♦ Cloud Computing: New Opportunities for Telecom Providers

Sape J. Mullender and Thomas L. Wood

Introduction

Since the first electronic computers were built in the 1950s, each decade was characterized by its own model for computing:

- *1950s*. Computers were carefully tended by teams of experts, who would start the computer, load a program, run it and examine the output. This cycle would then repeat itself.
- *1960s*. Batch processing was invented, which allowed one program to run, while results from the previous run would be printed and the next program would be loaded.
- *1970s*. Timesharing was introduced. A computer could be used, interactively, by many users simultaneously. This was the decade of Multics and UNIX* and, of course, IBM's VM370, the grandfather of today's virtualized processors.
- *1980s*. IBM introduced the personal computer and MS-DOS*. For a while, multiprogramming and access control became things of the glorious past.
- *1990s*. Personal computers were hooked up to the Internet. This led to network file systems, distributed systems, and a lot of electronic mail.
- *2000s*. The noughties became the decade of the World Wide Web. We might even characterize it as the decade of the browser, because browsers became our Internet access agents.
- *2010s*. The current decade will be the decade of the cloud, the decade in which our personal devices (which often have become mobile) use services in the Internet to get work done.

Search engines were among the first applications that became Internet services, mostly because searching the whole web for pages meeting certain criteria is just not possible from a personal computer. Now, there are many services that depend on support from large, distributed applications running in the Internet: services supporting social networks, uploading and watching video clips, news reporting, and document sharing.

Dedicated infrastructure for supporting services in the network allows service providers to get closer to their customers and to move services to where they are used. Infrastructure providers can offer computer facilities for hire, giving service providers instant access to additional capacity when the situation demands it.

These possibilities have not gone unnoticed and cloud computing has come to the forefront, both in corporate strategies and in academic research.

Bell Labs Technical Conference

The issue before you was not created through the usual process of refereeing the submitted extended abstracts and full papers. This issue represents an experiment undertaken by the editors of the BLTJ to reduce the time between the submission of material and its presentation to the Bell Labs community.

Extended abstracts were submitted in the spring on 2011 and the accepted papers were then presented at the first Bell Labs Technical Conference, which was held in Antwerp, Belgium on October 18 and 19 of

Bell Labs Technical Journal 17(2), 1–4 (2012) © 2012 Alcatel-Lucent. Published by Wiley Periodicals, Inc.
Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/bltj.21550

2011. All the papers you see in this issue were presented at the conference and both the conference and the process for creating this issue were judged to be a great success, to be repeated, perhaps once a year.

Issue Content

This issue of the *Bell Labs Technical Journal* shows that there is a variety of research related to cloud computing at Bell Labs.

Although the call for contributions was quite broad in scope, we were surprised to find that the recurring theme in all contributions that were selected for publication was, in some sense, performance. There are papers on fault tolerance in cloud computing, on maximizing performance and minimizing latency, and on delivering telecommunications service from the cloud. All are related to questions on the subject of giving performance guarantees.

It was not our intention to look for such papers; it appears that this is the stuff that Bell Labs researchers are concerned about. And, frankly, we're delighted. Dependability—doing the right thing, on time—is important to almost all of our products.

The issue before you contains nine papers, which we've divided into three groups of three. The first group addresses fault tolerance in cloud computing and consists of the following papers.

"Adaptive Application Scaling for Improving Fault-Tolerance and Availability in the Cloud," by Ganesan Radhakrishnan. This paper introduces a model for monitoring the performance of a set of distributed applications competing for shared resources. The goal is to detect potential resource scarcity and act before resources run out, so that the applications will continue to meet their service-level agreements.

"The Inherent Difficulty of Timely Primary-Backup Replication," by Pramod Koppol, Kedar S. Namjoshi, Thanos Stathopoulos, and Gordon T. Wilfong. Primary-backup replication is the work horse of reliable systems because existing, non-fault-tolerant applications can easily be converted to use it and then tolerate (single) crashes. In this paper, the authors analyze the limitations of this approach, particularly the unavoidable latency it introduces.

"Predictable Cloud Computing," by Sape J. Mullender. Processor and network virtualization are the most common tools for building cloud infrastructure. These tools, however, present huge obstacles to making distributed applications truly reliable and to making them perform in real time. This paper presents an alternative to processor and network virtualization that does give real time and independent-failure guarantees, while still allowing legacy applications to run in the cloud.

The second group of papers is about response time and latency in cloud computing and consists of the following papers.

"NIX: A Case for a Manycore System for Cloud Computing," by Francisco J. Ballesteros, Noah Evans, Charles Forsyth, Gorka Guardiola, Jim Mckie, Ron Minnich, and Enrique Soriano. NIX is an operating system designed for multicore computers. It uses a few cores for operating system services and hands over the remaining cores to applications. The applications can use the processors without interruption, thus benefiting high performance and resource-intensive real time applications.

"3D Rendering in the Cloud," by Martin D. Carroll, Ilija Hadžić, and William A. Katsak. An unusual, but important use of cloud computing will be residential computing in the cloud: keyboard, mouse, and screen are at home, but the computer is in the cloud. This paper addresses high-performance graphics processing to remote displays by virtualizing graphical processing units.

"Latency in Cloud-Based Interactive Streaming Content," by Ron Sharp. This companion paper to the previous one analyzes the latency budget for graphical rendering to remote displays. It shows that residential computing in the cloud is a challenging, but feasible option.

The third group concerns telecommunication service in the cloud and consists of these papers:

"Scalable and Elastic Telecommunication Services in the Cloud," by Yuh-Jye Chang, Adiseshu Hari, Pramod Koppol, Antony Martin, and Thanos Stathopoulos. This paper investigates the issues and opportunities for moving standard telecommunication services into the cloud. The differences between, on one hand, web

applications which are already being cloudified everywhere and, on the other hand, telco applications is substantial and the paper analyzes these differences. It then investigates design options and performance limitations and comes up with a viable framework for telco services in the cloud.

"DMME: A Distributed LTE Mobility Management Entity," by Xueli An, Fabio Pianese, Indra Widjaja, and Utku Günay Acer. The mobility management entity is used in LTE cellular networks to keep track of the locations of mobile devices in the network and to manage delivery of incoming calls and handovers when mobiles move from base station to base station. This paper discusses a distributed and fault-tolerant implementation of the MME, capable of being cloudified.

"Mitigating High Latency Outliers for Cloud-Based Telecommunication Services," by Fangzhe Chang, Peter S. Fales, Moritz Steiner, Ramesh Viswanathan, Thomas J. Williams, and Thomas L. Wood. The final paper in this issue reports on a set of experiments to investigate the performance implications of running applications in the cloud. The cloud, after all, provides a shared computing resource in which loads placed by some applications will impact response time for others. Although, on average, performance is quite acceptable, there are outliers—cases in which a response takes a very long time to be produced. This paper concentrates on those outliers and ways to reduce their frequency of their occurrence.

This, in a nutshell, summarizes the contents of this issue of the *Bell Labs Technical Journal*. We hope you'll enjoy reading it.

***Trademarks**

MS-DOS is a registered trademark of Microsoft Corporation.

UNIX is a registered trademark of The Open Group.

(Manuscript approved April 2012)

SAPE J. MULLENDER is director of the Network Systems Lab at Bell Labs in Antwerp, Belgium. He has worked extensively in operating systems, multimedia systems, and, in recent years, wireless systems research. He was a principal designer of the Amoeba distributed system; he led the European Union's Pegasus project, which resulted in the design of the Nemesis multimedia operating system; and he made valuable contributions to work on the Plan 9® and Inferno® operating systems. He received a Ph.D. from the Vrije Universiteit in Amsterdam, The Netherlands, where he also was formerly a faculty member. He currently holds a chair part time in the Computer Science Department at the University of Twente. Dr. Mullender has published papers on file systems, high-performance remote procedure call (RPC) protocols, migratable object location in computer networks, and protection mechanisms, and he was involved in the organization of a series of advanced courses on distributed systems.

THOMAS L. WOOD is a director in Bell Labs' Enabling Computing Technologies research domain and is based in Holmdel, New Jersey. Hired into Bell Labs' Government Communication Center, he has been with the company for over 25 years, and has worked on a variety of projects including large-scale control systems, image processing, and real time media processing. He led a team that created Voice over Internet Protocol (VoIP), IP traffic-shaping technology, and a hardware architecture that was deployed as part of a fiber-to-the-home solution. The technology was adapted and deployed as part of the company's Line Access Gateway product. Mr. Wood also served as a Brookings Congressional Fellow in the office of Senator Bill Frist. He has a B.S.E.E. from Rensselaer Polytechnic Institute in Troy, New York, and an M.S.C.S. from Columbia University in New York City. ◆