Correlating flow-based network measurements for service monitoring and network troubleshooting

Author(s):
Schatzmann, Dominik

Publication Date:
2013

Permanent Link:

https://doi.org/10.3929/ethz-a-009786857 →

Diss. ETH No. 20868
TIK-Schriftenreihe Nr. 136

Correlating Flow-based Network Measurements for Service Monitoring and Network Troubleshooting

A dissertation submitted to
ETH Zurich

for the degree of
Doctor of Sciences

presented by

DOMINIK SCHATZMANN

Master of Science ETH in Electrical Engineering and
Information Technology
born July 30, 1981
citizen of Zurich, ZH

accepted on the recommendation of
Prof. Dr. Bernhard Plattner, examiner
Dr. Wolfgang Mühlbauer, co-examiner
Dr. Xenofontas Dimitropoulos, co-examiner
Prof. Dr. Rolf Stadler, co-examiner

2012

Abstract

The resilience of network services is continuously challenged by component failures, mis-configured devices, natural disasters and malicious users. Therefore, it is an important but unfortunately difficult task of network operators and service administrators to carefully manage their infrastructure in order to ensure high availability. In this thesis we contribute novel service monitoring and troubleshooting applications based on flow-based network measurements to help operators to address this challenge.

Flow-level measurement data such as IPFIX or NetFlow typically provides statistical summaries about connections crossing a network including the number of exchanged bytes and packets. Flow-level data can be collected by off-the-shelf hardware used in backbone networks. It allows Internet Service Providers (ISPs) to monitor large-scale networks with a limited number of sensors. However, the range of security or network management related questions that can be answered directly by using flow-based data is strongly limited by the fact that only a small amount of information is collected per connection. In this work, we overcome this problem by correlating and analyzing sets of flows across different dimensions such as time, address space, or user groups. This hidden information proves very beneficial for flow-based troubleshooting applications.

Using such an approach, we show how flow-based data can be instrumented to effectively support mail administrators in fighting spam. In more detail, we demonstrate that certain spam filtering decisions performed by mail servers can be accurately tracked at the ISP-level using flow-level data. Then, we argue that such aggregated knowledge from multiple e-mail domains does not only allow ISPs to remotely monitor what their "own" servers are doing, but also to develop and evaluate new scalable methods for fighting spam.

To assist network operators with troubleshooting connectivity problems,

we contribute FACT, a system that implements a flow-based approach for connectivity tracking that helps network operators to continuously track connectivity from their network towards remote autonomous systems, networks, and hosts. In contrast to existing solutions, our approach correlates solely flow-level data, is capable of online data processing, and is highly efficient in alerting only about those events that actually affect the studied network. Most important, FACT precisely reports which address spaces are currently not reachable by which clients – an information required to efficiently react on connectivity issues.

In order to introduce such innovative and productive troubleshooting applications, we improved the entire value chain from low-level data processing all the way up to knowledge extraction.

First, we contribute FlowBox, a modular flow processing library. Its design exploits parallelization and is capable of processing large volumes of data to deliver statistics, reports, or on-line alerts to the user with limited delay. In addition we address the challenge of dissecting large data volumes in real-time into service-related flow sets. Our method identifies Server Sockets (SeSs), i.e., communication endpoints that offer a specific network service to multiple clients (e.g., webmail service, Skype, DNS service). Our approach is application-agnostic, copes with errors of flow data (e.g. imprecise timing) and greatly reduces Internet background radiation.

Second, to infer the underlying network application behind a SeS, we show how correlations across flows, protocols, and time can be leveraged for novel techniques to classify the service provided by a SeS. Although we limit our study to the classification of webmail services such as Outlook Web Access, Horde, or Gmail, our approach is very promising for other classes of network applications, too. Furthermore, we discuss how to achieve service classification with reasonable accuracy and coverage, yet limited effort. Our approach is based on the idea that the service classification of a SeS stays stable across multiple flows and across different users accessing the SeS. This allows us to reduce the overall collection and processing efforts by applying a time and address space sampling scheme during the data mining.

We evaluated our work using flow traces collected over the last 8 years at the backbone of an ISP that covers approximately 2.4 million IP addresses. The large-scale nature of our measurements is underlined by high peak rates of more than 80,000 flows per second, 3 million packets per second, and more than 20 Gbit/s of traffic.

Kurzfassung

Kommunikationsnetze werden kontinuierlich durch Ausfälle von Komponenten, falsch konfigurierten Geräte, Naturkatastrophen oder böswilligen Benutzer herausgefordert. Daher ist es eine wichtige, aber leider schwierige, Aufgabe der Netzbetreiber und Service-Administratoren, ihre Infrastruktur sorgfältig zu verwalten bei gleichzeitiger Gewährleistung einer hohen Verfügbarkeit der Dienste. Um den Betreibern zu helfen, diese Herausforderung besser zu meistern, stellt diese Arbeit neue flow-basierte Anwendungen zur Netzwerk-Fehlerdiagnose vor.

Flow-basierte Messdaten wie IPFIX oder NetFlow liefern eine statistische Zusammenfassung der Verbindungen, welche ein Netzwerkelement durchqueren. Solche Flowdaten können von üblicher Hardware an strategischen Punkten in Netzen aufgezeichnet werden und ermöglichen dadurch eine effiziente Überwachung auch von grossen Netzen. Jedoch können nur wenige Fragestellungen direkt mit Flowdaten beantwortet werden, da pro Verbindung nur wenig Information aufgezeichnet wird. Wir überwinden dieses Problem durch das Korrelieren und Analysieren von Gruppen von Flows über diverse Dimensionen wie Zeit, Adressraum oder Benutzergruppen. Dadurch erschliessen wir bisher ungenützte Information für Anwendungen zur Netzwerk-Fehlerdiagnose.

In dieser Arbeit zeigen wir zunächst, wie Flowdaten korreliert werden können, um Mail-Administratoren bei der Bekämpfung von Spam zu helfen. Genauer zeigen wir, dass die Entscheidung eines einzelnen Mailservers, eine Mail zu filtern, in den Flowdaten sichtbar ist. Dies erlaubt es solche Entscheidungen auf der ISP Stufe zu sammeln und zu vergleichen. Das gesammelte Wissen kann nicht nur zur Leistungsoptimierung diverser Einstellungen der eigenen Server, sondern auch für die Entwicklung und Evaluierung neuer skalierbarer Methoden zur Bekämpfung von Spam verwendet werden.

Weiter zeigen wir, wie Flowdaten korreliert werden können, um Netzwerk-Administratoren bei der Identifikation von Konnektivitäts-Problemen zu helfen. Im Gegensatz zu existierenden Lösungen setzt der vorgestellte Ansatz ausschließlich auf Flowdaten. Überdies ist das Verfahren sehr effizient bei der Entdeckung von genau solchen Konnektivitäts-Problemen, welche die Benutzer auch tatsächlich betreffen.

Damit solche innovativen flow-basierten Anwendungen entwickelt werden können, muss die gesamte Wertschöpfungskette von der low-level Datenverarbeitung bis hin zur Sammlung des Wissens verbessert werden.

Dazu präsentieren wir im ersten Teil der Arbeit unseren flow-basierten Ansatz zur Identifizierung von Kommunikationsendpunkten, welche Netzwerkdienste wie z.B. Webmail, Skype oder DNS zur Verfügung stellen. Die Identifikation der Endpunkte ist notwendig, um grosse Mengen an Flowdaten effizient in kleinere,dienstspezifische Mengen zu unterteilen. Die Auswertung des Ansatzes mit Hilfe gesammelter Flow Daten eines grossen Netzes belegt, dass unser Ansatz für die Echtzeit-Verarbeitung von Messdaten geeignet ist.

Im zweiten Teil werden die Eigenschaften von Kommunikationsendpunkten untersucht und Verfahren entwickelt, die den Dienst bestimmen, der von einem Endpunkt angeboten wird. Mithilfe dieser Information kann später die eigentliche Problemdiagnose vereinfacht werden, da man gezielt Flows bestimmter Netzwerkdienste untersuchen kann. Zum Beispiel wird gezeigt, wie die Korrelationen zwischen Benutzern, Protokollen, und Zeitinformationen der Flows genutzt werden kann, um neue Methoden zur Bestimmung des Dienstes zu entwickeln. Obwohl diese Studie auf die Identifizierung von Webmail-Diensten wie Outlook Web Access, Horde oder Google Mail beschränkt ist, ist der Ansatz vielversprechend, auch für andere Dienste. Ferner wird diskutiert, wie sich der Aufwand zur Bestimmung des Dienstes reduzieren lässt, ohne zuviele Details Preis zu geben. Unser Ansatz beruht darauf, dass der gleiche Dienst über eine längere Zeit für verschiedene Benutzer angeboten wird. Somit kann die Bestimmung dieses Dienstes auch anhand einer Stichprobe erfolgen. Durch Zwischenspeichern des Resultats kann eine Wiederholung des ressourcenintensiven Arbeitsschritts der Bestimmung des Diensts verhindert werden.

Wir haben unsere Arbeit mit Hilfe von Flowdaten ausgewertet, welche über die letzten 8 Jahre im Netz eines Schweizer ISPs aufgezeichnet wurden. Dessen Netz beherbergt circa 2.4 Millionen IP Adressen und erreicht Übertragungraten von von mehr 20 Gbit/s und Flowraten von mehr als $80,000$ Flows/s .