Green Clouds: The Next Frontier

Parthasarathy Ranganathan, HP Labs
Partha.ranganathan@hp.com

We are entering an exciting era for systems design. In addition to continued advances in performance, next-generation designs are also addressing important challenges around power, sustainability, manageability, reliability, and scalability. At the same time, a confluence of emerging technologies (such as photonics, non-volatile storage and 3D stacking), and new workloads (around cloud computing, unstructured data, and virtualization) offer additional new opportunities and challenges. The confluence of these trends motivates a rethinking of system design -- motivated by holistic considerations and cross-cutting traditional design boundaries. In this talk, we will examine what such a rethinking means for the basic systems building blocks of the future and their associated management. Focusing on representative examples from recent research, we will discuss the potential for dramatic (10-100X) improvements in efficiency with such future designs, and the challenges and opportunities for future research.

Predicting the future – challenges and opportunities

What can we predict for computing systems 10 years from now?

Historically, the first computer to achieve petascale computing (10^12 or one trillion computing operations per second) was demonstrated in the late nineties. About 10 years later, in mid-2008, the first terascale computer was demonstrated at 1000 times more performance. Extrapolating these trends, one can expect an exascale computer approximately around 2018. That is a staggering one million trillion computing operations per second and a thousand fold improvement in performance than any computer we currently have. Moore's law (often describes as the trend where computing performance doubles every 18 months) has traditionally helped address such performance challenges in the past, to petascale and more recently to terascale computing, but the transition from terascale computing to exascale computing is likely to pose some new challenges that we need to address.

The first challenge is around what is commonly referred to as the "power wall". Power consumption is becoming the key constraint limiting the design of future systems. This problem manifests itself in several ways, in the amount of electricity consumed by systems, in the ability to cost-effectively cool the system, in reliability, etc. For example, recent reports indicate that the electricity costs for powering and cooling cloud datacenters can often be several millions of dollars per year, often exceeding the costs spent on buying the hardware! The industry analyst firm IDC estimated that worldwide investment on power and cooling was close to 40 billion dollars last year. This emphasis on power has started having a visible impact on the design of computing systems with the emphasis shifting from optimizing performance to optimizing energy efficiency – performance achieved per watt of power consumed in the system. This has in part, been responsible for the emergence of multi-core computing as the dominant way to design microprocessors. Additionally, there has been a growing recognition that true energy efficiency optimized designs have to consider the power consumed by the computing system as well as the supporting equipment. For example, in a datacenter, for every watt of power consumed in the server, an additional half to one watt of power is consumed in the equipment responsible for power delivery and cooling (often referred to as the burdened costs of power and cooling).

Looking ahead, going beyond energy efficiency,*sustainability* is emerging to be an important issue. The electricity consumption associated with information technology (IT) equipment is responsible for 2% of the total carbon emissions in the world, more than that of the entire aviation industry. But more importantly, increasingly, IT is being used as the tool of choice to address the remaining 98% carbon emissions from other non IT industries (consider for example., the use of video conferencing to avoid travel, or the use of cloud services to avoid transportation or excess manufacturing). One way to improve sustainability is to consider the total lifecycle of the system – including both the supply and the demand side. In other words, in addition to the amount of energy used in operating a system, it is important to consider the amount of energy used in *making* the system as well.

But sustainability is just one of the new set of "ilities" that pose challenges going into the future. Another key challenge pertains to manageability. Manageability can be defined as the collective processes of deployment, configuration, optimization, and administration during the lifecycle of an IT system. To illustrate this challenge, let us consider, as an example, the potential infrastructure in a future cloud datacenter. Examining recent trends, one can assume 5 global data centers, each datacenter including 40 modular containers, with 10 racks per container, and 4 enclosures per rack, and 16 blades per enclosure. If each blade server had two sockets each with 32 cores, and 10 virtual machines per core, this example cloud vendor will have a total of 81,920,000 virtual servers at their service to operate their services. Each one of these 80+ million servers, in turn, requires several classes of operations – for bring-up, for day-to-day operations, diagnostics, tuning, all the way down to retirement! While there has been a lot of prior work on managing computer systems, manageability at such scale poses new challenges that need to be addressed. Reliability is yet another challenge. Technology scaling trends at a circuits level and the trends towards ever more on-chip integration at the micro-architecture level are expected to lead to higher incidence of error rates (both transient and permanent faults). It will consequently be important to design systems that can operate reliably and provide good uptime even when built out of unreliable components. Finally, it will be important to address these challenges within the constraints of recent business trends. One trend that is particularly important is around the emphasis and use of high-volume components to lower costs.

In this paper, we assert that this combination of challenges – low power, sustainability, manageability, reliability, costs –is likely to influence how we think of system design to achieve the next 1000-fold performance for the next decade. At the same time, there are interesting opportunities that are opening up as well.

From a workload point of view, there has been a fundamental shift in terms of data-centric workloads. The amount of data being created is exploding, growing significantly faster than Moore's law. For example, the size of the largest data warehouse in the Winter Top Ten Survey has been growing at a cumulative annual growth rate of 173%. The amount of online data is estimated to have increased nearly 60-fold in the last seven years. Data from richer sensors, digitization of offline content, and new applications like twitter, search, etc., will only increase data growth rates. Indeed, it is estimated that only 5% of the world's offline data has been made online so far. The emergence and rapid growth of data as a driving force in computing has led to a corresponding growth in data-centric workloads. These workloads focus on different aspects of the data lifecycle – e.g., capture, classify, analyze, maintain, archive – and pose significant challenges for the computing, storage, and networking elements of future systems. Among these, an important recent trend (coupled closely with the growth of large-scale internet web services) has been the emergence of complex analysis at immense scale. Traditional data-centric workloads like web serving and online transaction processing (e-commerce) are being superseded by workloads like real-time multimedia streaming and conversion, history-based recommendation systems, searches of text, images and even videos, and deep analysis of unstructured data (e.g., Google Squared). Compared to traditional enterprise workloads, emerging data-centric workloads have changed a lot of assumptions about system design. These workloads typically operate at larger scale (hundreds of thousands of servers) and on more diverse data (e.g., structured, unstructured, rich media) with I/O intensive, often random data access patterns and limited locality. In addition, these workloads have been characterized by a lot of innovation in the software stack targeted at increased scalability and commodity hardware (e.g., Google MapReduce/BigTable).

Concurrently, recent trends point to several potential technology disruptions in the horizon. On the compute side, recent microprocessors have favored multicore designs emphasizing multiple simpler cores for greater throughput. This is well matched with the large-scale distributed parallelism discussed earlier in data-centric workloads. Operating cores at near-threshold voltage has been shown to significantly improve energy efficiency [36]. Similarly, recent advances in networking, particularly around optics, show a strong growth in bandwidth for communication between different compute elements at various levels of the system design. Significant changes are also expected in the memory/storage industry. Recently, new non-volatile RAM (NVRAM) memory technologies have been demonstrated those significantly improve latency and energy efficiency compared to Flash and Hard Disk. Some of these NV memories, such as phase-change RAM (PCRAM) and Memristors have been demonstrated to have the potential to replace DRAM with competitive performance and better energy efficiency and technology scaling. At the same time, several studies have postulated the potential end of DRAM scaling (or at least a significant slowing down) over the next decade, further increasing the likelihood of DRAM being replaced by these NVRAM memories in future systems.

Inventing the future – cross-disciplinary holistic system design

In this paper, we argue that the confluence of all these trends – the march towards exascale computing and the associated challenges, the opportunities with emerging large-scale distributed data-centric workloads, and the potential disruptions from emerging technology advances – offers a unique opportunity to rethink traditional system design. In particular we believe that this next decade of innovation will be characterized by a holistic emphasis that cross-cuts traditional design boundaries – across different layers of the design, from chips to datacenters, across different fields in computer science, including hardware, systems, and applications, and across different engineering disciplines – computer engineering, mechanical engineering and environmental engineering. We envision that in the future, rather than focusing on the design of single computers, we will focus on the design of computing elements. Specifically, future systems will be composed of (1) simple building blocks that are efficiently co-designed across hardware and software that are (2) composed together into computing ensembles as needed when needed. We refer to these ideas as designing disaggregated dematerialized system elements bound together by a composable ensemble management layer. Below, we present three illustrative examples from our recent research that demonstrate the potential for the dramatic improvements possible with such a rethinking.

Cross-layer power management: The past few years has seen a surge in interest in enterprise power management with several solutions that individually address different aspects of the problem. Going forward in the future, many (or all) of these solutions are likely to be deployed together for better coverage and increased power savings. Currently, the emergent behavior from the collection of individual optimizations may or may not be globally optimal, or even stable, or correct! A key need, therefore, is a carefully designed coordination framework that is flexible and extensible and minimizes the need for global information exchange and central arbitration. In this first example, we discuss how a collaboration effort between computer scientists, thermo-mechanical engineers, and control engineering experts led to a novel coordination solution that addresses this need. Our design is based on carefully connecting and overloading the abstractions in current implementations to allow the individual controllers to learn and react to the effect of other controllers the same way they would respond to changes in workload demand variations. This enables formal mathematical analysis of stability, and provides flexibility to dynamic changes in the controllers and system environments. We demonstrate a specific coordination architecture for five individual solutions using different techniques and actuators to optimize for different goals at different system levels across hardware and software and show that this solution can provide significant advantages to existing state-of-the-art.

Dematerialized datacenters: Our second example discusses a collaboration between computer scientists, environmental engineers, and mechanical engineers to build a sustainability-aware new datacenter solution. Unlike prior studies that focus purely on operational energy consumption as a proxy for sustainability, we use the metric of lifecycle exergy destruction to systematically study the environmental impact of current designs across the entire lifecycle including embedded impact factors related to material use and manufacturing. Based on the insights provided by this study, we propose a new solution co-designed across system architecture and physical organization/packaging, including (1) new material-efficient physical organization, (2) environmentally-efficient cooling infrastructures, and (3) effective design of system architectures to reuse components – all working together to improve sustainability. A detailed evaluation of our proposed solution using a combination of sustainability models, computational fluid-dynamics modeling, and full-system computer architecture simulation, demonstrates significant improvements in sustainability even compared to an aggressive future configuration.

From microprocessors to nanostores: Our last example discusses a disruptive new system architecture for future data-centric workloads. Leveraging the memory-like and disk-like attributes of emerging non-volatile technologies, we propose a new building block for data-centric system design, called nanostores. A nanostore is a single-chip computer that includes 3-D stacked layers of dense silicon non-volatile memory with a layer of compute cores and a network interface. A large number of individual nanostores communicate over a simple interconnect and run a data-parallel execution environment like MapReduce to support large-scale distributed data-centric workloads. The key aspects of our approach are large-scale distributed parallelism and balanced energy-efficient compute in close proximity to the data. Our results show that nanostores can achieve orders of magnitude higher performance at dramatically better energy efficiency and have the potential to enable new data-centric applications that were previously not possible.

While the results from all these examples are promising, we believe we have only scratched the surface of what is possible. There are a lot more opportunities for further optimizations, including for hardware-software co-design (for example, new interfaces and management of persistent data stores) and other radical rethinking of system designs (for example, bio-inspired “brain” computing). Overall, we believe that future is bright and exciting and offers rich opportunities for more innovation by the broader engineering community, particularly around cross-disciplinary research that cuts across traditional design boundaries.