the INSTITUTE for
ENERGY EFFICIENCY
UC SANTA BARBARA

THE KAVLI FOUNDATION

Report summarizing the findings of a joint UC Santa Barbara Institute for Energy Efficiency Technology Roundtable / Kavli Futures Symposium on:

Scalable, Energy-Efficient Data Centers and Clouds

February 2012
Santa Barbara, California

Acknowledgements

This report summarizes the results of the joint Institute for Energy Efficiency Technology Roundtable / Kavli Futures Symposium on Scalable, Energy-Efficient Data Centers and Clouds, which was held at UC Santa Barbara on November 30 & December 1, 2011. As co-hosts of the event, the Institute for Energy Efficiency and The Kavli Foundation thank the roundtable sponsors, Intel, AMD and Hewlett Packard, for their generous support. Thanks are also extended to the participants, listed below, for their contributions to the discussion and outputs.

Participants

Name	**Organization**
Rod Alferness	UC Santa Barbara
Dan Blumenthal	UC Santa Barbara
John Bowers	UC Santa Barbara
Fred Chong	UC Santa Barbara
Miyoung Chun	The Kavli Foundation
Jud Cooley	Oracle
John D’Ambrosia	Dell, Ethernet Alliance
Stuart Elby	Verizon
Garth Gibson	Carnegie Mellon University
Jeff Henley	Oracle
Chris Johnson	Google
Krishna Kant	George Mason University, NSF
Dean Klein	Micron Technology
Michael Pinter	Southern California Edison
Partha Ranganathan	Hewlett-Packard
Steve Reinhardt	AMD
Adel Saleh	UC Santa Barbara
Nathan Schrenk	Facebook
Larry Smarr	UC San Diego
Dimitri Stiliadis	Alcatel-Lucent
Luke Theogarajan	UC Santa Barbara
Rich Uhlig	Intel
David Yen	Cisco

Sponsors

Table of Contents

Roundtable chair:
Fred Chong, UC Santa Barbara

Report authors:
Partha Ranganathan, Hewlett Packard
Fred Chong, Hassan Wassel, Martijn Heck and Adel Saleh, UC Santa Barbara

Executive Summary

In an era when massive data will enable unprecedented opportunities in business and science, cloud and data center facilities face significant challenges in the scaling of performance and energy consumption. To face these challenges, UC Santa Barbara's Institute for Energy Efficiency and The Kavli Foundation convened leading experts for a highly interactive, two-day roundtable on November 30 & December 1, 2011 to shape the direction of future technology research.

The overarching problem identified by the roundtable is that of exponentially-growing, massive global data – 1000X growth within the next 13 years. The World Economic Forum recently focused on the opportunities that can be created by exploiting this increasing flood of global data (World Economic Forum, 2012). Unfortunately, the International Technology Roadmap for Semiconductors (Semiconductor Industry Association, 2010) has transistor density doubling only every 3 years after 2013 (a scaling phenomenon commonly known as Moore's Law). Under these assumptions, Moore's Law will provide no more than 25X improved computational efficiency in the same 13 years, leaving at least a 40X gap between compute growth and data growth. The actual gap is likely to be much worse, as transistor energy efficiency has begun to improve slower than density. Without significant improvements in energy efficiency, tomorrow's data centers and clouds will either require substantially more investment and energy, or fail to realize the value of the world's data.

The roundtable identified four key themes that need to be addressed to improve energy efficiency in data centers. These themes arose from an initial context established with technology-focused talks on large-scale applications, computation, communication and storage. The four key themes are:

1. Investing in understanding the newly emerging workload space to ensure proper matching of architecture design to application requirements (Section 2 of this report);
2. Provisioning data center resources for predicted future workloads and applications, as well as improving the energy-proportionality of individual machines to correlate power consumption to load (Section 3);
3. Improving vertical integration within the software stacks of communication, storage and runtime-systems (Section 4);
4. Establishing standards for hardware and software communications to allow the integration of new technologies and component designs (Section 5).

1. Background

Data is growing exponentially and will continue to do so in the years to come. Between 2002 and 2009, data traffic grew by a factor of 56 (M. Mayer, 2009). Between 1998 and 2005, data centers grew in size by 173% per year (Figure 1). This growth in data is leading to a bottleneck both in computation power and in energy use. Computation power is not keeping pace with the growth in data; while data traffic grew by a factor of 56 from 2002 to 2009, computing power grew only by a factor of 16 due to Moore's Law. Figure 1 illustrates clearly the gap between growth in data and computation power.

This rapid growth in data is being fuelled by both human and non-human sources. By March 2011, YouTube users were uploading 48 hours of video every minute, corresponding to around 100 Terabytes of data per day. At the other end of the spectrum, automated sources such as sensor networks regularly summarize and stream information to data repositories for analysis and storage. The sources of data are diverse: they can be structured (census data and government statistics), semi-structured (XML and emails), unstructured (video and audio content), and real time (traffic reports). These varied sources of data are attractive to application designers, who in turn invent and develop new applications every day. Applications will continue to grow and become increasingly diverse in nature, further increasing computation needs and creating more data.

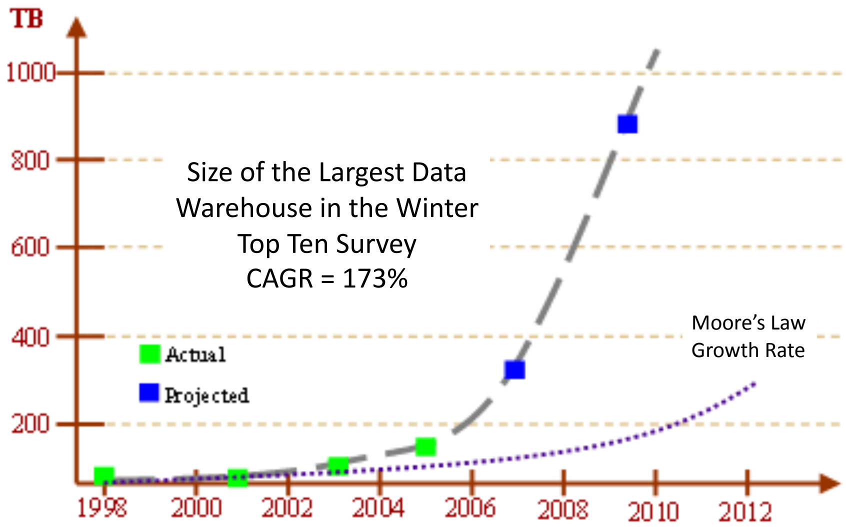

Figure 1: Projection of size of the largest data center reflecting the data growth trends.

Projecting a continued exponential growth in data at 173% per year, it is anticipated that within 13 years data centers will need 1000 times more computation power and, at least theoretically, will consume 1000 times more energy. Technology scaling trends are expected to be able to reduce anticipated energy consumption by no more than a factor of 25 (Semiconductor Industry Association, 2010). The remaining factor of around 40 will require a substantial inter-disciplinary effort to tackle the problem of architecting new data centers from device research, component design and ensemble architecture design.

2. Invest in Understanding the New Emerging Workload Space

Improper matching of application requirements to architecture design can lead to significant energy inefficiencies. A systematic and detailed understanding of current and emerging workloads can go a long way in addressing such inefficiencies. Both scalable clouds and data-centric workloads are considered below.

2.1 Scalable Clouds

Cloud architectures offer the potential for significant energy savings by virtue of being able to apply energy optimizations at a scale compared to smaller system deployments. For example, prior studies have shown significantly improved energy efficiencies for the cooling and power delivery infrastructure (as measured by the PUE metric) for cloud data centers compared to smaller deployments (Hoelzle & Barroso, 2009). However, there are several challenges in validating and unlocking such energy efficiency advantages.

2.1.1 Study and optimize the cloud as a global system

An important challenge is to characterize cloud systems as a global system. Recent trends towards considering infrastructure design for cloud systems as "warehouse-scale" computers (Hoelzle & Barroso, 2009) are a promising first step. However, a more holistic focus is needed that addresses the global tradeoffs with designing a cloud system. For example, beyond the servers and intra-datacenter networking it will be important to consider the long-haul optics and the networking infrastructure associated with the cloud. Such a global focus may lead to new designs, including for example designs where carriers optimize their infrastructure for a larger group and new economics and regulations that are better aligned with an integrated value-chain. A good illustrative example is the work done by Tucker et al. on optimizing the energy consumption of global cloud systems (Baliga, Ayre, Hinton, & Tucker, 2010). A holistic global characterization of the cloud will also lead to several interesting new research directions. These include the questions of what is the appropriate architecture for the global system and what are the new applications for this global system.

2.1.2 Study sociological implications of cloud adoption and corresponding research challenges

While almost all applications show a strong potential to be migrated to the cloud, significant roadblocks exist in terms of privacy and security, especially in the context of a multi-tenancy system. More broadly, there are several sociological issues tied with adoption of the cloud for a broader class of workloads. At the individual level, it is important to understand and address personal reservations in moving private data to the cloud and translate them to research challenges appropriate to the broader community. At the enterprise level, it is important to reconcile corporate policies and regulatory compliance with moving data and applications to the cloud as well as address issues pertaining to managing private and public clouds together.

2.1.3 High-performance computing (HPC) workloads on the cloud

A specific example of increased adoption of cloud workloads that was discussed was "HPC-on-the-cloud". HPC systems have traditionally been harbingers of new architectural ideas and present an important market for migrating to the cloud. Current roadmaps to future exascale systems are primarily directed at using specialized systems like GPUs or FPGAs to achieve significant energy efficiency improvements, with associated challenges around the software model. However, an interesting research opportunity exists around building a cloud exascale computer, with several corresponding open questions. It is unclear yet what the architecture of such a cloud exascale computer would be, or how its energy efficiency performance will compare to traditional cloud environments and to current HPC designs. Another open issue is the required software model that is needed to leverage cloud environments.

In addition, discussions also highlighted other important research opportunities for increasing the adoption of cloud workloads. Notably, unlike the well-trained management workforce for existing data centers, it will be important to understand and optimize the management of future cloud data centers. Bandwidth and latency issues, particularly in the context of multi-tenancy workloads with multiple service-level agreements, will continue to be important. Availability of cloud services with stricter guarantees of uptime will also become more important with broader adoption.

2.2 Data-centric Workloads

While it is clear that data-centric workloads will be an important component of future data centers, many details are unknown about the nature of such emerging data-centric workloads.

2.2.1 Characterize value from data

An interesting trend noted in the group discussions was that as the volume of available data increases, the amount of valuable insight extracted from the data, on a per-byte basis, also correspondingly decreases. This metric, dubbed the "information quality" of data, is also likely to show more heterogeneity across the total amount of data sets of the future compared to the past. This can have significant implications on system design, particularly requiring significantly more efficient ways to extract insights from low-quality large volumes of data. An important first step will be to formalize such metrics and measure them for emerging data-centric workloads.

2.2.2 Characterizing other data-centric behavior

Another similar metric that will be important for the characterization of future data-centric workloads is the frequency of how the data is used. This is called "hot versus cold data". Typically, a piece of data starts off being "hot" and gradually decays into being "cold" over time, i.e. it will be used or addressed less often over time. Consider, for example, a twitter feed that is most used in the first few days of its occurrence and gradually becomes more archival over time. But data can also fluctuate between being hot and cold based on other characteristics. For example, data about flu symptoms can exhibit seasonal variations in frequency of usage. Characterizing the nature of variations in the frequency of how data is accessed and used will be important to understand future memory hierarchy optimizations.

The data center may be unintentionally plagued by data duplication as an inevitable effect of receiving data from millions of sources. Actions taken within the data center to identify and remove duplication carry cost and performance penalties although successful de-duplication will potentially result in cost and performance gains.

There are several other important characteristics of data that could tie into new insights for future data centers. An open question is how various categories of data (e.g. structured, unstructured, real-time or media) contribute to the overall data explosion and how they differ in their system requirements. Another issue is how we can understand metadata and provenance associated with data, potentially generating them automatically. Also a question is how we can understand tradeoffs with multiple versions of the same data. For example, one participant mentioned that they could have up to 8000 popular versions of a single video title due to the combinatorial explosion in the number of formats and number of languages. One option is then to store such multiple versions of a single piece of data. Another option is to perform translation and transcoding on-demand and on the fly, thereby trading off computation for storage. The related issue is then where such computation would be done, i.e. in the cloud, in the edge of the network, or at the access device.

2.2.3 Sensors and instrumentation: an emerging challenging data-centric workload category in HPC

Another specific example that was discussed is the emerging category of HPC workloads around sensors and instrumentation. A decade back, the most challenging data sets in HPC were around one terabyte a day from satellite data from NASA. Currently, applications like gene sequencing, DNA analysis, spectroscopy and tomography easily produce two orders of magnitude more data. More importantly, unlike an internet application where large datasets are primarily cold, in many of these applications large volumes of data need to be accessible at the same time. For example, a genomic application that compares genome sequences across a large sample of individuals will need to operate on large volumes of data over and over again. The challenge is how to do this in a sustainable manner. Another example is global sensing and sensor networks around the planet for applications like climate change. This again leads to a large volume of data, with a large fraction concurrently hot. It can be concluded that such ubiquitous sensor networks and advanced instrumentation will create research opportunities around new architectures, new platforms and new applications, particularly with co-design opportunities across the stack.

2.3 Takeaways

- Study cloud computing as a global system.
- Study the social and economic forces around cloud computing.
- Characterize information quality and life cycle in emerging applications.
- Design architectures for huge data sources to meet the needs of future applications.

3. Resource Provisioning and Energy Proportionality

Resource provisioning and energy proportionality are key factors in data center design. The following discusses their potential impact on energy efficiency and concludes with design considerations.

3.1 Resource Provisioning

Data centers are large computing facilities built to last for 15 to 20 years. These systems are built for applications that have very diverse resource requirements, yet they have to live within the same facility. Some of the applications are storage- and network-bound, such as video streaming applications; others are latency-sensitive and computing-intensive applications such as search operations. The loads that exercise a data center vary across the time of day and are related to real-life events, such as celebrity weddings or deaths. Therefore, resource provisioning is an extremely hard design issue in data center design. However, it is a very important design criterion. Under-provisioning means that resources will be the bottleneck and performance degradation will cause lost customers or missed deadlines. Over-provisioning in system resources on the other hand means lost capital and in some cases lost power consumption. Accurate provisioning is extremely challenging since applications vary in their demands but they must be co-located in one facility to amortize the cost of the data center. In other words, one size does not fit all and volume drives cost. This is the rationale for why a resource ensemble in which machines and applications share a pool of resources might provide a solution to the provisioning problem.

Resource sharing has been a pillar in computer system design ever since the time-sharing machines of the sixties of last century. Sharing of resources allows for dynamic re-allocation between applications without the disruption of service or any physical re-allocation. Storage sharing in the data center is already deployed either in terms of Storage-Area-Network or distributed file systems such as Google-File System (GFS) (Ghemawat, Gobioff, & Leung, 2003). These systems trade performance loss arising from added communication latency and bandwidth requirements for flexibility and reliability. The same approach can be applied to DRAM sharing but to a finer granularity, i.e. sharing within a rack rather than sharing within a data center as proposed in the Disaggregated Memory Project (Lim, Chang, Mudge, Ranganathan, Reinhardt, & Wenisch, 2009). Such finer granularity sharing could be significantly more efficient given emerging optical communications technologies. Optics provides high bandwidth and lower sensitivity to physical distance, enabling resource sharing across larger ensembles.

New trends in silicon technology require that only a very small portion of the chip can be switching simultaneously. This means that even the current multi-core trend cannot scale beyond a certain limit. Consequently this opens the window for a hardware accelerator that is essentially an over-provisioning of functionality that will be used infrequently. It is unclear what type of accelerators will be useful in a data center setting but one can imagine that video encoders and decoders will be useful for a video streaming service. This approach will decrease energy consumption since hardware accelerators will be more energy-efficient than their software counterparts. One might think of this approach as over-provisioning of computational resources.

Provisioning communication resources is done by sharing the networking equipment between end-hosts, which can be dynamically allocated between hosts and applications. In this case, the bisection bandwidth of the network is shared between these applications and can be dynamically allocated with Quality-of-Service guarantees. However, this approach does not solve the problem of accurate provisioning of end-host bandwidth requirements. This can be solved using two methods: virtualization and dynamic data rate. Virtualization enables more than one end-host to share the physical end-host network port. This then enables dynamic resource allocation by allocating more or less VMs per physical end-host. Dynamic data rate capability exists in certain technologies such as InfiniBand, which allows dynamic allocation of network bandwidth according to the bandwidth requirement of the applications.

3.2 Energy Proportionality

Energy proportionality is a key requirement in data center design, for both systems and components. Energy proportionality is significant because most data center machines are exercised at loads of 30%-40%, as shown in Figure 2 (Hoelzle & Barroso, 2009). However, computing systems are currently designed to be most efficient under peak load. Energy proportionality means that power consumption is directly related to the load. An optimized and linear relation between both systems is the goal. There are two primary ways to achieve energy proportionality: turning off components that are not used and consolidating work in fewer components, or designing energy-proportional devices and systems.

Workload consolidation conserves energy by moving work to fewer machines and turning off the unused machines. Such an approach works tends to work well for batched utilization of data centers, but can be poorly suited for online applications. Additionally, parallel applications, where many machines must cooperate to compute results, can be problematic. For example, the index data for web search applications is distributed across all machines. None of them can be turned off in case a query needs data on one of these machines, not even under very low load. This is especially true for latency-sensitive applications. One solution is disaggregating memory from the computing modules and shutting down computing resources in proportion to load. In general, the control theory for such systems exists but the ensemble of many of these controllers and time constants is yet to be explored. As stated above, predicting load is very challenging since it often depends upon unexpected external events.

Device and system architecture research can also help in energy proportionality if the devices exert energy only when they are used. For example, DRAM chips require static refreshing circuit to be running whether data are being either read or written or not. This limits the energy proportionality of DRAM. Novel memory technologies such as FLASH, PCM and memristors do not need to be refreshed and thus are more energy proportional. Most of these technologies, however, have challenges relating to limited write endurance and read disturb. To address some of these challenges, designers may be able to trade data retention for write latency and endurance by designing devices with lower energy barriers to change state.

In communication, dynamic data rate channels and switches allow for more energy-proportional communication. This is harder to achieve in optical communication in which laser power consumption is static power consumption.

3.3 Takeaways

- Resource ensembles can be a solution to resource provisioning under variable load and diverse evolving application requirements.
- Optics can enable more flexible resource ensembles.
- Energy proportionality across system layers and components requires research into system architecture that uses novel technologies to achieve higher proportionality.

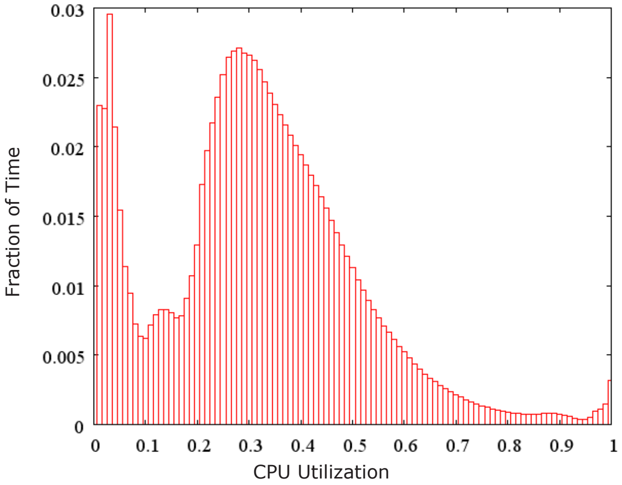

Figure 2: Histogram of CPU utilization in a Google data center

4. Vertical Integration

Trends in technology show that latency is improving, and that the bottleneck is moving to the abstraction layers of the software stack. Primarily, storage and communication technologies are improving, providing new opportunities for taking a fresh look at the software stack of data storage and communication. Abstraction has been a pillar for computer science systems design in enhancing productivity and enabling the interoperability of different systems from different manufacturers. This design principle is re-examined in data center design because latency of hardware components is improving to the extent that software latency is becoming the bottleneck in system performance. Improving performance means direct energy savings because the system spends less time idle and the job is done in less time, leading to lower energy consumption.

4.1 Storage Stack

Storage is traditionally divided into two categories: magnetic-based, large-space, slow storage (high latency and low bandwidth) hard disk drives (HDD) versus relatively faster (low latency and high bandwidth) dynamic memory technologies. Disks are accessed on the block-level (512 bytes) while DRAM is accessed using cache-line granularity (64 bytes). Figure 3 shows the traditional software stack supporting different storage technologies. It shows that main-memory using DRAM is accessed directly by the application. Even with security mechanisms like virtual memory, it is still accessed by the application on cache-line basis.

On the other hand, disks are accessed through File Systems that are based on block-based interfaces suitable for rotating magnetic disks. Storage-class memory technologies are introduced, such as Flash memory, phase-change memory (PCM), spin-torque magnetic memory (STT-Mem) and memristors. These are approaching the latency of DRAM and some of them are word-addressable rather than block-addressable. Table 1 describes projections for these technologies for the next 12 years. This presents an opportunity for system designers to change these interfaces in order to save energy and improve performance.

Flash memory is accessed on a block basis and its success was based on its evolutionary introduction as a secondary storage medium based on the same software stack used for magnetic disks. However, a number of benefits can be gained by re-engineering that software stack. For example, these file systems were optimized for latencies on the order of milliseconds, while Flash memory and other SCM devices can be accessed with latencies on the order of micro-seconds. There are proposals for an evolutionary path that leads to energy and latency savings for data center workloads that use key-value store. This is done by modifying the Flash-based solid-state-disks (SSDs) to provide a key-value store abstraction while that SSD is organized as a cache for a magnetic disk. It is worth noting that disks are not going away anytime soon because the cost per bit is much lower in disks as compared to these abovementioned technologies. The main question is how to design systems with these technologies rather than which one will replace magnetic disks.

12 year goals	DRAM	NAND	PCM	MRAM
Feature (nm)	**8?**	**2x**	**1x**	**1x**
Area (F^2)	**4?**	**1.5x**	**1.5x**	**1.5x**
Read Latency (ns)	**<10**	**1x**	**5x**	**1x**
Write Latency (ns)	**<10**	**100K x**	**5x**	**1x**
Log(cycles)	**16**	**0.3x**	**0.5x**	**1x**

Table 1: Comparison of non-volatile memory technologies vs. DRAM in 12 years

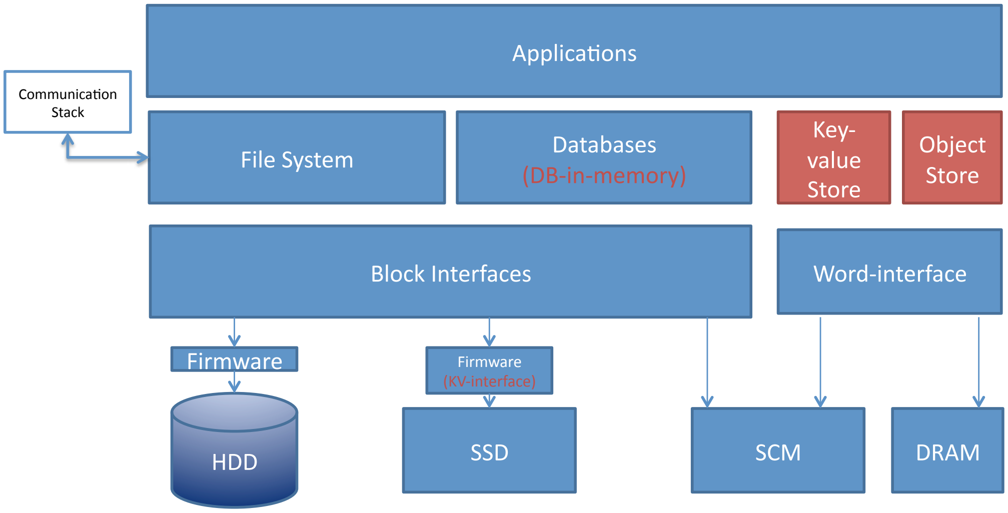

Figure 3: Storage Stack

Other novel SCM memories provide opportunities for either replacing DRAM or cooperating with it in a vertical or horizontal organization because it is word-addressable. Persistent word-addressable memory provides both an opportunity and a challenge to system designers. The opportunity lies in the fact that having persistent memory avoids the static power consumption for refreshing circuitry usually used in DRAM to maintain the data. However, system designers have always assumed the volatility of main memory which enabled system restarts to recover from memory errors. The persistence of memory needs to be expressed on the instruction-set level (ISA) for the operating system designer to be able to make use of it or redesign parts of the system that assumes memory volatility. Moreover, SCM may be used as disk-caches to be accessed using database-in-memory interfaces or file systems. They can be also be used as a complementary memory to place cold data, thereby saving refresh circuitry power consumption of DRAM. All of these proposals must be studied with DC workloads in mind.

4.2 Communication Stack

The latency of network switches within data centers is rapidly improving and is expected to reach 0.3 – 0.4 microseconds by the end of 2012. Figure 4 shows the different layers of the networking stack. It shows the three layers designed for the Internet, namely the Link, Network and Transport layers. These layers were designed with the Internet in mind, where the interoperability of components using different technologies and made by different manufacturers is crucial. The lack of a reliable link layer means that packet drops can be caused by congestion or by unreliable communication channels. This requires a transport layer that exponentially backs off of transmission, resulting in lost bandwidth. However, data center networks have very different characteristics. In data centers, channels are almost lossless, thus flow control protocols should be built on this assumption. DCTCP is a proposal that tries to use early feedback on buffer status to avoid packet drops due to congestion. It tries to solve the incast pattern problem (derived from MapReduce applications), where many sources target one destination synchronously by keeping buffer levels low at all times.

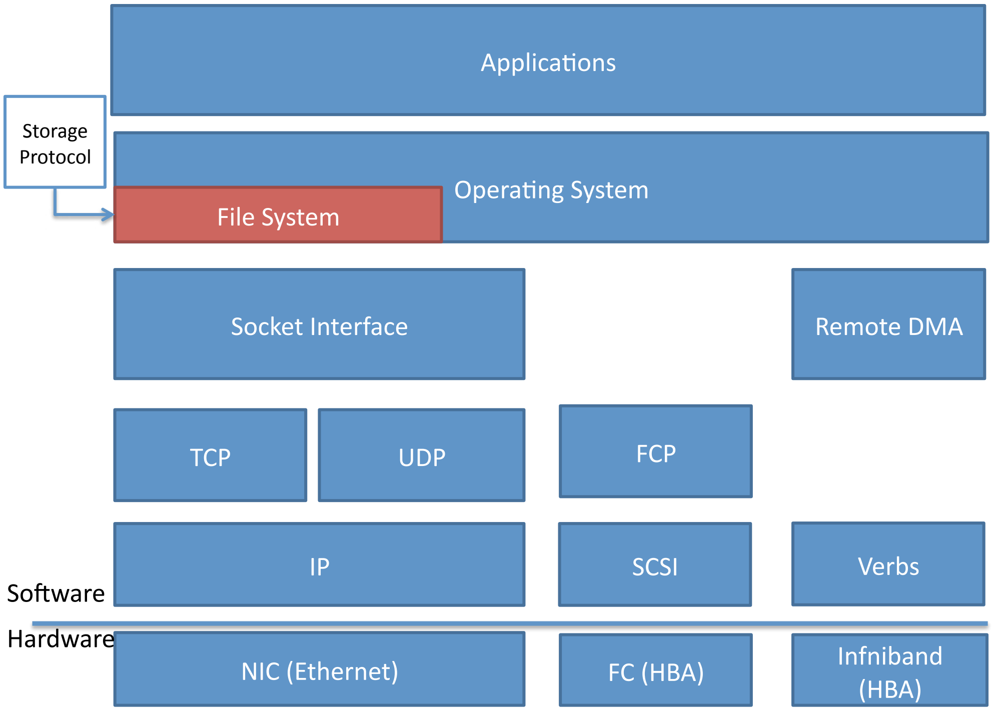

Figure 4: Communication Stack

Low switching latency changes the bottleneck of communication latency from the hardware to the software stack. Further optimizations, such as possibly collapsing some layers into fewer layers, will reduce this latency. Unlike the Internet, all hosts in a data center are under one management, which makes this approach feasible. An extreme approach would be to try to offload the entire software stack into the hardware network interface. However, this approach seems less favorable since all of these protocols are still evolving for the data center environment.

Another approach would be to provide a programming abstraction for the programmer aside from conventional TCP Sockets. Examples include a remote-memory-access interface or remote key-value store interface. These interfaces might be built on TCP/UDP sockets or use IP directly, thus saving latency. These interfaces would generate small-sized packets that would require very low overhead networks.

4.3 Software Stack

Data centers run internet services applications that are usually written in scripting languages such as Ruby, Python and PHP. These languages provide great productivity advantages to the programmers at the expense of lower performance. Balancing these two odds presents a challenge to cloud system providers because latency results in customer stratification and eventually revenue. For example, Facebook developed a PHP-to-C compiler to allow developers to write their code in PHP while the tool generates C code that can be compiled and run without the need of PHP interpreter. This allowed Facebook to run their web services on a lower number of machines and thus save energy.

4.4 Impact on Energy Efficiency

Lower latency can save energy in two main ways. First, lower latency means that it takes fewer cycles to run a certain job, which in turn means that the energy overhead of the saved cycles is reduced. In essence, lower latency means lower energy consumption per job but not necessarily lower power consumption overall. However, in a scale-out data center, executing the same number of jobs on fewer machines (i.e. shutting down extra machines) will save power. A second way to exploit this performance improvement is by accommodating more work and data from the users using the same power budget.

4.5 Takeaways

- New technologies introduced to the storage stack require a fresh look at the software-hardware architecture and interface, exploiting performance improvements to improve energy efficiency.
- Since communication switching latencies are improving within data centers, the communication software stack that was designed for the Internet should be reconsidered.
- Reduced latency will save energy either by running machines for less time or by consolidating more work in fewer machines.

5. Communications for Hardware and Software and Implications

Advances at the architecture and component level are critical to meet future demand. To achieve this, standards must be set for memory, processor, interconnects and photonics so that components from different vendors can be plugged in.

5.1 Processor Scalability

Processor scalability improves both performance and power utilization. By designing for a target application, energy efficiency can be optimized. It is unclear which type of processors will emerge as the standard. In theory, the single-task performance of a single processor is better than that of ten processors that are ten times slower (Amdahl's Law). In practice, the situation is more complex and will depend on workload. For example, multiple smaller and energy-efficient processors are preferred in smartphones. The choice between many small processors and a few large processors will impact the entire server architecture, from memory to interconnect. Again, the trade-off will be application-specific. A better analysis of energy versus performance is needed at the interfaces. However the trend is that multilayer collapse will continue, with the network interface controller disappearing and the processor interfacing directly with the network (e.g. through Ethernet). The network will have intelligent edge processing, whereas the center will take care only of the transport. This leads to lower latency, lower power and more compact networks. Packet processing at the edge will be needed in such a configuration.

5.2 Interconnects

The usually cited figure that communication accounts for 5% of the power consumed by a data center is misleading. This number only covers networking energy, but not communication within the machine between the processor, I/O, and memory. New optical interconnect technologies are currently being considered for implementation in data centers, including Lightpeak as an optical interconnect cable, and the Helios hybrid electrical-optical switch architecture (Farrington, et al., 2010). Lightpeak is a potential architecture changer as a low-cost photonic link, for example to memory. The question of how much of the electrical backplane can be replaced is again dependent on the application, since applications such as internet search, Facebook and YouTube are all radically different both in traffic and processing needs. Currently, a typical 3-stage network has tens of thousands interconnects, switches and transmitter/receiver endpoints. In terms of energy efficiency, the electrical-optical and optical-electrical conversions are bottlenecks. To a large extent, these conversions can be eliminated by transferring to an all-optical switching fabric. More specifically, for a typical 2-stage network the following sequence of ("o" for optical and "e" for electrical) interfaces can be found: "EeoeeoeeoeeoeE". By eliminating the NIC and moving to all-optical switching fabrics, this sequence can in principle be reduced to "EooooE". This is expected to happen within the next 10 years. Energy efficiency is a clear driver for optics in the data center. Also the I/O from processor to memory can in principle be optical. The largest bottleneck for implementation is currently to demonstrate the feasibility first, but reducing the number of boxes will also be key.

5.3 Network Protocol

Another open question is what network protocol will be used: for example, Ethernet or Infiniband. Infiniband is a good example of an energy-proportional network protocol. Energy proportionality is an essential component of data center energy efficiency, as discussed above. In general this is considered to be a (solvable) control theory problem. The most advantageous solution might be application-based switching between network protocols. Again, this means that the optimum design is application-dependent.

5.4 Section Conclusions

Although performance is a key requirement for the customer, energy costs arise as a critical issue, and energy-efficient operation is becoming more important. The optimum communication design for performance and efficiency is application-dependent. Applications drive the required infrastructure, and the infrastructure in turn drives power consumption. However, energy efficiency is hard to measure, since the metrics are not clear. One way is to examine standard or reference applications such as search, key value lookup or business applications, and measure performance. This issue has significant research potential and a set of benchmarks should be developed.

5.5 Takeaways

- Optical communication will play a greater role in improving performance and saving energy within data centers.
- Optics will eventually replace electrical signaling at distances from chip-to-chip and larger.

6. Research Infrastructure

A recurring issue in the roundtable discussion was how best to do scholarly research on data centers and clouds. It became evident that a partnership between academia and industry is needed but it was not clear how best to achieve this partnership, given corporate liability for any leaks of customer data. Sanitized versions of user data or traces may be hard to obtain because the anonymization process is not straightforward. Google recently attempted to deal with this issue by publishing traces of system events from its data center. However, they did not include any network traces. Another possible approach is to create a set of benchmarks or models that capture the essence of the applications. Creating these benchmark suites and models will require interaction between academia and industry since these applications need to represent not only the state of the art, but also future applications and capabilities that are not possible using today's data centers.

These application models and benchmarks, in conjunction with cloud and data center simulators, will enable researchers to explore ideas without the burden of prototyping that might not capture the scale of the cloud. An important point that is that past mistrust in network simulations was due to the lack of representative workloads rather than the lack of accuracy of the simulators. Cloud and data center simulators, however, need to be customizable to accommodate innovation in hardware devices and system architectures. Metrics need to be related to the work, such as query per second, rather than to the system itself, such as aggregate computational power. Additionally, these simulators will have to utilize current cloud systems to be able to simulate any system of sufficient scale in a reasonable time.

A strategic partnership between academia and industry is crucial to achieve to create an open system of platforms. Innovation will flourish using these platforms because it will substantially shorten the evaluation cycle for new ideas. Academia will benefit from system insights that companies have and industry will build on the intrinsic innovative nature of academic research that is not bound by current markets or technologies.

Takeaway

- A partnership between academia and industry is necessary to build a research infrastructure that will expedite innovation in the cloud computing and data center space. Government funding agencies will need to facilitate these efforts.

7. Beyond Energy Efficiency

The roundtable focused on the pressing need for energy efficiency in data centers with an emphasis on maintaining performance to meet business needs. Looking further into the future, sustainable computation may become an increasingly important goal. To increase sustainability beyond energy-efficient computing, Energy Adaptive Computing (EAC) globally manages computations and resources across distributed data centers (Kant, Murugan, & Du, 2012). In particular, EAC attempts to optimize for carbon footprint using locally-available renewable energy sources, taking account of the often intermittent nature of these sources. An even more comprehensive view of designing sustainable data centers can be found using life-cycle analysis, where the carbon footprint of manufacturing, use, and disposal of components can be considered (Chang, Meza, Ranganathan, Shah, Shih, & Bash, 2012).

8. Conclusion

The consensus among the roundtable's industry participants was that cost is their primary consideration and that energy efficiency is simply one component of cost. However, to harness the expected flood of future data, clouds and data centers will need scale in efficiency by an anticipated 40X over the next 13 years, making energy efficiency an imminent issue. Such improved efficiencies will require a better understanding of exascale data processing, a cross-cutting redesign of large-scale systems, and an integration of emerging technologies. All of these efforts will require substantial interdisciplinary effort, industry-academia partnerships, and sustained federal funding.

Bibliography

Baliga, J., Ayre, R., Hinton, K., & Tucker, R. (2010). Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport. *Proceedings of the IEEE* , 149 - 167 .

Chang, J., Meza, J., Ranganathan, P., Shah, A., Shih, R., & Bash, C. (2012). Totally Green: Evaluating and Designing Servers for Lifecycle Environmental Impact. *International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM Press.

Farrington, N., Porter, G., Radhakrishnan, S., Bazzaz, H. H., Subramanya, V., Fainman, Y., et al. (2010). Helios: a hybrid electrical/optical switch architecture for modular data centers. *Proceedings of the ACM SIGCOMM 2010 conference on SIGCOMM*. New Delhi, India: ACM.

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. *Proceedings of the nineteenth ACM symposium on Operating systems principles.* Bolton Landing, NY, USA.

Hoelzle, U., & Barroso, L. A. (2009). *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines.* Morgan and Claypool Publishers.

Kant, K., Murugan, M., & Du, D. (2012). Enhancing Data Center Sustainability Through Energy Adaptive Computing. *ACM Journal of Emerging Technologies in Computing Systems,* (to appear).

Lim, K., Chang, J., Mudge, T., Ranganathan, P., Reinhardt, S., & Wenisch, T. (2009). Disaggregated memory for expansion and sharing in blade servers. *International Symposium on Computer Architecture*. ACM Press.

M. Mayer. (2009, August 13). *The Physics of Data, Talk given at Xerox PARC*. Retrieved from http://www.parc.com/event/936/innovation-at-google.html

Semiconductor Industry Association. (2010). *International Technology Roadmap for Semiconductors.*

World Economic Forum. (2012). *Big Data, Big Impact: New Possibilities for International Development.* Geneva: World Economic Forum.

THE KAVLI FOUNDATION

the INSTITUTE for ENERGY EFFICIENCY

UC SANTA BARBARA

Printed on New Leaf Reincarnation Matte paper, designated Ancient Forest Friendly, containing 100% recycled fiber and 60% post-consumer waste, processed chlorine free, and manufactured with electricity that is offset with Green-e® certified renewable energy certifiates.