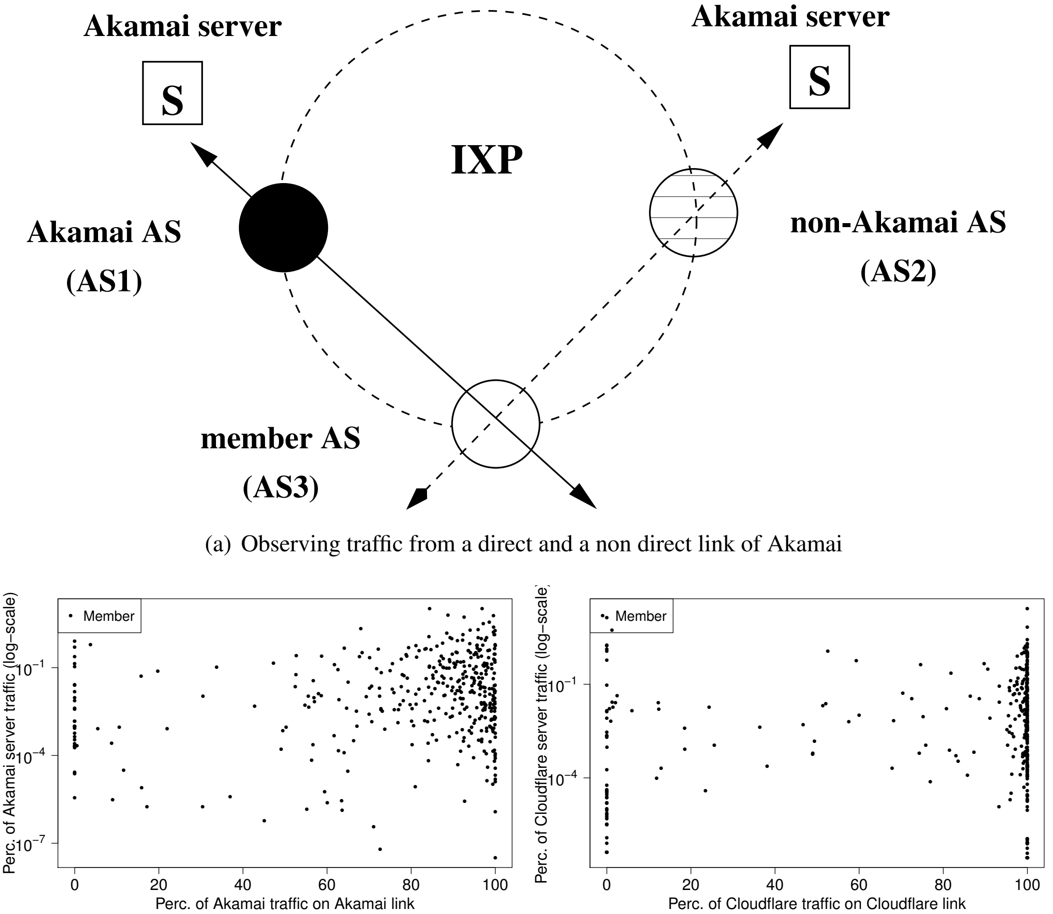
[3] While in a few isolated cases, the ASs that host servers from a given organization are part of that organization (e. g., see [63]), hand-checking the 143 organizations with more than 1000 servers confirmed that in almost all cases, these ASs are genuine third-party networks that are run and operated independently from the organization whose servers they host.

IPs belonging to a total more than 350 different organizations (highlighted in Figure 8.6(c) with a square).

8.4.3 New Reality: Links are Heterogeneous

In Section 8.4.2, we show that organizations take advantage of network diversity and purposefully spread their infrastructure across multiple networks. This development creates very fluid and often transparent network boundaries, which in turn causes havoc when trying to attribute the right traffic to the right network. The issues are illustrated in the cartoon Figure 8.7(a). The figure shows the traditional AS-level perspective, whereby Akamai is a member AS (AS1) of this IXP, and so are a generic AS3 and another generic (non-Akamai) AS2, and the Akamai AS peers at this IXP with AS3 which, in turn, peers also with AS2. This traditional AS perspective is enriched with member-specific details that specify that there is an Akamai server behind/inside (non-Akamai) AS2 and behind/inside the Akamai AS. Note that in terms of the traditional AS-level view, the question of how much Akamai traffic is seen at this IXP is clear-cut and can be simply answered by measuring the traffic on the peering link between AS3 and the Akamai AS. However, when accounting for the fact that there is an Akamai server behind/inside the non-Akamai member AS2, answering that same question becomes more involved. It requires measuring the traffic on the (Akamai) peering link between AS3 and the Akamai AS as well as accounting for the Akamai traffic on the (non-Akamai) peering link between AS3 and (non-Akamai) AS2.

Clearly, for accurately attributing traffic to the responsible parties in today's network, the trend towards network heterogenization creates problems for the traditional AS-level view of the Internet. To illustrate the extent of these problems, we show in Figure 8.7(b) what we observe at the IXP for Akamai. Recall that Akamai (AS20940) is a member of the IXP and peers with some 400 other member ASs. In the traditional view, accounting for Akamai traffic traversing the IXP simply means capturing the traffic on all the peering links between Akamai and those member ASs. Unfortunately, this simple view is no longer reflecting reality when Akamai servers are hosted inside or "behind" (non-Akamai) IXP member ASs. To capture this aspect, Figure 8.7(b) shows for each IXP member that peers with Akamai (indicated by a dot) the percentage of Akamai traffic on the direct peering link to Akamai (x-axis) vs. the percentage of total Akamai-server traffic for this member AS (y-axis). Under the traditional assumption, all dots would be stacked up at $x = 100$, reflecting the fact that to account for Akamai-related traffic, all that is needed is to measure the Akamai peering links. However, with Akamai servers being massively deployed in third-party networks, including many of the other member ASs of the IXP, we observe that some members get all their Akamai-related traffic from ASs other than the (member) Akamai AS ($x = 0$), even when that traffic is sizable ($y >> 0$). Moreover, the scattering of dots across Figure 8.7(b) succinctly captures the diverse spread of traffic across the direct peering link vs. the other member links. In terms of numbers, Akamai sends $11.1\,\%$ of its traffic not via its peering links with the member AS. Put differently, traffic from more than $15K$ out of the $28K$ Akamai servers that we identified in our IXP data is seen at the IXP via non-IXP member links to Akamai. The same holds true for other major CDNs but also for relatively new players such as CloudFlare. Figure 8.7(c) shows the same kind of plot as Figure 8.7(b) for CloudFlare. It demonstrates that despite adhering to



(a) Observing traffic from a direct and a non direct link of Akamai

(b) Perc. of Akamai traffic vs. perc. of Akamai traffic via direct link

(c) Perc. of CloudFlare traffic vs. perc. of CloudFlare traffic via direct link

Figure 8.7: AS link heterogeneity: Traffic via direct member link relative to other member links.

very different business models (i. e., Akamai deploys servers inside ISPs vs. CloudFlare operates its own data centers), the two CDNs have similar usage patters as far as their peering links are concerned.

Looking beyond Akamai, we observe that different services from the same organization use their servers differently resulting in different usage patterns of the peering links. For example, for Amazon CloudFront, Amazon's "CDN part", almost all traffic is send via the IXP's Amazon links. However, for Amazon EC2, the "cloud part", a sizable fraction comes via other IXP peering links. We also noticed that for most cases where we see the use of the non-IXP member links, the percentage of traffic in those links increases during peak times. This may be due to reasons such as load balancing, performance improvement, or cost savings. Lastly, how our view of the usage of the IXP's public peering links is impacted by private peerings that may be in place between member ASs of the IXP remains unexplored.

9 Peering at Peerings

The Border Gateway Protocol (BGP) is an information-hiding protocol. This has come in full view with the recently reported independent discovery by [43] and [87] that there are easily an order of magnitude more peering links in today's Internet compared to what earlier studies reported. Moreover, part of this surprising finding was the realization that the vast majority of these newly discovered peerings are difficult to detect because they are established at the 350^+ Internet eXchange Points (IXP) that are currently in operation world-wide.

As discussed in more detail by [67], one of the main reasons for these surprises has been the fact that many IXPs have started to offer the use of their *route server (RS)*[1] as a free value-added service to their members. In essence, using an IXP's RS greatly simplifies routing for its members – a member AS has to set up just a single BGP session to connect to possibly hundreds of other member ASs. That is, most members use an IXP's RS to establish *multi-lateral* peerings (i. e., one-to-many) as compared to *bi-lateral* peerings (i. e., one-to-one). At the same time, the increasing deployment and use of RSs by IXPs has had the effect of hiding much of the information concerning those members' peering policies from the traditionally collected and widely-used public BGP measurements (e. g., RIPE RIS, Route-views). As a result, trying to understand and, more importantly, reason about who is peering with whom at which IXP (and how and why) has become increasingly difficult.

The main contribution of this chapter is an original analysis that combines a detailed control plane view of IXP-related peerings with a corresponding rich data plane view to study the operations of IXPs and reason about the different peering decisions that networks make when they connect at today's IXPs. To perform this analysis, we rely on our ongoing collaboration with three European IXPs. Two of the IXPs operate a RS and gave us access to a wealth of RS-specific BGP data, and all three IXPs made the traffic datasets that they routinely collect from their public switching infrastructures available to us (see Chapter 6).

What distinguishes this study from prior work in this area is that it is the first work that correlates IXP-provided control plane measurements in the form of RS-specific BGP data with IXP-provided data plane measurements in the form of SFLOW data to scrutinize IXP-peerings beyond their mere existence. To illustrate, the papers [43] and [87] and related studies (e. g., [98, 111]) are all about connectivity (e. g., existence of peerings, number of peerings at an IXP or Internet-wide). In fact,

[1] Note, while the term "route server" often refers to a "route collector/monitor" or "Route-views server", we use it exclusively to refer to an IXP route server [91, 95] as defined in Section 9.1.2.

while [43] relied on proprietary data plane measurements from a large European IXP, it did not use or have access to any of the IXP's RS-specific BGP data. On the other hand, [87] relied on this new BGP data that can be obtained from some of the IXPs that operate a RS, but did not have access to any IXP-provided data plane measurements.

Taking full advantage of this data, we contrast our proposed traffic-aware analysis of the Internet peering ecosystem with the largely traffic-agnostic approaches that have traditionally been pursued. In the process, we highlight the important role that the IXPs' RSs have started to play for inter-domain routing and report on a number of empirical findings that shed new light on the Internet peering ecosystem as a whole and on our three IXPs, in particular. These findings include:

- RSs are enablers that offer new and complex peering options to IXP members and also provide them with open access to a significant portion of Internet routes.
- While RS usage varies across IXPs and IXP members, there exist some patterns of RS usage among IXP members with similar business types (with exceptions).
- Due to the popularity of RSs, multi-lateral peering dominates bi-lateral peering in terms of number of peerings but not in terms of traffic, but some of the top traffic contributors use mainly multi-lateral peering.
- RS usage contributes to a more dynamic peering landscape by enabling a proliferation of multi-lateral peerings and facilitating the use of both bi-lateral and multi-lateral peering sessions between two members at the same time.
- Recent developments in present-day RS design identify new research opportunities, including a feasibility study for adapting the RS paradigm for intra-domain routing in large ISPs.

9.1 Trends in IXP Operations

9.1.1 Background and Motivation

Recall, IXPs offer a shared switching fabric where the members can exchange traffic with one another once they have established peering connections (public peering) between them. This generic service provided by IXPs is an example of a *(positive) network effect* (also known as *network externality* [110]) because the more members an IXP has, the more valuable that IXP is to each member. Especially in Europe where many IXPs operate on a not-for-profit basis [67], this observation has led to significant innovations at IXPs in the form of constantly expanding service offerings.

One important such new service offering has been the free use of IXP-operated RSs. With more IXPs operating RSs, the membership base of those IXPs keeps growing and more of their member ASs start using the RS service. The resulting proliferation of multi-lateral peerings creates new options for the member ASs to peer with one another which, in turn, causes them to reconsider their existing peering arrangements. At the same time, the ease with which networks can join IXPs and start using them has let to a trend whereby one and the same AS can be a member at multiple IXPs, often within the same city or geographic region.

Thus, while we are observing an Internet peering ecosystem in flux, a pre-requisite for studying the current system to ultimately reason about its evolutionary path is to have a basic understanding of "all things concerning RSs". Given the central role that IXP RSs play in this chapter, we describe in this section the basic operations of an IXP RS and discuss the main reason for the observed bandwagon effect.

9.1.2 The IXP Route Server as Enabler

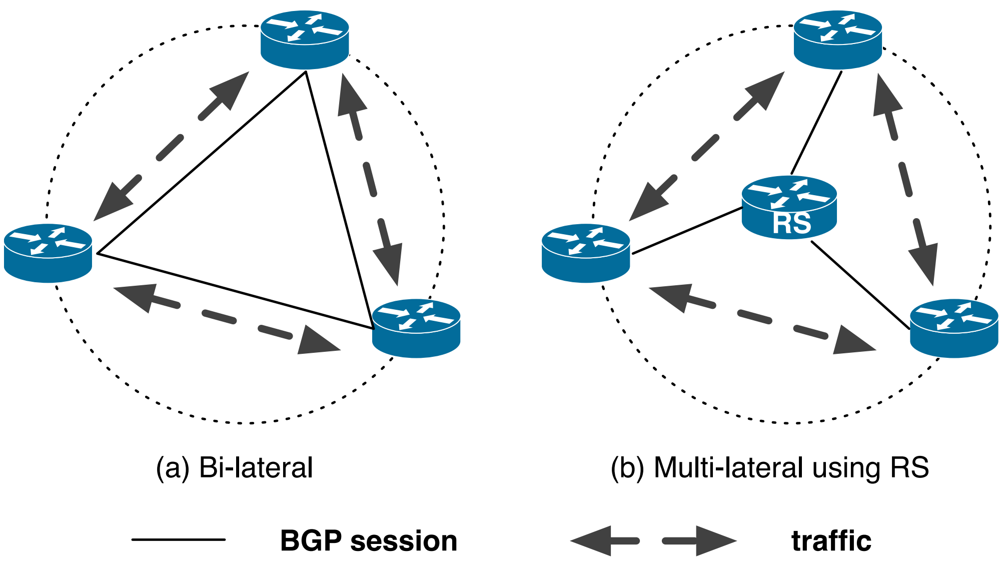
The typical way to establish connectivity between two ASs is to establish a direct eBGP session between two of their respective border routers. Initially, if two IXP member ASs wanted to exchange traffic via the IXP's switching fabric, they had to establish a *bi-lateral (BL)* eBGP peering session at the IXP. However, as IXPs grew in size, to be able to exchange traffic with most or all of the other member ASs at an IXP and hence reap the benefits of its own membership, a member's border router had to maintain more and more individual eBGP sessions. This started to create administrative overhead, operational burden, and the potential of pushing some router hardware to its limit.

To simplify routing for its members and to address these scalability challenges, IXPs introduced RSs and offered them as a free value-added service to their members. In short, an IXP Route-Server (RS) is a process that collects routing information from the RS's peers or participants (i. e., IXP members that connect to the RS), executes its own BGP decision process, and re-advertises the resulting information (i. e., best route selection) to all of the RS's peer routers. Figure 9.1 shows the flow of control plane information (BGP sessions) and data packets (data plane) for both traditional bi-lateral peering as well as peering via the RS at an IXP. The latter is referred to as *multi-lateral (ML)* peering because it typically involves more than two BGP partners.

Thus, IXP RSs greatly simplify routing for the IXP's members. A member AS just connects to the RS via a single eBGP session to set up BGP peering with all other IXP members that peer with the RS.[2] Clearly, this lowers the maintenance overhead, in particular for small ASs. Note, however, that using the RS (i. e., ML peering) does not preclude BL peering by one and the same member AS. In particular, larger ASs can take advantage of the RS while still having the option to establish BL peerings with selectively-chosen IXP members. For example, if a large member AS finds the capabilities of the RS to be insufficient for its needs (e. g., with respect to traffic engineering or filtering) or prefers to have more control over the peerings with its most important peers, it can use BL peerings with the latter and ML peerings with those members that peer with the IXP's RS.

Note that the IXP RS is not involved in the data path. Moreover, it executes the BGP decision process based on the information it learns from its peers. Thus, there is no guarantee that the best route selected by the RS is the same as the one that the peer would have selected given a full set of BL peering sessions. In this sense, while using a RS solves a scalability problem, it has the potential of creating a new problem known as the *hidden path problem* [95]. For example, if different peers of the IXP's RS advertise routes for the same prefix, only one route will be selected and re-advertised to all other peers. If this route cannot be propagated to a particular peer (e. g.,

[2] For redundancy purposes, IXPs typically operate two RSs and encourage their members to connect to both RSs.



(a) Bi-lateral

(b) Multi-lateral using RS

Figure 9.1: IXP peering options.

because of export filtering restrictions), that peer will not be able to receive any route for this prefix, even though an alternative path was advertised to the RS by a different peer. While this hidden path problem was an issue with earlier versions of the RS software, recent advances have largely overcome the problem, and in Section 9.1.4, we describe a popular Internet routing daemon that addresses this problem by providing support for maintaining peer-specific routing tables.

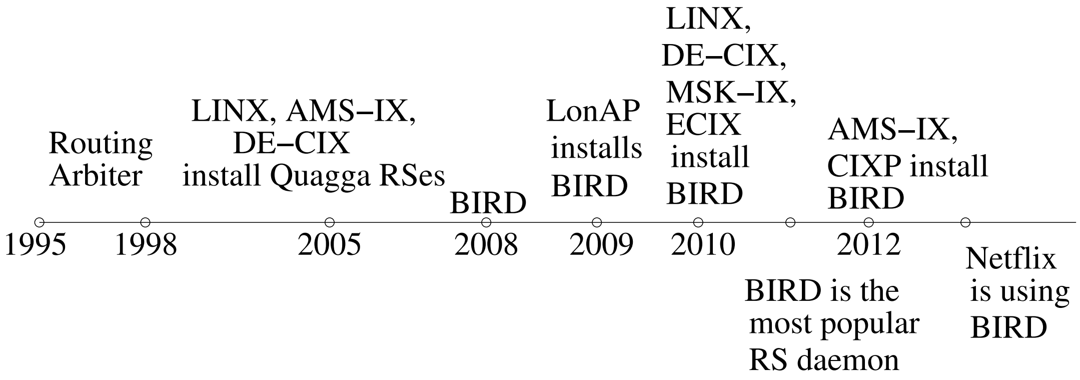


Figure 9.2: Route server deployment time line.

9.1.3 On the Deployment History of Route Servers

The first RSs were designed and deployed in the US as part of the decommissioning of the NSFNET around 1995 [89]. However, the service never took off due to the small size of the largest US IXPs at that time (i. e., MAE-East's membership approached some 20 ASs in mid-1990).[3]

It took another 10 years and the increasing popularity of some European IXPs for the RS concept to re-emerge and gain momentum. In fact, around 2005 some European IXPs with more than 100 member ASs (e. g., LINX in London, AMS-IX in Amsterdam, and DE-CIX in Frankfurt) started to offer RS capabilities to their members. The de-facto RS of this initial roll-out was Quagga [30]. However, it suffered from the hidden path problem and had numerous operational problems (e. g., prefix filtering, resource consumption, performance limitations) [94, 106, 96, 118]. OpenBGPD [24] addressed most of these problems and started to be deployed around 2008. Difficulties with its maintenance prevented OpenBGPD from being widely adopted and deployed [118].

In 2008, the BIRD [35] project was launched, and it took just three years for BIRD to become the most popular IXP route server [14]. Figure 9.2 shows a time line when some of the larger IXPs migrated to BIRD. In fact, the success of BIRD goes beyond RSs. Since 2013, it is the core routing component of the Netflix Open Connect Appliance [23].

9.1.4 BIRD: A Popular Router Daemon for RSs

BIRD is an open-source software routing daemon. The project was launched in 2008, is developed by CZ.NIC Labs, and has been actively supported by the IXP community. In the following, we describe a BIRD configuration that has been abstracted from the Euro-IX RS example [14] and is the basis of the one in operational use by one of the IXPs with which we have an ongoing collaboration.

Like all routing daemons, BIRD maintains a Routing Information Base (RIB) which contains all BGP paths that it receives from its peers – the Master RIB. However, when using BIRD as RS, it can be configured to (i) maintain peer-specific RIBs and (ii) use them instead of the Master RIB for peer-specific BGP best path selection (see Figure 9.3). More precisely, each member AS that peers with the RS maintains an eBGP session with the RS, which results in a peer-specific RIB. When IXP member AS X advertises a prefix to the RS, it is first put into the AS X-specific RIB. Next, if this prefix passes the AS X-specific import filter, it is added to the RS' Master RIB. If the export filter of AS X allows it, then this prefix will also be added to each AS Y-specific RIB, where AS Y is any other IXP member that peers with the RS. Then, the RS performs a peer-specific best path selection and exports the prefix by re-advertising it to AS Y.

IXPs typically insist on applying import filters to ensure that each member AS only advertises routes that it should advertise. For this purpose, the IXPs usually rely on route registries such as

[3] The RIPE RS [53] was a precursor of the RS proposed in [89].

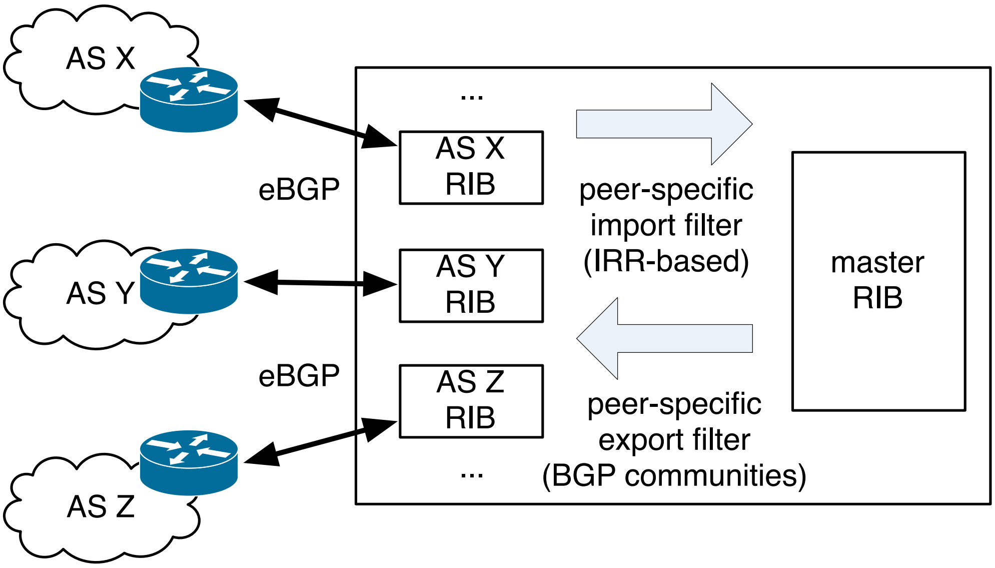

AS X
AS Y
AS Z
eBGP
eBGP
...
AS X RIB
AS Y RIB
AS Z RIB
...
peer-specific import filter (IRR-based)
peer-specific export filter (BGP communities)
master RIB

Figure 9.3: BIRD route server: Example setup.

IRR [20]. This policy limits the likelihood of unintended prefix hijacking and/or advertising bogus prefixes including private address space.

With respect to export filters, they are typically used by the IXP members to restrict the set of IXP member ASs that receive their routes. The commonly used vehicle for achieving this objective is the use of RS-specific BGP community values [86]. By using export filters, peers of the RS can express policies.

Note that using peer-specific RIBs overcomes the hidden path problem, because the BGP decision process is executed for each peer independently. If the best route for a certain prefix can not be re-advertised to some particular members (e. g., because of export filters), the RS can still select and advertise an alternative route to those particular members – presuming that another, non-blocked route is available. Maintaining separate RIBs for each member AS has a cost in terms of memory requirements and processing capabilities. However, recent stress tests and real deployment performance reports for BIRD at large IXPs [78, 106, 96] show that maintaining about 400 RIBs, each with up to 100K prefixes, consumes roughly 4 GB of memory and can be run on commodity servers.

9.1.5 Looking Glasses and Route Servers

The fact that network operators typically cannot see the results of their route announcements within their own network makes debugging BGP routing difficult. To address this problem, the community has set up BGP looking glasses (LGs) and route monitors (RMs) in many different locations across the Internet. LGs are servers, co-located on routers, where anyone can execute a limited number of commands. LGs are typically offered as a public service. On conventional LGs the commands include ping, traceroute, as well as some BGP commands. The latter provides details about all eBGP sessions of the router. In the past, LGs have also been used by researchers to gain better visibility into the AS-level topology and IXP-specific peering fabrics [90, 51].

LGs can also be co-located with RSs at IXPs. In this case, LGs act as proxies for executing commands against the Master RIB of the RS. The additional LG capabilities may include commands which list (a) all prefixes advertised by all peers and/or (b) the BGP attributes per prefix. Such LGs (referred to as LG-RS below) are already available at some IXPs, including DE-CIX and LonAP, and their capabilities have been recently used to explore AS connectivity at IXPs [87, 98]. However, the deployment of these LGs is still limited, and major IXPs such as AMS-IX do not offer publicly accessible LGs.

9.2 IXP Specific Data Sources

In this section, we describe the three European IXPs with which we have fruitful ongoing collaborations and that have been very forthcoming with sharing their IXP-internal datasets with us. We provide details about these datasets as well as the IXP-external datasets that are at our disposal for this study.

9.2.1 Profiles of three IXPs

The three IXPs differ greatly in size (e. g., number of members or traffic) and service offerings that they provide for their members. Table 9.1 summarizes these differences, focusing in particular on the features relevant for this chapter.

Large IXP (L-IXP): With close to 500 members and peak traffic that exceeded 3 Tb/s in late 2013, this IXP is one of the largest IXPs in Europe and worldwide. It operates a layer-2 switching fabric that is distributed over a number of co-locations/data centers within a metropolitan area.

Members can connect on ports ranging from 1 Gb/s to 100 Gb/s. Its membership includes the full spectrum of networks, including CDNs, content providers, hosters, the whole range of ISPs (from tiertier11 to regional to local providers), and resellers. This IXP provides arguably one of the richest and most advanced service offerings, including SLAs, remote peering, and black-holing. One of its key service offering is the free use of its RS. It operates two (for redundancy) IPv4 and IPv6 RSs using a BIRD configuration similar to the one discussed in Section 9.1.4 with advanced LG support.

Medium IXP (M-IXP): This medium-sized IXP operates a widely-distributed layer-2 switching fabric to which members can connect to in several locations. Half of its members are also members at the L-IXP. As of late 2013, this IXP had 101 members and its peak traffic exceeded 250 Gb/s. It offers the free use of its RS (again, two for each IPv4 and IPv6) without peer-specific RIBs and a LG that supports the execution of a limited set of commands.

IXP	L-IXP	M-IXP	S-IXP	L-IXP & M-IXP
Member ASs	496	101	12	50
tiertier11 ISPs*	12	2	1	2
Large ISPs*	35	4	1	4
Major Content/ Cloud Providers*	17	5	0	5
RS (IPv4/IPv6)	BIRD Multi-RIB	BIRD Single-RIB	No	/
Public RS-LG (IPv4/IPv6)	Yes	Yes, limited commands	No	/
Member ASs using the RS	410	96	/	43

Table 9.1: IXP profiles.

Importantly, in terms of size, membership, service offerings, and operations, this IXP is typical for many of the medium-sized European IXPs.

Small IXP (S-IXP): This is a national IXP with a single site in the country's capital. As of late 2013, it had 12 member ASs and saw peak traffic of about 20 Gb/s. It offers a very limited set of service, and most importantly for our work, it does not operate a RS. Its members are mainly the local and national ISPs. While a few global players are present, none of the CDNs and content providers are. This IXP is typical for the many small European IXPs that generally serve a specific local need and are not driven by growth.

9.2.2 IXP-internal Data: Route Server

For the two IXPs that operate a RS (i. e., L-IXP and M-IXP), we have access to the data from their BIRD deployment. In the case of L-IXP, we have weekly snapshots of the peer-specific RIBs, starting in June 2013. For M-IXP, we have several snapshots of the Master-RIB, starting in December 2013. In addition, for L-IXP, we have all BGP traffic to and from its RS that is captured on the network interface via tcpdump.

9.2.3 IXP-internal Data: Traffic

For each of our three IXPs, we have access to data plane measurements in the form of traffic that is routinely collected from the IXPs' public switching infrastructures. More precisely, for each IXP, the available datasets consist of massive amounts of SFLOW records [144], sampled from their public switching infrastructure. The measured SFLOW records contain Ethernet frame samples

*The numbers reflect only the well-known and easy-to-identify networks by business type to illustrate the existence of common members. No attempt at a complete classification has been made.

that have been collected using random sampling (i. e., 1 out of 16K at L-IXP and M-IXP, and 1 out of 32K at S-IXP). SFLOW captures the first 128 bytes of each sampled frame. Thus, they contain full Ethernet, network- and transport-layer headers, as well as some bytes of payload for each sampled packet. For further details about relevant aspects of these collected SFLOW records (e. g., absence of sampling bias, removal of irrelevant traffic), we use a similar technique as the one outlined in [43].

In this work, we rely on four continuous weeks of collected SFLOW for each IXP. For L-IXP we cover the period in August/September 2013, for M-IXP and S-IXP we cover the period of December 2013. In addition, we rely on week-long snapshots collected at the L-IXP dating back to 2011.

9.2.4 Public IXP-specific Data

In our work, we also rely on a number of widely-used public datasets, including BGP data from the route collectors from RIPE RIS, Route-views, PCH [28], and Abilene [19]. We refer to them as RM BGP data. We also use a more recently obtained collection of BGP data that was recorded with the help of LGs. In particular, we use the data from the RS-LG at the L-IXP in accordance with the method described in [87]. Moreover, we rely on data obtained from several LGs that query the routing tables of member-routers at the L-IXP.

9.3 Connectivity: Bi-/Multi-lateral

Relying on our IXP-provided measurements, we show in this section how we can get close to recovering the actual peering fabrics at the IXPs. We also illustrate what portions of those actual peering fabrics can and cannot be recovered when using the different public BGP data.

9.3.1 Connectivity: IXP-provided Data

To determine if IXP members AS X and AS Y are using a ML peering at the IXP, we rely on the IXP-provided RS data. More specifically, for the L-IXP, we first check if AS X and AS Y peer with the RS. If so, we next check in the peer-specific RIB of AS Y for a prefix with AS X as next hop. If we find such a prefix, we say that AS X uses a ML peering with AS Y. If we also find AS Y in the peer-specific RIB of AS X as next hop, we say that the ML peering between AS X and AS Y is *symmetric or bi-directional*; otherwise, we say that the ML peering between AS X and AS Y is *asymmetric*.

Given that the RS at the M-IXP only uses the Master RIB but no peer-specific RIBs, we re-implement the per-peer export policies based upon the Master RIB entries to determine peerings via the RS. More specifically, if there is a route for a prefix in the Master RIB with AS X as next hop, we postulate a ML peering with all member ASs that peer with the RS, including AS

Multi-lateral Peerings: RS RIBs				
	IPv4		IPv6	
	symmetric	asym.	symmetric	asym.
L-IXP	65,599	14,153	34,596	5,086
M-IXP	3,140	594	1,173	434
Bi-lateral Peerings: Inferred from Data-Plane				
	bi-/multi	bi-only	bi-/multi	bi-only
L-IXP	14,673	5,705	4,256	3,727
M-IXP	399	61	223	75
S-IXP	/	55	/	20
Total Peerings				
L-IXP	85,457		43,409	
M-IXP	3,795		1,682	
S-IXP	55		20	
Visibility in the RS Looking Glass				
L-IXP	all multi-lateral		all multi-lateral	
M-IXP	none		none	
S-IXP	/		/	

Table 9.2: Multi-lateral and bi-lateral connectivity.

Y, unless the community values associated with the route explicitly filter the route via the peer-specific export filter to AS Y.

For each of our three IXPs, to determine if IXP members AS X and AS Y are using a BL peering at the IXP, we rely on the IXP-provided traffic measurements. In particular, to conclude that AS X and AS Y established a BL peering at the IXP, we require that there are SFLOW records in the IXP-provided traffic data that show that BGP data was exchanged between the routers of AS X and AS Y over the IXP's public switching infrastructure. We cannot however differentiate between asymmetric and symmetric BL peerings with these data plane measurements.

Note that our methodologies yield a lower bound for BL peerings and an upper bound for ML peerings, but there is evidence that these bounds are in general very tight. For example, with respect to BL peerings, our method is not significantly biased by the SFLOW sampling rate because the numbers are very stable once we use data from more than two weeks. Indeed, Figure 9.4 shows that for the L-IXP, the additional BL peerings seen in the third (fourth) week are less than 1% (0.5%). As far as ML peerings are concerned, our method does not account for the fact that some RS peers might reject the advertisements of the RS. While we have rarely encountered this behavior, it nevertheless can result in some over-counting by our method. At the same time, we find member ASs that use one and the same link for ML as well as for BL peering.

The routers' IP addresses have to be within the publicly known subnets of the respective IXP.

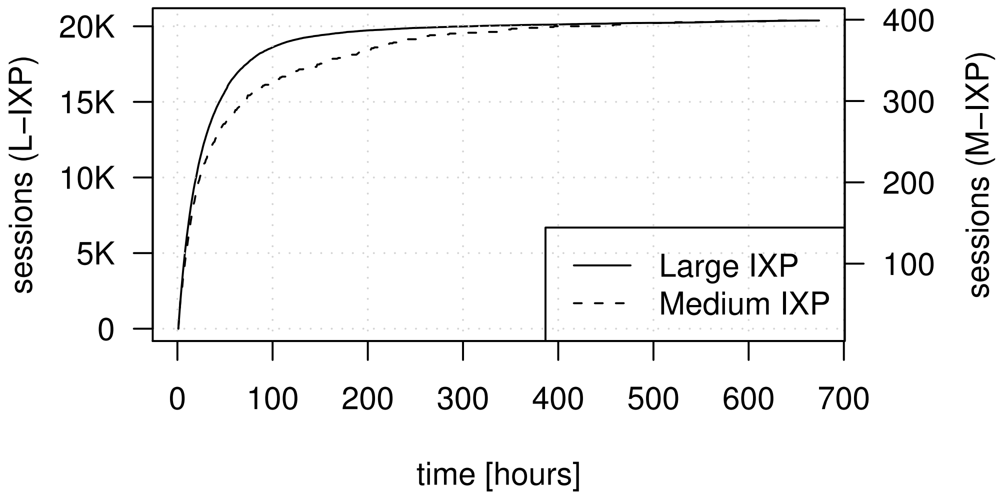


Figure 9.4: Inferred bi-lateral BGP sessions over time.

Our best efforts to reconstruct the actual ML and BL peering fabrics of each of our three IXPs is summarized in Table 9.2. To highlight the power of the IXP-provided datasets, we further break down (where possible) each of the ML and BL peering fabrics into links that are used for either IPv4 or IPv6 and in a symmetric or asymmetric manner. For each IXP, we also tally the total number of peerings that result from applying our methodologies.

9.3.2 Connectivity: Public BGP Data

To assess the ability of the various public BGP datasets to provide visibility into the peering fabrics of our three IXPs and compare with our findings in Section 9.3.1 above, we first mine the RS-LG BGP data that has recently been considered and used in [87] and is publicly available. The results are shown in Table 9.2. There are three main take-aways. First, when restricting the study to the L-IXP that has an RS-LG with the capabilities listed in Section 9.1.5, the new methodology proposed in [87] is indeed successful in recovering the full-fledged ML peering fabric of that IXP. Second, as the example of the M-IXP shows, having a RS-LG that lacks the necessary querying capabilities is of no help for inferring the IXP's ML peering fabric. Third, irrespective of the capabilities of the RS-LG, the RS-LG BGP data cannot be used to reveal any peerings of the L-IXP's and M-IXP's BL peering fabrics (recall that there is no ML peering at the S-IXP).

For completeness, we also obtained the number of peerings that can be discovered from mining the traditional and widely-used RM BGP data. The results confirm past findings that a majority of the peerings (i. e., some 70-80%) cannot be seen in those data [43, 51, 86, 98, 138]. In addition, we also notice a significant bias in this data towards BL peerings. Interestingly, this data does produce peerings between IXP member ASs that we do not see even in our most complete peering fabrics

Some portion of the ML peering fabric may be recovered but only with substantial additional efforts. For example, by using prefixes extracted from the RIPE RIS or Route-views data, a portion of the IPv4 prefixes on the Master RIB of the M-IXP can be recovered and used to submit queries to this RS-LG.

discussed in Section 9.3.1. One possible explanation for seeing such peerings in the public but not in the IXP-provided BGP data is that the member ASs in question engage in *private peering* at the IXP. As private peerings are handled completely separate from the IXP's public switching infrastructure, IXP-provided route server data does not include them.

In view of Table 9.1, the picture that emerges with respect to connectivity at the L-IXP and M-IXP is that their free RS service is very popular with their members, resulting in the number of ML peerings at those IXPs to be 4X and 10X larger than the number of established BL peerings. We also observe significant fractions of peerings for IPv6 and of asymmetric peerings.

10

Exploring EDNS-Client-Subnet Adopters

While Chapters 8 and 9 deal with an IXP as vantage point, this Chapter focuses on active measurements being combined with passive measurements at a large European ISP. The focus is on exploring the Internet actively, using new or recently adopted technologies to our advantage by creatively "exploiting" features of a protocol extension.

10.1 Introduction

In the current Internet, users who want to resolve hostnames typically use the local resolver provided by their Internet Service Provider (ISP). Unless the answer is cached, the ISP's domain name server (DNS) performs a recursive lookup to receive an authoritative answer which it can then cache. DNS is used by Large Content Delivery Networks (CDNs) as well as Content Providers (CPs) to map users to possible locations and consequently to appropriate servers [124, 155].

Unfortunately for the CDNs and CPs, the DNS query is not directly issued by the end-user but only by the local resolver (see [154] for details). Thus, the assumption underlying this solution is that the end-user is close to the local resolver. However, several studies [44, 119] have shown that this assumption does not always hold which, in turn, can lead to degraded end-user experience. In particular, the introduction of third party resolvers like Google Public DNS [18] and OpenDNS [25] has exaggerated this trend as end-users, taking advantage of these as local resolvers may experience poor performance [44, 127, 140]. This empirical evidence is mainly due to the fact that these popular third-party resolvers are typically not located within the end-users' ISP and are therefore often not close to the end-user.

The solution, as proposed by Google and others to the IETF [75], is the EDNS-Client-Subnet DNS extension (ECS). While traditional DNS queries do not include the IP address of the query issuer (except via the socket information), ECS includes IP address information of the original query issuer in the query to the authoritative name server. The IP address information can then be used by the CDN or CP to improve the mapping of end-users to servers. Thus, it is not surprising that major Internet companies (e. g., Google, EdgeCast, and OpenDNS) have already adopted ECS and have established the consortium "a faster Internet" [2]. The fact that ECS can indeed help improve end-user performance is highlighted by extensive active measurement studies [127, 140].

For the Internet measurement community, the adoption of ECS by some of the major Internet players turns out to offer unique but clearly unintended opportunities. To illustrate, we show how ECS can be used to uncover details of the operational practices of ECS adopters with almost no effort. Our key observation is that ECS allows anyone to issue queries on behalf of any "end-user" IP address for domains with ECS support. Thus, ECS queries can be used to uncover the sophisticated techniques that CDNs and CPs use for mapping users to servers (e.g, see [104, 103, 119]). Indeed, this currently hard-to-extract information can now be collected using only a single vantage point and relying on publicly available information. In the past, to obtain similar information, network researchers had to find and use open or mis-configured resolvers [41, 93, 151], required access to a multitude of vantage points in edge networks [127, 140], relied on volunteers [44, 45], analyzed proprietary data [105, 132], or resorted to searching the Web [149]. In summary, the three contributions to the area of Internet measurement are:

- We show that a single vantage point combined with publicly available information is sufficient to uncover the global footprint of ECS adopters.
- We demonstrate how to infer the DNS cacheability and end-user clustering strategies of ECS adopters.
- We illustrate how to uncover the assignment of users to server locations performed by major ECS adopters.

At the same time, this work is also intended to increase the awareness of current and future ECS adopters about which operational information gets exposed when enabling this recent DNS extension. Our software and the measurements will be made accessible to the research community.

10.2 Data Sets

The following two different kinds of datasets are used: (i) prefixes to be used as fake "client location" for the ECS queries, and (ii) popular ECS adopters.

10.2.1 Network Prefixes

In principle, one list of prefixes may be considered sufficient. However, because we want to uncover the operational practices of certain CDNs and CPs with regards to client localization and clustering, we explore different prefix sets of varying scopes and magnitudes.

Academic Network (UNI). This network includes a very diverse set of clients, ranging from office working spaces to student dorms and research facilities which complicates usage profiling. This network uses two /16 blocks, does not have an AS, and is localized to a single city in Europe.

Large ISP (ISP). The dataset includes more than 400 prefixes, ranging from /10 to /24, announced by a European tier-1 ISP. This ISP offers services to residential users as well as enterprise networks and also hosts CDN servers.

Large ISP, de-aggregated prefixes (ISP24). We also use the de-aggregated announced prefixes of our large ISP at the granularity of /24 blocks to investigate if the finer granularity lets us uncover additional operational details.

Popular Resolvers (PRES). This proprietary dataset consists of the 280K most popular resolver IPs that contacted a large commercial CDN. These resolvers are distributed across 21K ASs, 74K prefixes, and 230 countries.

RIPE. RIPE RIS [31] makes full BGP routing tables from a multitude of BGP peering sessions publicly available. This data includes 500K prefixes from 43K ASs. We always use the most recent routing table.

RouteViews (RV). RV [32] is another public BGP routing table source offered by the University of Oregon.

10.2.2 Content Provider Datasets

For our experiments, we need to identify ECS adopters and the corresponding hostnames. For this purpose, we utilize Alexa [3], a publicly available database of the top 1M second-level domains, as was done in [127]. Since the ECS extension does not allow us to directly find out if a name server is ECS enabled or not, we use the following heuristic. We re-send the same ECS query with three different prefix lengths. If the scope is non-zero for one of the replies, we identified an ECS-enabled server and hostname.

We obtain two groups of (domain names, name-servers) pairs. The first group fully supports ECS and accounts for 3 % of the second-level domain names. The second group, about 10 %, seems to be ECS-enabled but does not appear to use it for the tested domains. Thus, roughly 13 % of the top 1 million domains may be ECS-enabled. This number is only slightly larger than what was reported in a 2012 study [127]. Importantly, some of the big players among the ECS adopters include Google (and YouTube), EdgeCast, CacheFly, OpenDNS, HiCloud, and applications hosted in the cloud such as MySqueezebox.

To estimate the potential traffic affected by ECS, we use a 24-hour packet-level anonymized trace from a large European ISP. The trace is from a residential network with more than 10K active end-users and was gathered using Endace cards and analyzed with Bro [129]. It contains 20.3M DNS requests for more than 450K unique hostnames and 83M connections. While Alexa only includes the second-level domain names, this dataset allows us to identify full hostnames, which we use in a similar manner as above. In total, we find that roughly 30 % of the traffic involves ECS adopters. This highlights that while the number of ECS adopters is relatively small (less than 3 % of authoritative name servers), it includes some of the relevant players responsible for a significant fraction of traffic.

From the set of identified ECS adopters, we select a large CDN, two smaller CDNs, and an application deployed in the cloud to explore their operational practices. In particular, we focus on:

Google is a founding member of the consortium “a faster Internet” and one of the main supporters of ECSİt has adopted ECS in all their resolvers and name servers. Moreover, Google uses a sophisticated back-end with many data centers, edge-servers, and Google Global Cache (GGC) servers [16, 52]. Moreover, it is known that there can be thousands of Google servers behind a single Google IP [148].

EdgeCast is a large CDN that also offers streaming solutions. It is also one of the participants in the “a faster Internet” consortium.

CacheFly is another CDN that has adopted ECS.

MySqueezebox is a Logitech product that runs on top of Amazon’s Web cloud Service EC2.

10.3 Methodology

For our experiments, we take advantage of the ECS extension of the python DNS libraries provided by OpenDNS [25]. Based on this library, we have developed a framework which utilizes the above API to send ECS DNS queries with arbitrary IP/prefix (even full IPs) to authoritative name servers. By embedding this library into our test framework, we can handle failures and retries efficiently which would have been more complicated with a stand-alone utility like the patched dig tool [75].

We emphasize that a single vantage point is sufficient for performing our experiments. This is mainly due to the fact that with ECS, the answers we obtain do not depend on the IP of the resolver issuing the DNS request but only on the prefix included in the ECS resource record. As a result, we can use a single server, a commodity PC, from which we issue queries at the rate of 40 to 50 queries per second. With regards to the queries, we use hostnames from the Alexa list and the ISP traces and the source prefixes from our prefix datasets. Note that because a large fraction of IPv6 connectivity is still handled by 6to4 tunnels [73] and related techniques, we did not include IPv6 in this preliminary study.

For each query we issue we add an entry to our SQL database which includes all parameters including the timestamp, the returned records (answers) including TTL and returned scope. Before and after each experiment, we collect the most recent prefixes for each dataset. To speed-up the experiments, we unique the set of prefixes before starting an experiment.

10.4 Evaluation

To evaluate the (unintended) capabilities of ECS as a measurement tool, we explore how difficult it is to (i) uncover the footprint of ECS adopters, (ii) assess the impact of ECS on cacheability of DNS records, and (iii) determine how ECS adopters assign users to server locations. The reported results for Google, MySqueezebox, and CacheFly are from experiments performed on 03/26/13 and the ones for EdgeCast are from 05/07/13. All queries are sent for a single hostname (e. g.,

	Prefix set	Server IPs	Subnets	ASs	Countries
Google	RIPE	6,284	320	163	47
	RV	6,280	320	163	47
	PRES	5,940	301	156	46
	ISP	207	28	1	1
	ISP24	535	44	2	1
	UNI	123	17	1	1
EdgeCast	RIPE/RV/PRES	7	7	1	3
	ISP/ISP24/UNI	1	1	1	1
CacheFly	RIPE/RV/PRES	21	21	11	11
	ISP/ISP24/UNI	1	1	1	1
MySqueezebox	RIPE/RV/PRES	10	7	2	2
	ISP/ISP24/UNI	6	4	1	1

Table 10.1: ECS adopters: Uncovered footprint.

www.google.com) to one of the authoritative name servers of the respective service provider (e. g., ns1.google.com).

10.4.1 Uncovering Infrastructure Footprints

We first report on our experiences with using ECS to uncover the footprint of the four selected ECS adopters. Table 10.1 summarizes the number of unique server IPs, subnets, ASs, and locations (country-level only [133]). It shows that the footprint of Google is by far the most interesting one, with more than 6K server IPs across 163 ASs in 47 countries. While illustrative and also informative (e. g., we uncover more locations than previously reported [45, 148]), the main and more surprising finding is the simplicity with which we can uncover this infrastructure using ECS from a single vantage point in less than 4 hours.

For validation purposes, we checked each server IP—all of them served us the Google search main page. In addition, the reverse look-up revealed that while all servers inside the official Google AS use the suffix 1e100.net [88], those deployed in third-party ASs use different hostnames (e. g., cache.google.com, or names containing the strings ggc or googlevideo.com, respectively).

Concerning the used prefix sets, we observe some differences. Both the RIPE as well as the RV prefix sets are sufficiently complete to yield the same results. PRES is not sufficient to uncover the full set of Google Web servers, but it yields a major fraction of them in only 55 minutes per experiment. Alternatively, one can use a subset of the RIPE/RV prefix sets. Using a random prefix from each AS reduces the number of RIPE/RV prefixes to 43,400 (8.8 % of RIPE prefixes) and results in 4,120 server IPs in 130 ASs and 40 countries in 18 minutes (with 40 requests/sec). When doubling the number of selected prefixes to two per AS, we uncover 4,580 server IPs in 143 ASs, and 44 countries.

When relying on the ISP, ISP24, and UNI prefix sets, we see the impact of Google's sophisticated optimization techniques that try to map users to appropriate servers. In particular, we uncover a much smaller number of servers. Indeed, Google's techniques work reasonable well as most network prefixes are mapped to Web server IPs in a single AS. However, by using the de-aggregated prefix set of the ISP (i. e., ISP24), we are able to expand the coverage from 200 to more than 500 server IPs. More than 95 % of them are in the Google AS while the rest are in a neighbor ISP. A more careful investigation revealed that the prefixes served by the neighbor ISP are from a customer of this ISP whose prefix is not announced separately but only in aggregated form (i. e., together with other prefixes of the ISP). Google is able to infer this and redirects clients in this ISP to caches located in the neighbor ISP.

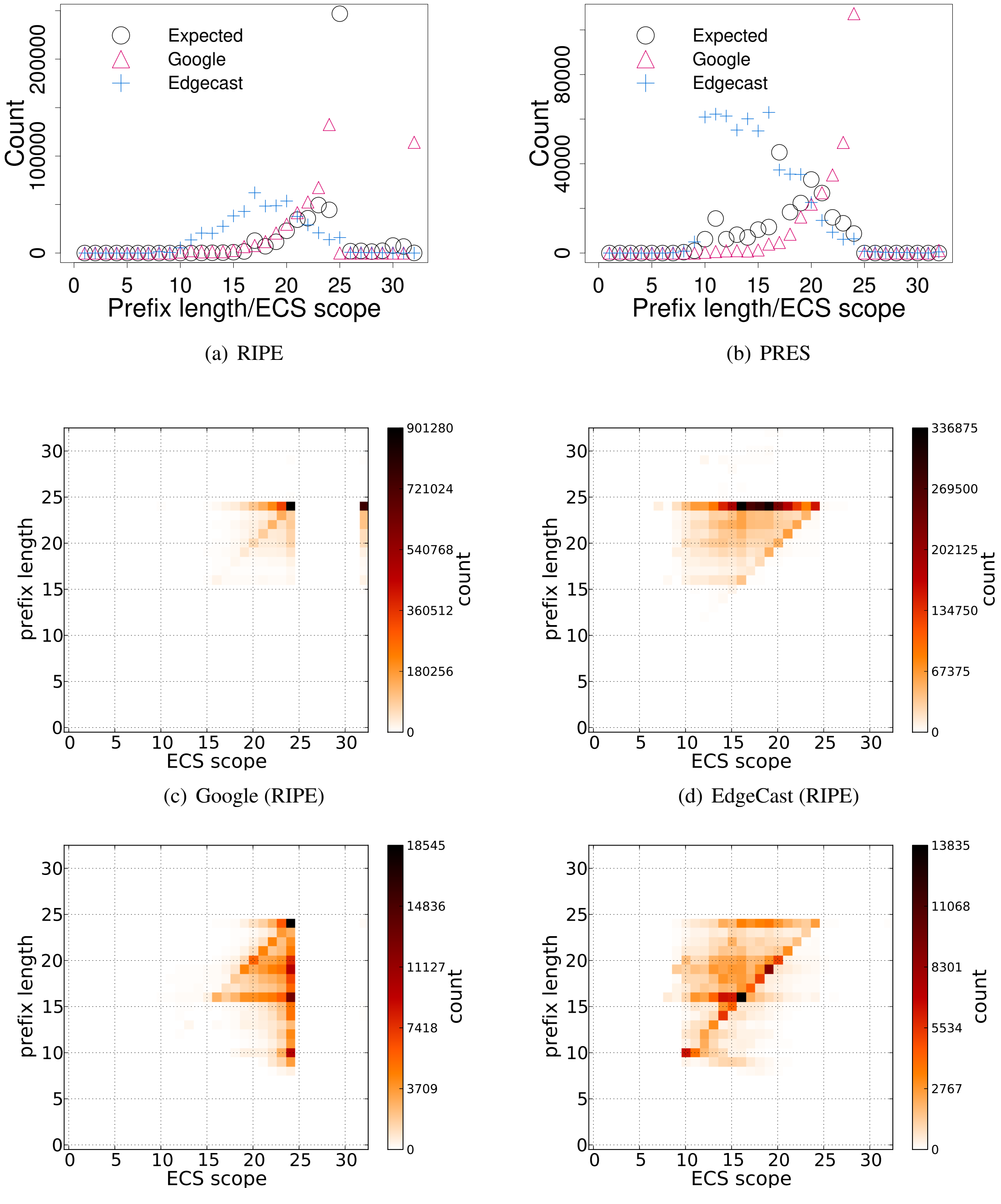
Of the ASs uncovered by using the RIPE, RV, PRES prefix sets, only 845 and 96 server IPs are in the ASs of Google and YouTube, respectively. All the others IPs are in ASs not owned by Google. This shows the profound impact of Google (GGC) caches which have been deployed by many ASs [16]. We repeat the experiments by using the Google Public DNS server and observed that the returned answers were almost always identical (99 %). This is not necessarily the case when using Google's Public DNS server for other lookups. However, Google's Public DNS server is forwarding ECS queries to white-listed authoritative DNS servers of other ECS adopters. Therefore, we can even (ab)use Google's Public DNS server as intermediary for measurement queries and thus (i) hide from discovery or (ii) explore if these ECS adopters use a different clustering for Google customers.

Table 10.1 also shows that the footprint of the other ECS adopters are "less" interesting, mainly because their footprint is not as widely distributed as the one of Google. Nevertheless, we see in principle similar results. Most of the infrastructure can be uncovered with the RIPE/RV/PRES prefix sets. The ECS adopters again use clustering such that the ISP, ISP24, and UNI prefixes are all mapped to a single server IP. Note that EdgeCast may use HTTP-based redirection which cannot be uncovered using only DNS. While EdgeCast uses a single AS, CacheFly, and MySqueezebox are utilizing infrastructures across multiple ASs. We also observe that both players map the UNI and ISP/ISP24 prefixes to infrastructures in Europe (e. g., MySqueezebox maps them to the European facility of Amazon EC2).

10.4.2 Uncovering DNS Cacheability

Next, we examine the meta information included in the ECS response: the scope. In principle, if the prefix length corresponds to a publicly announced prefix, one may expect that the returned scope is equal to the prefix length. However, this is not necessarily the case. Content providers often return either coarser- or finer-grained summaries; e. g., they respond either with aggregated or de-aggregated scopes (as compared to the prefix length) which indicates that they perform the end-user clustering for client to server assignment on a different granularity than the routing announcements may suggest.

We note, that the scope may have a major impact on the re-usability of the DNS response, i. e., the cacheability of DNS responses. While most of the responses have a non-zero TTL, a surprisingly large number have a /32 scope. An ECS scope of /32 implies that the answer is valid only for the specific client IP which issued the DNS request. In this section, we explore DNS cacheability for



(a) RIPE

(b) PRES

(c) Google (RIPE)

(d) EdgeCast (RIPE)

(e) Google (PRES)

(f) EdgeCast (PRES)

Figure 10.1: Prefix length vs. ECS scope for RIPE and PRES.

two ECS adopters in detail: Google and EdgeCast. The others are less interesting as CacheFly always uses a /24 scope and MySqueezebox is similar to EdgeCast.

Figure 10.1(a) shows the RIPE prefix length distribution (circles). In addition, it includes the returned scopes from using the RIPE prefixes to query our ECS adopters. We note, that the distributions vary significantly. There is massive de-aggregation by Google but massive aggregation by EdgeCast which operates a smaller infrastructure.

When executing back-to-back measurements for Google (e. g., 4 queries within a second), we find that the answer as well as the scope can change within seconds even though the TTL is 300 seconds. Overall, as seen in Figure 10.1(a), for almost a quarter of the queries, the returned scope is 32. This indicates that currently, Google severely restricts the cacheability of ECS responses or may want to restrict reuse of the answers to single client IPs. For approximately 27 % of the queries prefix length and scope are identical. For 41 % of the queries we see de-aggregation while there is aggregation for 31 %. We again note, that the returned scopes for the RIPE and RV prefixes are almost identical.

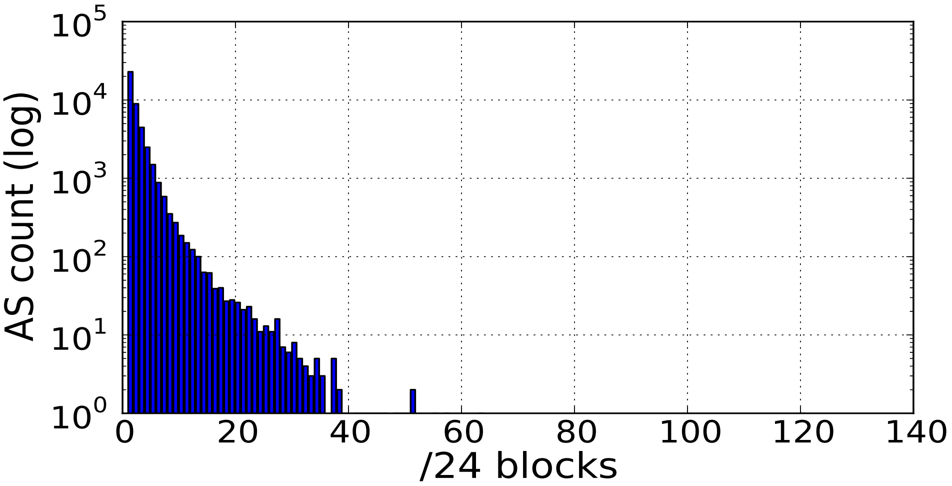
Exploring the scopes returned by EdgeCast may at first glance appear useless, because they only returned a single IP with a TTL of 180 seconds. However, EdgeCast is using significant aggregation for all prefix lengths across all prefix sets. For example, when using the RIPE prefixes, the return scope is identical for 10. 5 % but less specific for 87 %.

When using the ISP prefix set, the overall picture is similar even though the specific numbers vary. An initial study of the prefixes with scope 32 indicates that Google has succeeded in profiling at least some of the IP subnets of the ISP; e. g., Google returns scope /32 for all CDN servers of a large CDN provider inside the ISP. In future work, we plan to explore if there exists a natural clustering for those responses with scope /32.

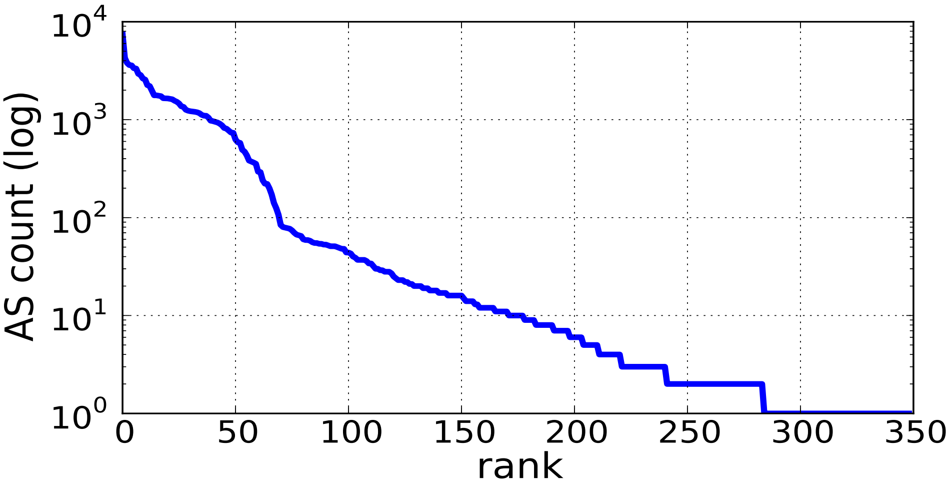
Given that the size of the UNI prefix set is limited, we issued queries from all contained IP addresses with prefix length 32. For this case, we have some evidence of per-subnet profiling as the returned scopes vary heavily from /32 to /15, even for neighboring IP addresses.

For the PRES prefixes, Figure 10.1(b) shows extreme de-aggregation. For more than 74 % of the prefixes, the scope is more restrictive than the prefix length, and in 17 % they are identical. Only few returned scopes are /32s. This may indicate that Google treats popular resolvers differently than random IP addresses. Google may already be aware of the problem regarding caching DNS answers as discussed in Section 2.6.2. For EdgeCast we see significant aggregation.

To highlight the relationship of prefix length in the query to scope in the reply, Figures 10.1(c) and 10.1(e) show heat-maps of the corresponding two-dimensional histograms. For the RIPE dataset we notice the two extreme points at scopes /24 and /32. For the PRES dataset, the heat-map highlights the de-aggregation. Figures 10.1(d) and 10.1(f) show the heat-maps for EdgeCast. While for the RIPE dataset we see the effect of the extreme prefix de-aggregation for Google very clearly, the picture for EdgeCast is more complicated as there is mainly aggregation. For the PRES dataset, the heat-map shows even more diversity as there is de-aggregation as well as aggregation. This results in a blob in the middle of the heat map.



(a) # of ASs mapped to # of /24s with Google servers.



(b) Ranking of # of ASs served by /24s with Google servers.

Figure 10.2: Google: User to server mapping (RIPE).

10.4.3 Uncovering User-to-Server Mapping

So far we have not yet taken advantage of the Web server IP addresses in the DNS replies. These allow us to uncover the user-to-server mapping employed by an ECS-enabled CDN or CP. In the following, to illustrate another example of the unintended measurement capabilities offered by ECS, we explore how Google performs this mapping (based on the RIPE prefix set) and examine how stable this mapping is.

Google returned 5 to 16 different IP addresses in each reply. Almost all responses ($> 90\,\%$) included either 5 or 6 different IP addresses. We did not find any correlation between the ECS prefix length or the returned scope and the number of returned IP addresses. All IP addresses from a single response always belong to the same /24 subnet. Thus, based on a single ECS lookup per prefix, we always find a unique mapping between query prefix and the server subnet from the DNS reply. Overall, we identified 349 unique /24s.

To understand the stability of the mapping, we aggregate the query prefixes to the AS level and ask if all client prefixes within an AS are mapped by Google to the same or a small number of server subnets. In short, the answer is yes. 22.9K ASs are always mapped to a single /24 which hosts Google servers. 9K (5K) ASs are mapped to two (3) /24s, respectively. However, there are also more than 4.5K ASs with around 335K prefixes for which the mapping varies drastically and which are served by more than 5 different /24s (e. g., 18 ASs are mapped to servers in more than 40 different /24s). For a more detailed account, see Figure 10.2(a). Manual inspection shows that this latter set of ASs includes some with a highly distributed footprint. Next, we consider aggregation to the AS level for the subnets that host Google servers. We find that a majority of the ASs (41K) is served exclusively by Google servers from a single AS. About 2K ASs are mapped to two ASs, and less than 20 ASs are served by servers from more than 7 ASs.

Next, looking at the data differently, we ask how many different ASs are mapped to a single /24 block that hosts Google servers. We find a highly skewed distribution (see Figure 10.2(b)). A small fraction of subnets is responsible for a large number of ASs. The top /24 block (173.194.33.0/24) serves 7.4K ASs. On the other extreme, there are 12 /24s that serve only prefixes from a single AS. Those are typically /24 blocks that host GGC servers. Among the ASs hosting these /24s we find that the official Google AS (AS15169) is by far the most often used AS with prefixes from more than 41.5K mapped to it. Another popular AS is YouTube (AS36040) even though we only queried for the Google Search home page. Of the 163 ASs that host Google servers, around 60 exclusively serve client prefixes from a single AS.

When we analyzed the returned IP addresses over time (May 3-4, 2013), we found that around 35 % of the prefixes are always served by a single /24 block over a period of 48 hours. Given the highly distributed infrastructure of Google one may have expected larger churn. 44 % of the query prefixes are mapped to two /24 and a very small percentage to more than five /24s. One possible explanation for this stable mapping is that Google uses local load-balancers [104, 148]. In future work, we plan to further uncover how the user-to-server mapping evolves in time not just for Google but also for other CDNs and CPs.

10.4.4 Summary

We show that the adoption of EDNS-Client-Subnet DNS extension by major Internet companies offers unique but most likely unintended measurement opportunities to uncover some of their operational practices. Using early ECS adopters like Google, EdgeCast, CacheFly and MySqueezebox as examples, our experimental study shows how simple it is (using a single vantage point as simple as a commodity PC) to (i) uncover the footprint of these CDN/CP companies, (ii) infer how they cluster clients, and (iii) reverse-engineer how they map users to servers. In addition, we point out potential implications that ECS can have on the cacheability of DNS responses by major DNS resolvers. This work also highlights the need to increase the awareness among current and future ECS adopters about the consequences of enabling ECS.

11 Conclusion

Investigating the Internet is still difficult, as the number of players and roles, and the complexity of business and interconnection increase quantitatively and change qualitatively. We tackle this increasing difficulty by examining multiple independent data sources, of which some, i. e., IXPs and ISPs, are proprietary, and others are generally available, i. e., public data repositories and active measurements.

The ISP business is well understood, but unfortunately mostly only by the ISPs themselves. Those companies tend not to publish too many data besides what is required by regulatory and market authorities or for the sake of public relations. However, we have had the chance to measure different levels of network traffic at one major European ISP, including packet-level traces at high speed, NETFLOW information, and BGP routing information. We have described the challenges and how to operate vantage points to collect packet-level traces. In addition, we have provided valuable best practice information for anyone planning such tasks. The ISP as a vantage point is still one of the most important entities for observing the Internet and its current development, because ISPs are and continue to be powerful players. In addition, the ISP business is changing. ISPs have greatly evolved, from providing bandwidth "bit-pipes" to becoming service providers that even compete in the fields of content delivery, i. e., media distribution and home media entertainment solutions. ISPs remain a critical point from which to observe what type of content is dominating or emerging for eyeballs, and what mechanisms and technologies, e. g., CDNs or P2P, dominate the network traffic.

In contrast, IXPs are a much more understudied subject. We first described in Section 3.2 the fairly unknown ecosystem of Internet Exchanges and how IXPs contribute to the overall picture. Our findings are that the business of IXPs, i. e., IX, NAP, and IBX, differs across market places, i. e., the EU, US, and Asia. A major reason for this is the different historical development in Internet operation in the different parts of the world. While the IXPs in Europe tend to be carrier- and data-center-neutral, they are strongly connected to either or both entities in the US. To the best of our knowledge, we are among the first to deploy research equipment to collect data and measure different aspects of IXPs, including the comparison with publicly available data. In the course of our work, we have had access to three differently sized IXPs, where we installed measurement equipment for passive data collection. In this work, we have focused on the largest among them, which is a major European exchange point of importance in more than just the local context. The type of data includes flow information and routing data. The flow information provides a sampled subset of all layer-2 frames crossing the public platform of the IXPs in SFLOW format, consisting

of collector statistics and up to 128-byte frame content, including source and destination IP and MAC addresses up to application information. Our main findings from the analysis of IXP data are as follows.
Internet Exchanges are well-localized physical locations and vantage points in the Internet infrastructure where one can observe traffic samples on a true Internet scale. Mining the data collected at one such vantage point reveals a network where heterogeneity abounds in every direction. Given that economic incentives drive many of the main commercial Internet players to either host third-party servers in their own network infrastructure or deploy their own servers, often in massive numbers, in strategically selected third-party networks we expect the observed trend towards increasingly heterogeneous networks and increasingly diverse usage of IXP peering links in particular, and AS links in general, to accelerate, especially in light of the growing importance of cloud providers. The observed heterogeneity calls for a new mental model for the Internet ecosystem that accounts for the observed network heterogenization, points towards measurements that reveal and keep track of this ongoing heterogenization process, and is rich and flexible enough to adapt to a constantly changing Internet environment.

Our analysis at an ISP and some IXPs is based on passive measurements. However, in many cases this information is not sufficient on its own, and therefore needs to be complemented by additional data. One source for achieving this enrichment is active measurements. Examples that we have covered are DNS data and HTTPS certificate data.

Active measurements can be much more than complementary. In Chapter 10, we showed and elaborated on how we collected information about the operational practices of some large operators, i. e., Google, EdgeCast, and CacheFly. Our measurement is efficient in that it can be conducted from any (even residential) broadband access point and it delivers information for which one would otherwise require a dense and globally distributed measurement network – which does not exist. Our findings uncover infrastructure footprints of CDN operators, DNS cache-ability, and user-to-server mappings. Our exploration methodology is confirmed to be unintended by the creators of the EDNS client-subnet extension, it cannot easily be mitigated, and it can easily be replicated.

11.1 Outlook

Our exploration of IXPs is not over – it has only begun. We have mainly looked at the data of the largest of the IXPs where we run measurements. However, we have found inconsistencies with public data that need to be examined closely. For example, we observe peerings in public data that we do not "see" at the IXPs' platform and vice versa.

Our recommendation for main directions of future work points to two aspects:

1. Continuing the work on IXP data and extended it in the dimensions of number of vantage points, data completeness, ground truth, and comparison to public sources.
2. Combining data from the various sources and vantage points towards a better understanding of Internet business relationships and the network itself.

(1) Although we could speculate about information filtering at public vantage points or private interconnects at IXPs, we need to develop a way of dealing with hidden information across different vantage points in order to approach provable truth. After an additional number of peerings were "discovered" in [43], the estimation of the number of peerings can be improved by monitoring route-servers. More work will definitively be done in the direction of examining additional vantage points, both qualitatively and quantitatively. Comparing different measurements of the same entity, i. e., control plane vs. data plane for a route public server, will be valuable in that regard.

(2) The other main direction of further research is a general combination of information, and should be combined with the mentioned exploration of more vantage points. Combination and correlation of data points and data types within a single vantage point are not completed yet as they concern our IXP research. One next step will be the examination of peerings vs. data flow at the IXP in order to learn about the type of usage of peerings among networks. Another interesting aspect is the thorough study of multilateral peerings vs. bilateral peerings, which can, as seen in Chapter 8, go well beyond the scope of the AS level.

Pursuing these two directions of research will allow for a higher quality of measurement data and will eventually contribute to higher-level insights, e. g., risk assessments, availability analysis, and regulatory needs.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Anja Feldmann for her continuous support of my PhD study and related research, and for her patience, motivation, and immense knowledge. Her guidance helped me throughout the entire research and writing of this thesis. Thank you for giving me the right impulses at the right time.

Beside my advisor, I would like to thank the other members of my thesis committee.
Dr. Walter Willinger provided me at all time with helpful comments and encouragement. He also asked those questions which inspired me to widen my research from various perspectives. Many thanks go also to Prof. Dr. Odej Kao and Prof. Dr. Jean-Pierre Seifert as reviewers and members of my committee.

I thank my fellow PhD candidates and postdocs for the stimulating discussions, for the sleepless nights during which we worked together before deadlines, and for all the fun we have had in these last years of cooperation. Particularly Dr. Nikolaos Chatzis has always been a reliable, open minded, and helpful peer and mentor to me.

Finally, I would like to thank my wife Cornelia, who has been overwhelmingly understanding, patient and supportive during the course of my studies, and who has taken good care of Emilia and Tabea. My thesis is dedicated to the three of you.

List of Figures

1.1 Internet Topology Evolution 1 17
1.2 Internet Topology Evolution 2 19

2.1 IPFIX message format 33
2.2 Example of ECS query and response. 34

3.1 AS level views of a topology 39
3.2 Definition of zones for content distribution via the Internet 40
3.3 Edge and Core by Connection End Points 41
3.4 Example of an IXP topology 49
3.5 Number of ISPs per country 53
3.6 Peering without and with Route server 54
3.7 Traffic volumes (daily) of AMS-IX 56
3.8 IXP traffic growth 57
3.9 Typical telco access network structure 62

4.1 Mixed hardware-software RAID60 75
4.2 Separation of Management and Tapping 77
4.3 Optical line tap setups 78
4.4 Schematics of remote monitoring with restricted signaling channels 88
4.5 Rotating measurement 98

5.1 Accessing an ISP measurement site 104
5.2 Monitoring and data flow 105

6.1 Accessing an IXP measurement site 113
6.2 IXP measurement site 114
6.3 IXP measurement site (improved) 114
6.4 Flow collection setup 117
6.5 Different access paths 118

7.1 Running an active measurement setup 120
7.2 Active measurement: architectural overview 121

8.1 Traffic filtering steps 126
8.2 Traffic share of Server IPs 129
8.3 Percentage of IPs per country 132
8.4 Churn of server hosting IPs and ASs 137
8.5 Churn of server traffic by region 139
8.6 Heterogeneity of organizations and ASs 143
8.7 AS link heterogeneity 146

9.1 IXP peering options. 152
9.2 Route server deployment time line. 152
9.3 BIRD route server: Example setup. 154
9.4 Inferred bi-lateral BGP sessions over time. 159

10.1 Prefix length vs. ECS scope for RIPE and PRES. 167
10.2 Google: User to server mapping (RIPE). 169

List of Tables

1.1 IXP-growth 21

3.1 Overhead components in packet trace data 44
3.2 Traffic and member statistics for AMS-IX 55

8.1 IXP summary statistics 131
8.2 Ranking by traffic and server-IPs 133
8.3 IXP as local yet global player 134

9.1 IXP profiles. 156
9.2 Multi-lateral and bi-lateral connectivity. 158

10.1 ECS adopters: Uncovered footprint. 165

Bibliography

[1] 2012 Report on European IXPs - Euro-IX. https://www.euro-ix.net/documents/1117-Euro-IX-IXP-Report-2012-pdf?download=yes.

[2] A Faster Internet Consortium. http://www.afasterinternet.com.

[3] Alexa top sites. http:///www.alexa.com/topsites.

[4] Amazon CloudFront - Amazon Web Services. http://aws.amazon.com/cloudfront/.

[5] AMS-IX Internet Peering. https://www.ams-ix.net/services-pricing/internet-peering.

[6] AMS-IX Mobile Peering. https://www.ams-ix.net/services-pricing/mobile-peering/grx.

[7] AMS-IX Route Servers. https://www.ams-ix.net/technical/specifications-descriptions/ams-ix-route-servers.

[8] AT&T Company Information – Networks. http://www.att.com/gen/investor-relations?pid=5711.

[9] CIDR Report. http://www.cidr-report.org/.

[10] Data Center Map. http://www.datacentermap.com/.

[11] DE-CIX Apollon. http://apollon.de-cix.net.

[12] Deutsche Telekom International Carrier Sales & Solutions (ICSS) – IP Transit. http://www.telekom-icss.com/iptransit.

[13] ECIX Launches First German 100GE Internet Exchange in Frankfurt. http://www.ecix.net/news/ecix-launches-first-german-100ge-internet-exchange-in-frankfurt.

[14] Euro-IX Resources: Traffic, Reports, and Best Practices. https://www.euro-ix.net/resources.

[15] European Internet Exchange Association. https://www.euro-ix.net.

[16] Google Global Cache. http://ggcadmin.google.com/ggc.

[17] Google Peering and Content Delivery. https://peering.google.com/about/peering_policy.html.

[18] Google Public DNS. https://developers.google.com/speed/public-dns.

[19] Internet2 Network Research Data. http://noc.net.internet2.edu/i2network/research-data.html.

[20] IRR - Internet Routing Registry. http://www.irr.net.

[21] JAP an.on Proxy Server. http://jap.inf.tu-dresden.de/.

[22] Like Netflix, Facebook is boosting its edge network. http://gigaom.com/2012/06/21/like-netflix-facebook-is-planning-its-own-cdn/.

[23] Netflix Open Connect. https://signup.netflix.com/openconnect.

[24] OpenBGPD Border Gateway Protocol. http://www.openbgpd.org/.

[25] OpenDNS. http://www.opendns.com.

[26] Orange and Akamai form Content Delivery Strategic Alliance. http://www.akamai.com/html/about/press/releases/2012/press_112012_1.html.

[27] Packet Clearing House. https://www.pch.net.

[28] Packet Clearing House routing archive. https://www.pch.net/resources/data.php.

[29] PeeringDB. https://www.peeringdb.com.

[30] Quagga Routing Suite. http://www.nongnu.org/quagga/.

[31] RIPE Routing Information Service. http://www.ripe.net/ris/.

[32] Routeviews Project – University of Oregon. http://www.routeviews.org/.

[33] Stratosphere, Big Data project. http://www.stratosphere.eu.

[34] Team Cymru. http://www.team-cymru.org/.

[35] The BIRD Internet Routing Daemon. http://bird.network.cz.

[36] Position Statement on the Review of the International Telecommunication Regulations (ITRs). https://www.euro-ix.net/documents/1042-EuroIX-EC-WCIT-Position-Statement-pdf, October 2012.

[37] Technology Transitions Policy Task Force Seeks Comment on Potential Trials. http://transition.fcc.gov/Daily_Releases/Daily_Business/2013/db0510/DA-13-1016A1.pdf, May 2013.

[38] Ripe atlas probes. https://atlas.ripe.net/, 2014.

[39] samknows - fcc project page. http://www.samknows.com/broadband/fcc_project, 2014.

[40] Adhikari, V., Guo, Y., Hao, F., Varvello, M., Hilt, V., Steiner, M., and Zhang, Z.-L. Unreeling Netflix: Understanding and Improving multi-CDN Movie Delivery. In *Proceedings of the Conference of the IEEE Computer and Communications Societies (INFOCOM)* (2012).

[41] Adhikari, V. K., Jain, S., Chen, Y., and Zhang, Z. L. Vivisecting YouTube: An Active Measurement Study. In *Proceedings of the Conference of the IEEE Computer and Communications Societies (INFOCOM)* (2012).

[42] Aditya, P., Zhao, M., Lin, Y., Haeberlen, A., Druschel, P., Maggs, B., and Wishon, B. Reliable Client Accounting for Hybrid Content-Distribution Networks. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI)* (2012).

[43] Ager, B., Chatzis, N., Feldmann, A., Sarrar, N., Uhlig, S., and Willinger, W. Anatomy of a Large European IXP. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)* (2012).

[44] Ager, B., Mühlbauer, W., Smaragdakis, G., and Uhlig, S. Comparing DNS Resolvers in the Wild. In *Proceedings of the Internet Measurement Conference (IMC)* (2010).

[45] Ager, B., Mühlbauer, W., Smaragdakis, G., and Uhlig, S. Web Content Cartography. In *Proceedings of the Internet Measurement Conference (IMC)* (2011).

[46] Aggarwal, V., Akonjang, O., and Feldmann, A. Improving user and isp experience through isp-aided p2p locality. In *Proceedings of 11th IEEE Global Internet Symposium 2008 (GI '08)* (Washington, DC, USA, April 2008), IEEE Computer Society.

[47] Akamai. Facts and Figures: Network Deployment. `http://www.akamai.com/html/about/facts_figures.html`.

[48] Alizadeh, M., Yang, S., Katti, S., McKeown, N., Prabhakar, B., and Shenker, S. Deconstructing datacenter packet transport. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks* (New York, NY, USA, 2012), HotNets-XI, ACM, pp. 133–138.

[49] Amazon. AWS Dashboard. `http://status.aws.amazon.com/`.

[50] Amazon. EC2 Public IP ranges. `https://forums.aws.amazon.com/ann.jspa?annID=1701`.

[51] Augustin, B., Krishnamurthy, B., and Willinger, W. IXPs: Mapped? In *Proceedings of the Internet Measurement Conference (IMC)* (2009).

[52] Axelrod, M. The Value of Content Distribution Networks. AfNOG 9, 2008.

[53] Bates, T. Implementation of a Route Server for Policy Based Routing across the GIX Project, 1993.

[54] Benson, T., Akella, A., and Maltz, D. A. Network traffic characteristics of data centers in the wild. In *Proceedings of the Internet Measurement Conference (IMC)* (New York, NY, USA, 2010), IMC '10, ACM, pp. 267–280.

[55] Bermudez, I., Mellia, M., Munaf, M., Keralapura, R., and Nucci, A. DNS to the Rescue: Discerning Content and Services in a Tangled Web. In *Proceedings of the Internet Measurement Conference (IMC)* (2012).

[56] Bischof, Z. S., Otto, J. S., Sánchez, M. A., Rula, J. P., Choffnes, D. R., and Bustamante, F. E. Crowdsourcing isp characterization to the network edge. In *Proceedings of the First ACM SIGCOMM Workshop on Measurements Up the Stack* (New York, NY, USA, 2011), W-MUST '11, ACM, pp. 61–66.

[57] Blog, F. A Continued Commitment to Security. http://www.facebook.com/blog/blog.php?post=486790652130.

[58] Blog, G. S. Making Search More Secure. http://googleblog.blogspot.com/2011/10/making-search-more-secure.html.

[59] Blog, N. N. Netflix Launches Today in Sweden, Denmark, Norway, Finland. http://nordicsblog.netflix.com/2012/10/.

[60] Blunk, L., Karir, M., and Labovitz, C. Multi-Threaded Routing Toolkit (MRT) Routing Information Export Format. RFC 6396 (Proposed Standard), Oct 2011.

[61] Boschi, E., Mark, L., Quittek, J., Stiemerling, M., and Aitken, P. IP Flow Information Export (IPFIX) Implementation Guidelines. RFC 5153 (Informational), Apr 2008.

[62] BRD. Telekommunikationsgesetz [TKG], 2004. http://www.gesetze-im-internet.de/tkg_2004/BJNR119000004.html.

[63] Cai, X., Heidemann, J., Krishnamurthy, B., and Willinger, W. Towards an AS-to-Organization Map. In *Proceedings of the Internet Measurement Conference (IMC)* (2010).

[64] Calder, M., Fan, X., Hu, Z., Katz, E.-B., Heidemann, J., and Govindan, R. Mapping the Expansion of Google's Serving Infrastructure. In *Proceedings of the Internet Measurement Conference (IMC)* (2013).

[65] Chatzis, N., Smaragdakis, G., Böttger, J., Krenc, T., and Feldmann, A. On the Benefits of Using a Large IXP as an Internet Vantage Point . In *Proceedings of the Internet Measurement Conference (IMC)* (2013).

[66] Chatzis, N., Smaragdakis, G., and Feldmann, A. On the Importance of Internet eXchange Points for Today's Internet Ecosystem. http://arxiv-web3.library.cornell.edu/abs/1307.5264v2.

[67] Chatzis, N., Smaragdakis, G., Feldmann, A., and Willinger, W. There is More to IXPs than Meets the Eye. *ACM SIGCOMM Computer Communication Review 43*, 5 (2013).

[68] Cisco. Visual Networking Index (VNI) and Forecast. http://www.cisco.com/en/US/netsol/ns827/networking_solutions_sub_solution.html.

[69] Claise, B. Cisco Systems NetFlow Services Export Version 9. RFC 3954 (Informational), Oct 2004.

[70] Claise, B. Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. RFC 5101 (Proposed Standard), Jan 2008. Obsoleted by RFC 7011.

[71] Claise, B., Johnson, A., and Quittek, J. Packet Sampling (PSAMP) Protocol Specifications. RFC 5476 (Proposed Standard), Mar 2009.

[72] Clark, D. D., Sollins, K. R., and Wroclawski, J. Tussle in cyberspace: Defining tomorrow's internet, 2002.

[73] Colitti, L., Gunderson, S. H., Kline, E., and Refice, T. Evaluating IPv6 adoption in the Internet. In *Proceedings of the International Conference on Passive and Active Measurement (PAM)* (2010).

[74] Communication, P.

[75] Contavalli, C., van der Gaast, W., Leach, S., and Lewis, E. Client subnet in DNS requests (IETF draft). `http://tools.ietf.org/html/draft-vandergaast-edns-client-subnet-01`.

[76] Cooper, D., Santesson, S., Farrell, S., Boeyen, S., Housley, R., and Polk, W. Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. RFC 5280 (Proposed Standard), May 2008. Updated by RFC 6818.

[77] Dingledine, R., Mathewson, N., and Syverson, P. Tor: The second-generation onion router. In *Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13* (Berkeley, CA, USA, 2004), SSYM'04, USENIX Association, pp. 21–21.

[78] Filip, O. BIRD's flight from Lisbon to Prague. RIPE 60.

[79] Fixme. Akamai and AT&T Forge Global Strategic Alliance to Provide Content Delivery Network Solutions. `http://www.akamai.com/html/about/press/releases/2012/press_120612.html`.

[80] Flach, T., Dukkipati, N., Terzis, A., Raghavan, B., Cardwell, N., Cheng, Y., Jain, A., Hao, S., Katz, E.-B., and Govindan, R. Reducing Web Latency: the Virtue of Gentle Aggression. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)* (2013).

[81] Foundation, M. Publicsuffix.org. `http://publicsuffix.org/`.

[82] Frank, B., Poese, I., Lin, Y., Smaragdakis, G., Feldmann, A., Maggs, B., Rake, J., Uhlig, S., and Weber, R. Pushing cdn-isp collaboration to the limit. *ACM SIGCOMM Computer Communication Review 43*, 3 (Jul 2013), 34–44.

[83] Frank, B., Poese, I., Smaragdakis, G., Uhlig, S., and Feldmann, A. Content-aware traffic engineering. In *Proceeings of ACM SIGMETRICS 2012 (poster session)* (New York, NY, USA, June 2012), ACM, pp. 413–414. Poster session; also published as Perfomance Evaluation Review, 40(1):413–414, http://dx.doi.org/10.1145/2318857.2254819.

[84] Fusco, F., Dimitropoulos, X., Vlachos, M., and Deri, L. Indexing Million of Packets per Second using GPUs. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)* (2013).

[85] Gerber, A., and Doverspike, R. Traffic Types and Growth in Backbone Networks. In *OFC/NFOEC* (2011).

[86] Giotsas, V., and Zhou, S. Improving the Discovery of IXP Peering Links through Passive BGP Measurements. In *Proceedings of the IEEE Global Internet Symposium* (2013).

[87] Giotsas, V., Zhou, S., Luckie, M., and Claffy, K. Inferring Multilateral Peering. In *Proceedings of the ACM Conference on Emerging Networking Experiments And Technologies (CoNEXT)* (2013).

[88] Google. What is 1e100.net? `http://support.google.com/bin/answer.py?hl=en&answer=174717`.

[89] Govindan, R., Alaettinoglou, C., Varadhan, K., and Estrin, D. Route Servers for Interdomain Routing. *Com. Networks 30* (1998).

[90] He, Y., Siganos, G., Faloutsos, M., and Krishnamurthy, S. A systematic framework for unearthing the missing links: Measurements and Impact. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI)* (2007).

[91] Hilliard, N., Jasinska, E., Raszuk, R., and Bakker, N. Internet Exchange Route Server Operations. IETF draft, draft-ietf-grow-ix-bgp-route-srver-operations-01, 2013.

[92] Hohlfeld, O. *Impact of buffering on quality of experience*. Dissertation, Technische Universität Berlin, 2013.

[93] Huang, C., Wang, A., Li, J., and Ross, K. Measuring and Evaluating Large-scale CDNs. In *Proceedings of the Internet Measurement Conference (IMC)* (2008).

[94] Hughes, M. Route Servers at IXPs – Bugs and Scaling issues with Quagga. UKNOF 13.

[95] Jasinska, E., Hilliard, N., Raszuk, R., and Bakker, N. Internet Exchange Route Server. IETF draft, draft-ietf-idr-ix-bgp-route-server-03, 2013.

[96] Jasinska, E., and Malayter, C. (Ab)Using Route Servers. NANOG 48.

[97] Josephsen, D. *Building a Monitoring Infrastructure with Nagios*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2007.

[98] Khan, A., Kwon, T., Kim, H. C., and Choi, Y. AS-level Topology Collection through Looking Glass Servers. In *Proceedings of the Internet Measurement Conference (IMC)* (2013).

[99] Kim, J., Schneider, F., Ager, B., and Feldmann, A. Today's usenet usage: Characterizing NNTP traffic. In *Proceedings of the IEEE Global Internet Symposium* (Mar. 2010).

[100] Knowledge, D. C. Who Has the Most Web Servers? http://www.datacenterknowledge.com/archives/2009/05/14/whos-got-the-most-web-servers.

[101] Kreibich, C. Broccoli: The bro client communications library. http://www.cl.cam.ac.uk/~cpk25/broccoli/manual/, 2004.

[102] Kreibich, C., Weaver, N., Nechaev, B., and Paxson, V. Netalyzr: illuminating the edge network. In *Internet Measurement Conference* (2010), M. Allman, Ed., ACM, pp. 246–259.

[103] Krishnamurthy, B., and Wang, J. On Network-aware Clustering of Web Clients. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)* (2001).

[104] Krishnan, R., Madhyastha, H., Srinivasan, S., Jain, S., Krishnamurthy, A., Anderson, T., and Gao, J. Moving Beyond End-to-end Path Information to Optimize CDN Performance. In *Proceedings of the Internet Measurement Conference (IMC)* (2009).

[105] Labovitz, C., Iekel, S.-J., McPherson, D., Oberheide, J., and Jahanian, F. Internet Inter-Domain Traffic. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)* (2010).

[106] Labrinidis, A., and Nguyenduy, E. Route Server Implementations Performance. 20th Euro-IX Forum, April 2012.

[107] Lee, C., Lee, D. K., and Moon, S. Unmasking the growing udp traffic in a campus network. In *Proceedings of the International Conference on Passive and Active Measurement (PAM)* (Berlin, Heidelberg, 2012), PAM'12, Springer-Verlag, pp. 1–10.

[108] Leighton, T. Improving Performance on the Internet. *Communications of the ACM 52*, 2 (2009), 44–51.

[109] Levin, D., Wundsam, A., Mehmood, A., and Feldmann, A. Berlin: The berlin experimental router laboratory for innovative networking. In *Proceedings of the 6th International Conference on Testbeds and Research Infrastsructures for the Development of Networks and Communities (TridentCom 2010, poster session)* (Berlin / Heidelberg / New York, May 2010), T. Magedanz, A. Gavras, N. H. Thanh, and J. S. Chase, Eds., vol. 46 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST)*, Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer, pp. 602–604. Poster.

[110] Liebowitz, S. J., and Margolis, S. E. Network Externality: An Uncommon Tragedy. *J. Econ. Perspectives 8*, 2 (1994).

[111] Luckie, M., Huffaker, B., Dhamdherei, A., Giotsas, V., and kc claffy. AS Relationships, Customers Cones, and Validations. In *Proceedings of the Internet Measurement Conference (IMC)* (2013).

[112] Mahimkar, A., Chiu, A., Doverspike, R., Feuer, M. D., Magill, P., Mavrogiorgis, E., Pastor, J., Woodward, S. L., and Yates, J. Bandwidth on demand for inter-data center communication. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks* (New York, NY, USA, 2011), HotNets-X, ACM, pp. 24:1–24:6.

[113] Maier, G., Feldmann, A., Paxson, V., and Allman, M. On dominant characteristics of residential broadband internet traffic. In *IMC '09: Proceedings of the 2009 Internet Measurement Conference* (New York, NY, USA, November 2009), ACM Press, pp. 90–102.

[114] Maier, G., Feldmann, A., Paxson, V., Sommer, R., and Vallentin, M. An assessment of overt malicious activity manifest in residential networks. In *Proceedings of the eighth Conference on Detection of Intrusions and Malware and Vulnerability Assessment (DIMVA '11)* (Berlin / Heidelberg, Germany, July 2011), vol. 6739 of *Lecture Notes in Computer Science (LNCS)*, IEEE, Springer, pp. 144–163.

[115] Maier, G., Schneider, F., and Feldmann, A. A first look at mobile hand-held device traffic. In *Proceedings of the International Conference on Passive and Active Measurement (PAM)* (2010).

[116] Maier, G., Schneider, F., and Feldmann, A. Nat usage in residential broadband networks. In *Proceedings of the International Conference on Passive and Active Measurement (PAM)* (Berlin / Heidelberg, Germany, 2011), N. Spring and G. F. Riley, Eds., vol. 6579 of *Lecture Notes in Computer Science (LNCS)*, Springer, pp. 32–41.

[117] Maier, G., Sommer, R., Dreger, H., Feldmann, A., Paxson, V., and Schneider, F. Enriching network security analysis with time travel. *SIGCOMM Comput. Commun. Rev. 38*, 4 (Aug 2008), 183–194.

[118] Malayter, C. Route Servers, Mergers, Features, & More. NANOG 51.

[119] Mao, Z., Cranor, C., Douglis, F., Rabinovich, M., Spatscheck, O., and Wang, J. A Precise and Efficient Evaluation of the Proximity Between Web Clients and Their Local DNS Servers. In *USENIX ATC* (2002).

[120] Maxmind. GeoLite Country. `http://dev.maxmind.com/geoip/legacy/geolite`.

[121] Mehmood, A., Hohlfeld, O., Levin, D., Wundsam, A., Ciucu, F., Schneider, F., Feldmann, A., and Braun, R.-P. The routerlab: Emulating internet characteristics in a room. In *Proceedings of 11th ITG Conference on Photonic Networks (11. ITG-Fachtagung Photonische Netze)* (Berlin / Offenbach, Germany, May 2010), VDE-Verlag, pp. 201–208. Poster Session.

[122] Mehmood, A., Sarrar, N., Uhlig, S., and Feldmann, A. How happy are your flows: an empirical study of packet losses in router buffers. Technical report, Technische Universität Berlin, Fakultät für Elektrotechnik und Informatik, Berlin, Germany, May 2012. No. 2012/07.

[123] Neuman, C., Yu, T., Hartman, S., and Raeburn, K. The Kerberos Network Authentication Service (V5). RFC 4120 (Proposed Standard), Jul 2005. Updated by RFCs 4537, 5021, 5896, 6111, 6112, 6113, 6649, 6806.

[124] Nygren, E., Sitaraman, R. K., and Sun, J. The Akamai Network: A Platform for High-performance Internet Applications. *SIGOPS Oper. Syst. Rev.* (2010).

[125] Oliveira, R., Pei, D., Willinger, W., Zhang, B., and Zhang, L. The (in)completeness of the observed internet as-level structure. *IEEE/ACM Trans. Netw. 18*, 1 (Feb 2010), 109–122.

[126] Oliveira, R., Pei, D., Willinger, W., Zhang, B., and Zhang, L. The (In)completeness of the Observed Internet AS-Level Structure. *IEEE/ACM Trans. Netw. 18*, 1 (2010).

[127] Otto, J. S., Sánchez, M. A., Rula, J. P., and Bustamante, F. E. Content delivery and the natural evolution of DNS - Remote DNS Trends, Performance Issues and Alternative Solutions. In *Proceedings of the Internet Measurement Conference (IMC)* (2012).

[128] Pang, R., Paxson, V., Sommer, R., and Peterson, L. Binpac: A yacc for writing application protocol parsers. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement* (New York, NY, USA, 2006), IMC ’06, ACM, pp. 289–300.

[129] Paxson, V. Bro: A system for detecting network intruders in real-time. *Computer Networks Journal 31*, 23–24 (1999). Bro homepage: `www.bro-ids.org`.

[130] Phaal, P., Panchen, S., and McKee, N. InMon Corporation’s sFlow: A Method for Monitoring Traffic in Switched and Routed Networks. RFC 3176 (Informational), Sep 2001.

[131] Plonka, D., and Barford, P. Flexible Traffic and Host Profiling via DNS Rendezvous. In *SATIN* (2011).

[132] Poese, I., Frank, B., Ager, B., Smaragdakis, G., and Feldmann, A. Improving Content Delivery using Provider-aided Distance Information. In *Proceedings of the Internet Measurement Conference (IMC)* (2010).

[133] Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., and Gueye, B. IP Geolocation Databases: Unreliable? *ACM SIGCOMM Computer Communication Review 41*, 2 (2011).

[134] Polfliet, S., Ryckbosch, F., and Eeckhout, L. Optimizing the datacenter for data-centric workloads. In *Proceedings of the International Conference on Supercomputing* (New York, NY, USA, 2011), ICS ’11, ACM, pp. 182–191.

[135] Reggani, A., Schneider, F., and Teixeira, R. An end-host view on local traffic at home and work. In *Proceedings of Passive and Active Measurements Conference (PAM ’12)* (March 2012).

[136] Richter, P., Smaragdakis, G., Chatzis, N., Böttger, J., Feldmann, A., and Willinger, W. Peering at Peerings: On the Role of IXP Route Servers. In *Proceedings of the Internet Measurement Conference (IMC)* (2014).

[137] Roesch, M. Snort – lightweight intrusion detection for networks. In *Proceedings of the Systems Administration Conference (LISA)* (1999).

[138] Roughan, M., Willinger, W., Maennel, O., Pertouli, D., and Bush, R. 10 Lessons from 10 Years of Measuring and Modeling the Internet's Autonomous Systems. *IEEE Journal Selected Areas in Communications 29*, 9 (2011).

[139] University of oregon route views project. http://www.routeviews.org/.

[140] Sánchez, M. A., Otto, J. S., Bischof, Z. S., Choffnes, D. R., , Bustamante, F. E., Krishnamurthy, B., and Willinger, W. Dasu: Pushing Experiments to the Internet's Edge. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI)* (2013).

[141] Sandvine. Global Internet Phenomena Report. http://www.sandvine.com/news/global_broadband_trends.asp.

[142] Sarrar, N. Ein bittorrent analyzer für das bro nids. Projekt, Technische Universität Berlin, Berlin, Germany, February 2008.

[143] Schneider, F., Agarwal, S., Alpcan, T., and Feldmann, A. The new web: Characterizing ajax traffic. In *Proceedings of the International Conference on Passive and Active Measurement (PAM)* (New York, NY, USA, April 2008), vol. 4979 of *Lecture Notes in Computer Science*, Springer-Verlag Berlin Heidelberg, pp. 31–40.

[144] InMon – sFlow. http://sflow.org/.

[145] Stapleton-Gray, R., and Woodcock, W. National Internet Defense – Small States on the Skirmish Line. *Communications of the ACM 54*, 3 (2011).

[146] Streibelt, F., Böttger, J., Chatzis, N., Smaragdakis, G., and Feldmann, A. Exploring EDNS-Client-Subnet Adopters in your Free Time. In *Proceedings of the Internet Measurement Conference (IMC)* (2013).

[147] Sundaresan, S., de Donato, W., Feamster, N., Teixeira, R., Crawford, S., and Pescapè, A. Measuring home broadband performance. *Commun. ACM 55*, 11 (Nov 2012), 100–109.

[148] Tariq, M., Zeitoun, A., Valancius, V., Feamster, N., and Ammar, M. Answering What-if Deployment and Configuration Questions with Wise. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)* (2009).

[149] Trestian, I., Ranjan, S., Kuzmanovic, A., and Nucci, A. Unconstrained Endpoint Profiling (Googling the Internet). In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)* (2008).

[150] Triukose, S., Al, Z.-Q., and Rabinovich, M. Content Delivery Networks: Protection or Threat? In *ESORICS* (2009).

[151] Triukose, S., Wen, Z., and Rabinovich, M. Measuring a Commercial Content Delivery Network. In *World Wide Web Journal* (2011).

[152] Vallentin, M., Sommer, R., Lee, J., Leres, C., Paxson, V., and Tierney, B. The nids cluster: Scalable, stateful network intrusion detection on commodity hardware. In *Proceedings of the Symposium on Recent Advances in Intrusion Detection (RAID)* (2007).

[153] Vixie, P. Extension Mechanisms for DNS (EDNS0). RFC 2671 (Proposed Standard), Aug 1999. Obsoleted by RFC 6891.

[154] Vixie, P. DNS Complexity. *ACM Queue 5*, 3 (2007), 24–29.

[155] Vixie, P. What DNS is Not. *Commun. ACM 52*, 12 (2009), 43–47.

[156] Wilson, C., Ballani, H., Karagiannis, T., and Rowtron, A. Better never than late: Meeting deadlines in datacenter networks. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)* (New York, NY, USA, 2011), SIGCOMM '11, ACM, pp. 50–61.