Agathe Benichou
Professor Pfaffmann
CS203: Computer Architecture
Due : Friday December 9th 2016

The Future of Machines

There are many technologies which, upon their development and advancement, will shape the future of the world. Revolutionary technologies such as Artificial Intelligence, Virtual Reality, and the Internet of Things will have an immense impact on how ordinary people live their lives. There are also more subtle technologies that won't affect ordinary humans as directly but will make a vast impact on the future of Computers. These technologies, such as cloud based computing, the Google cluster, and possibly open source Instruction Set Architectures, are advancing at a faster rate than ever before. The implications that these technologies will bring are far-reaching to the development of computers and servers.

There are four articles which express four different pieces of technology: *Web Search for a Planet: The Google Cluster Architecture, Profiling a Warehouse Scale Computer, Proprietary vs Open Instruction Set Architectures* and *Unlocking Ordered Parallelism with the Swarm Architecture*. *Web Search for a Planet: The Google Cluster Architectures*, discusses the unique server that Google built to maintain and increase the speed of their Google search queries. *Profiling a Warehouse Scale Computer* explores cloud based and cloud support computing and examines Google data centers to identify where server performance can be improved. *Proprietary vs Open Instruction Set Architectures* examines the debate on whether Instruction Set Architectures should be privately owned and operated or open to the public to edit and expand on. The final article, *Unlocking Ordered Parallelism with the Swarm Architecture* explores the cutting edge parallel architecture that is Swarm which uses ordered parallelism to outperform sequential implementations. How will the technologies discussed in these articles affect the future of Computers?

How will the growth of the Google impact server development and web search development? The Google Cluster discussed in *Web Search for a Planet: The Google Cluster Architecture* is distributed and operated worldwide. In this article, Google describes how their web search application is unique and achieves superior performance at a fraction of the cost compared to other systems. Google architecture combines more than 15,000 commodity class PCs with fault tolerant software to create what is known as the Google Cluster. This massive infrastructure of clusters is distributed worldwide in order to handle Google's constant stream of queries. Search engineers generally require a great deal of computation per request but Google argues that its infrastructure, compared to having fewer, more expensive high-end servers, is optimal for their search engine. Not only does this architecture give them the best price per performance but it also utilizes easy parallelization so that different queries can run on different processors and a single query can use multiple processors. Since these commodity class PCs (and machines in general) are inherently unreliable, the reliability is provided by replicating the query services across many machines within the clusters and automatically detecting and handling machine failures. Since queries are managed by parallelizing individual requests, this eliminates dependencies and decreases server failure. The Google Clusters, which are made by Google themself, focuses on tolerating software failures and uses replication to reinsure availability and reliability. Google believes that their PC-based Cluster architecture is the optimal tradeoff that results from using unreliable commodity PCs while parallelizing the search over many machines to ensure reliability. This server architecture is an interesting approach for dealing with Google's intense web search application. One would think that the most favorable approach to dealing with this massive amount of data would be implementing several high-end and expensive servers. The fact that our everyday computers can work together (of course, with specialized software) to support thousands of queries per second coming from all around the world demonstrates the strong capabilities of server parallelism. The success of the Cluster would drop dramatically if the server didn't replicate the query services across many machines using parallelism. This automated handling of machine failures as well as the parallelism that the query is executing among several machines is an important factor to the development of servers. This

unique concept can be expanded to technologies other than servers to possibly improve the speed and performance of various machines.

How can cloud based and cloud support computing improve in order to advance data centers? The results found in the study examined in *Profiling a Warehouse Scale Computer* identifies where future server performance and cost savings can be improved in data center software and hardware. This article discusses the results from a study profiling the performance of more than 20,000 Google machines over a 3-year period which served the requests of billions of users. Data centers are the platform of choice for modern applications, especially with the growth of cloud based computing. Cloud computing relies on sharing computer resources and uses networks of large servers typically running on low-cost consumer PCs with specialized software, exactly like the Google Clusters, to spread data-processing. However, this has created new challenges for computer architecture that powers large internet servers. The hardware behind the cloud, known as warehouse scale computers or WSCs, emphasizes system design for internet services. WSCs were created to solve computing problems that were too large to fit on a single server. These data centers consist of a large cluster of computers, each of which communicates with other servers through remote procedures. On the data side of these machines, the article reports that 50-60% of all core cycles stalled which is a consistent result. This has created the idea of adopting weaker cores in data centers; since the cores are mostly stalled on memory, there is not a need to waste core speed and energy on them. The analysis of execution cycles suggests a high fraction of stalls are due to instruction caches and these stalls are so severe that they completely drain the cores front-end causing full instruction starvation. The article advocates that “the high incidence of instruction misses and the growing instruction cache working set encourages more emphasis on reducing or protecting instruction cache working sets.” The fast growth of public clouds gives significant interest in looking for longer-term performance optimizations using the suggested approach. Public clouds are currently being used by small and large companies, academics and web development. If the warehouse scale computers within a data center can be altered so that most frequently used components are sped

up, then this might create a permanent shift towards cloud computing. This brings up concerns about cloud computing such as cyberattacks, information theft or government intrusion.

How will the open vs proprietary ISA argument resolve? How will an open source ISA develop and affect the technology world? These questions are discussed in *Proprietary vs Open Instruction Set Architectures* which explores the reasons why Instruction Set Architectures have been proprietary and why the switch to open source Instruction Set Architectures will impact the connection between hardware and software. An ISA is one of the most important interfaces in a computer system because it divides the software from the hardware. Open source software can create competition and creativity so many people believe that having an open source ISA would benefit technology. Dave Christer, who spent two decades contributing to AMD x86s processor argues that there is no need for open software if ISA owners are responsive to the needs of their customers, which entails adding extensions that make sense for the customer base. He believes that commercial ISAs operate as very effective standards, standards provide stability, stability supports strong ecosystems and strong ecosystems enable a large amount of applications. On the other hand, David Patterson, who led the design and implementation of RISC I, argues ISAs have been proprietary for business reasons but there is no real technological reason for the lack of free, open ISAs. Companies with patents on ISAs are quick to sue if you infringe on their patents, even with a licence. This kills competition and stops innovation. He believes that a 21st century ISA should be minimal and modular with a smart chip that is customized to the application. Clearly, the computing industry has been revolutionized by open source software such as Linux, Ubuntu, and Firefox. These softwares are just as widely used as Windows and Apple computers which are proprietary. A successful ISA can speed up the hardware of a computer and overall improve the use of any machine which has interacting hardware and software. However, there are security risks to having the internal structure of an ISA exposed to anyone. If a large company is using an open source software, then it is easy for a hacker to discover how their machines operate which potentially exposes the information of all of that company's customers.

How will the concept of ordered parallelism continue to grow with the Swarm Architecture in order to improve architectures? The development and performance of Swarm Architectures is discussed heavily in *Unlocking Ordered Parallelism in Swarm Architecture* whose goal is to design efficient architectural support for ordered parallelism. This innovative architecture outperforms its sequential implementations. Multicores are spreading widely but still providing limited architectural support for parallelization so it is crucial to explore new architectural mechanisms to efficiently exploit as many types of parallelism as possible. The more parallelism is explored, the more parallel systems become more versatile and easier to program. Many believe that the only way to improve performance is using parallelism. Parallelizing a program includes two main steps: dividing the work into tasks and enforcing synchronization among tasks with potential data dependencies to ensure correct behavior. The two different classes of parallelism are ordered and unordered which both place different demands on the system. In unordered parallel programs, available tasks can execute and complete in any order. In ordered parallelism, tasks must follow some sort of order; even when tasks create new children tasks. Ordered parallelism is more demanding on the system but it is simpler than unordered parallelism. Swarm exploits ordered parallelism by relying on a co-designed execution model. Tasks dynamically create children tasks and schedule them to run at a future time which results in a different task creation and execution orders. Swarm consists of time stamped tasks where each task can access arbitrary data and crete children tasks with any timestamp greater than or equal to its own. Swarm's techniques could be useful in making automatic parallelization practical as well as combined with existing autopar alization techniques to contribute to the speedup of machines. While Swarm is one of the first architecture of its kind, many believe for it to be the future since it will greatly increase the speed in machines.

All of the technologies discussed in the four articles will impact the future of computers and servers, both on the software and hardware side. Mankind is obsessed with making technologies faster, smaller, better. This involves further developing the hardware and software of cloud computing used in architectures such as the Google Cluster which are stored in giant data warehouses. The potential expansion of ISAs that would stem from open source software

could introduce a significant improvement on the performance of cloud computing. While cloud computing increases efficiency, flexibility and has automatic software recovery, there are many dangers associated with it. The cloud is an off-premise system where users outsource their data needs to a third party provider. This entails trusting your data for someone else to look after, they have full control of your data. Technologies such as Google Drive and DropBox are widely used by all types of users (individuals, families and companies) to store volumes of private information on the same cloud system. Anytime data is stored on the Internet, there is a risk for a cyberattack. While these companies have strong encryption on their cloud systems to prevent hackers and security breaches are rare, there is still always the possibility.

In the end, mankind is both scared and intrigued by the unknown. This is especially true in regard to advancements in technology which has affected the world so much in the last couple of decades. The growth and intent of cloud computing is impressive, but there are major security risks which is one of the biggest inhibitors to wider adoption. The rise of Google clusters, the extensive spread of cloud computing, the imminent motion of open source Instruction Set Architecture and the expansion of employing ordered parallelism in Architectures; only time will tell how these technologies will affect the future.

Bibliography

Barroso, L.a., J. Dean, and U. Holzle. "Web Search for a Planet: The Google Cluster Architecture." *IEEE Micro* 23.2 (2003): 22-28. Web.

Beal, Vangie. "What Is Cloud Computing? A Webopedia Definition." *What Is Cloud Computing? A Webopedia Definition*. N.p., n.d. Web. 08 Dec. 2016.

Jeffrey, Mark C., Suvinay Subramanian, Cong Yan, Joel Emer, and Daniel Sanchez. "Unlocking Ordered Parallelism with the Swarm Architecture." *IEEE Micro* 36.3 (2016): 105-17. Web.

Heary, Jamey. "The Dangers of Cloud Computing." *Network World*. Network World, 28 May 2010. Web. 08 Dec. 2016.

Hill, Mark D., Dave Christie, David Patterson, Joshua J. Yi, Derek Chiou, and Resit Sendag. "Proprietary versus Open Instruction Sets." *IEEE Micro* 36.4 (2016): 58-68. Web.

Kanev, Svilen, Juan Pablo Darago, Kim Hazelwood, Parthasarathy Ranganathan, Tipp Moseley, Gu-Yeon Wei, and David Brooks. "Profiling a Warehouse-scale Computer." *ACM SIGARCH Computer Architecture News* 43.3 (2015): 158-69. Web.

"How Technology Is Shaping Our Future And Our Careers." *Www.linkedin.com*. N.p., n.d. Web. 8 Dec. 2016.

"6 Technology Mega-trends Shaping the Future of Society." *Www.weforum.org*. N.p., n.d. Web. 8 Dec. 2016.

"8 Reasons to Fear Cloud Computing." *Business News Daily*. N.p., 01 Oct. 2013. Web. 08 Dec. 2016.

"Top Ten Major Risks Associated With Cloud Storage." *Cloudwards*. N.p., 17 Aug. 2015. Web. 08 Dec. 2016.

"What Are the Real Security Risks of the Cloud?" *Security Intelligence*. N.p., 29 May 2016. Web. 08 Dec. 2016.