International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 07, July 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Large Scale Data Analytics of User Performance for Humanizing Substance Deliverance

S.Joseph Gabriel[1], C.Venkatesan[2], P.Rizwan Ahmed[3]

[1]*Associate Professor & Head of Computer Science, Mazharul Uloom College, Ambur (India)*

[2]*Research Scholar, Mazharul Uloom College, Ambur (India)*

[3]*Assistant Professor & Head of Computer Application, Mazharul Uloom College, Ambur (India)*

ABSTRACT

The Internet is fast becoming the de facto content delivery network of the world, supplanting TV and physical media as the primary method of distributing larger files to ever-increasing numbers of users at the fastest possible speeds. Recent trends have, however, posed challenges to various players in the Internet content delivery ecosystem. These trends include exponentially increasing traffic volume, increasing user expectation for quality of content delivery, and the ubiquity and rise of mobile traffic.

For example, exponentially increasing traffic—primarily caused by the popularity of Internet video—is stressing the existing Content Delivery Network (CDN) infrastructures. Similarly, content providers want to improve user experience to match the increasing user expectation in order to retain users and sustain their advertisement based and subscription-based revenue models. Finally, although mobile traffic is increasing, cellular networks are not as well designed as their wireline counterparts, causing poorer quality of experience for mobile users. These challenges are faced by content providers, CDNs and network operators everywhere and they seek to design and manage their networks better to improve content delivery and provide better quality of experience.

Index Terms: data analytics, machine learning, user behavior, user experience, content delivery

I. INTRODUCTION

Internet today is largely a content driven network. Starting from simple data transfer between two computers directly connected by a wire, the complexity of content delivery over the Internet has come a long way to include several complex applications such as adaptive video streaming, peer-to-peer file sharing, massively multiplayer online gaming, cloud storage, and cloud-based computation. Over the years, there have been several innovations to support the growth of content delivery, both in protocols used for delivering content, as well as in the infrastructure to support and improve new content delivery applications. Not able protocol innovations include optical technology for very high speed connectivity at the physical level , high-speed variants of IEEE 802.3 Ethernet specification at the data link layer [1], IP multicast at the network layer, specialized transport layer mechanisms such as SCTP and DCCP specifically for streaming media and video, and application layer protocols such as Real Time Streaming Protocols and Dynamic Adaptive Streaming over HTTP [12]. There have also been several infrastructure innovations, especially in the design of content delivery systems. Beginning with caches placed in front of content servers to return frequently-requested content, content providers began distributing these caches globally close to the request origin using Content Distribution

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Networks (CDNs). Other content delivery system design innovations include peer to peer content delivery such as BitTorrent and its variants, and hybrid P2P-CDNs, where, in addition to the content servers, clients of the CDN also contribute content they have downloaded to peer clients.

These protocol and infrastructure innovations have greatly improved content delivery over the Internet making it very robust. Today, as a result, tens of millions of people consider the Internet to be a necessity. They work, bank, communicate, plan travel, find food, and seek entertainment using Internet-based services. However, with the ubiquity of Internet-connected devices, and with online services demanding high bandwidths and low latency, there are challenges faced by all players in the ecosystem. Users expect instant, high-quality connectivity to their services from any device, and the players in the content delivery ecosystem know that better performance is tightly correlated with higher revenues. But challenges such as increasing load on the network and heterogeneity of technologies especially in cellular data networks (UMTS, LTE) may not provide the best quality of experience to users at all times and on all devices. The players in the ecosystem know that it could be costly to not deal with these challenges ahead of time. For instance, a single second of downtime for a content delivery service like Google or Amazon costs these services several hundred thousands of dollars in revenue.

II. CONTENT DELIVERY ECOSYSTEM

We begin with a brief overview of the different players in the content delivery ecosystem in the Internet today.

Each of these players have access to rich data that can be used towards improving content delivery.Content providers encompass a wide variety of media and e-commerce players who provide content on the Internet primarily for revenue. These include news websites (e.g., CNN), social networking websites (e.g., Facebook, Yelp), and also video providers (e.g., HBO, ABC). Content providers want to maximize their revenues from subscription-based and advertisement-based business models while trying to minimize content distribution costs. To this end, content providers have business arrangements with CDNs (e.g., Akamai, Limelight) to distribute their content across different geographical locations. Similarly

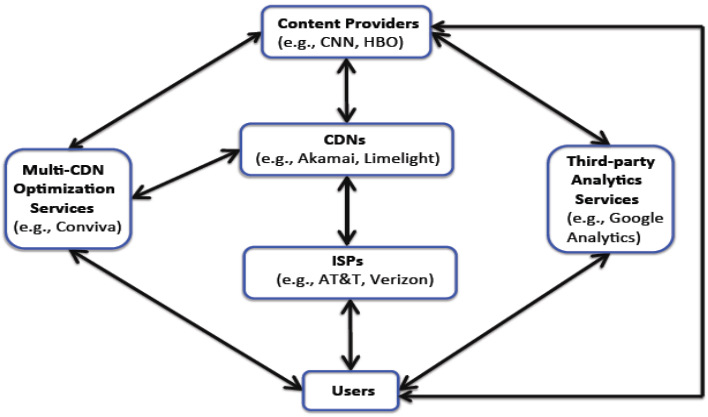

Figure 1.1: Overview of the Internet content delivery ecosystem

more recently they also have contracts with third-party analytics services (e.g., Google Analytics, Ooyala [28]) and optimization services (e.g., Conviva [11]) to understand and

improve user experience. Content Distribution Networks (e.g., Akamai, Limelight) consist of distributed system of servers allocated across different geographical regions for serving content to end users with high performance

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

and availability. CDNs provide content providers a cost-effective mechanism to offload content from their infrastructure. Hence CDNs need to allocate their resources (e.g., server and bandwidth capacity) efficiently across user population. CDNs aim to design their delivery infrastructure to minimize their delivery costs while maximizing their performance. Towards this end there have been many studies and proposals on efficient design of the CDN infrastructure. Although CDNs primarily serve content using dedicated servers operated by them, more recently there have been proposals for other designs including hybrid models that make use of peer-to-peer mechanisms on user-owned devices and also federation across multiple CDNs. CDNs collect a large amount of logs daily on user behavior. Tailoring CDN design based on the user behavior to improve content delivery with minimal costs is an interesting problem faced by CDNs.

Internet Service Providers (ISPs) form the backbone of the Internet by delivering the content from the CDNs and content providers to the end users. Traffic on the Internet has been increasing exponentially over the years. In particular, with the advent of smartphones and new wireless technologies such as 3G and 4G, mobile traffic is on the rise. But, unlike the other players, ISPs do not have access to detailed client-side or server-side logs making it more challenging to extract user behavior information from network traces alone.

III. BIG DATA ANALYTICS

Big data analytics is now extensively used in fields of computer science such as recommendation systems, search and information retrieval, computer vision and image processing, and is making its foray into the real world in terms of business intelligence, healthcare and supply chain analysis. It is also used even within the domain of networks in areas such as network security.

Several technology innovations in the past decade were essential in being able to analyze massive volumes of data. The MapReduce framework is perhaps the innovation that heralded the area of big data analytics, and open-source versions of MapReduce such as Hadoop, and the distributed HDFS filesystem allow researchers to rapidly gather insights from more data that can fit on any single machine. Hadoop, Hive and recent advancements such as Spark make short work of analyzing massive quantities of data. Keeping up with the infrastructure developments, there have been algorithms and libraries that are specifically suited to data mining and machine learning, ranging from the more traditional tools such as Weka and Scikit-learn to tools built for big data such as Graphlab.

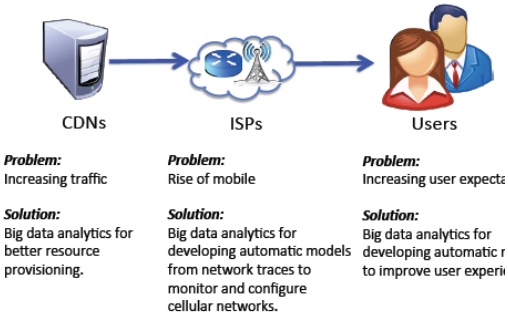

Figure 1.2: Flow of information during content delivery from CDNs to ISPs to Users. We look at how we can use large-scale data analytics to help improve content delivery at each point in the flow.

IV. CONCLUSION

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

In this research paper we showed that applying large scale data analytics is a step forward towards solving some of the main challenges faced by the various players in the content delivery. We showed that large scale data analytics and machine learning algorithms can be used as an effective tool to characterize user behavior in the wild to inform various content delivery system design decisions. This chapter concludes the dissertation with a summary of the approach and contributions followed by a discussion of the lessons learned and remaining open problems in this space.

REFERENCES

[1] 100Gbps Ethernet Task Force. http://www.ieee802.org/3/ba/.

[2] Use RTMFP for developing real-time collaboration applications. http://labs. adobe.com/technologies/cirrus/.

[3] Akamai investor summit 2013. http://www.akamai.com/dl/investors/ 2013_ir_summit_presentation.pdf, .

[4] Akamai Net Session. http://www.akamai.com/client/, .

[5] Driving Engagement for Online Video. http://goo.gl/pO5Cj, .

[6] Census Bureau Divisioning. http://www.census.gov/geo/www/us_regdiv. pdf.

[7] Buyer's Guide: Content Delivery Networks. http://goo.gl/B6gMK.

[8] Cisco Report on CDN Federation - Solutions for SPs and Content Providers To Scale a Great Customer Experience.

[9] Cisco visual networking index: Global mobile data forecast update 2013-2018. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index vni/white_paper_c11-520862.html, .

[10] Cisco study. http://goo.gl/tMRwM, .

[11] Conviva. http://www.conviva.com.

[12] MPEG - DASH. http://dashif.org/mpeg-dash/.

[13] DCCP. https://tools.ietf.org/html/rfc4340.

[14] Firebug. getfirebug.com.

[15] Google page speed. https://developers.google.com/speed/pagespeed/, .

[16] Graphlab. http://graphlab.com/.

[18] Hadoop. http://hadoop.apache.org/.