LLNL-TR-474891

Exascale Hardware Architectures Working Group

S. Hemmert, J. Ang, P. Chiang, B. Carnes, D. Doerfler, M. Leininger, S. Dosanjh, P. Fields, K. Koch, J. Laros, J. Noe, T. Quinn, J. Torrellas, J. Vetter, C. Wampler, A. White

March 21, 2011

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Exascale Hardware Architectures Working Group

For the NNSA Workshop: From Petascale to Exascale: R&D Challenges for HPC Simulation Environments

Group Lead: Scott Hemmert
Participants and Contributors: Jim Ang, Brian Carnes, Patrick Chiang, Doug Doerfler, Sudip Dosanjh, Parks Fields, Ken Koch, Jim Laros, Matt Leininger, John Noe, Terri Quinn, Josep Torrellas, Jeff Vetter, Cheryl Wampler, Andy White

Introduction and Scope of Working Group

The ASC Exascale Hardware Architecture working group is challenged to provide input on the following areas impacting the future use and usability of potential exascale computer systems: processor, memory, and interconnect architectures, as well as the power and resilience of these systems. Going forward, there are many challenging issues that will need to be addressed. First, power constraints in processor technologies will lead to steady increases in parallelism within a socket. Additionally, all cores may not be fully independent nor fully general purpose. Second, there is a clear trend toward less balanced machines, in terms of compute capability compared to memory and interconnect performance. In order to mitigate the memory issues, memory technologies will introduce 3D stacking, eventually moving on-socket and likely on-die, providing greatly increased bandwidth but unfortunately also likely providing smaller memory capacity per core. Off-socket memory, possibly in the form of non-volatile memory, will create a complex memory hierarchy. Third, communication energy will dominate the energy required to compute, such that interconnect power and bandwidth will have a significant impact. All of the above changes are driven by the need for greatly increased energy efficiency, as current technology will prove unsuitable for exascale, due to unsustainable power requirements of such a system.

These changes will have the most significant impact on programming models and algorithms, but they will be felt across all layers of the machine. There is clear need to engage all ASC working groups in planning for how to deal with technological changes of this magnitude. The primary function of the Hardware Architecture Working Group is to facilitate codesign with hardware vendors to ensure future exascale platforms are capable of efficiently supporting the ASC applications, which in turn need to meet the mission needs of the NNSA Stockpile Stewardship Program. This issue is relatively immediate, as there is only a small window of opportunity to influence hardware design for 2018 machines. Given the short timeline a firm co-design methodology with vendors is of prime importance.

Technology Challenges

Improving energy and power efficiency is the most formidable challenge facing exascale architectures. Exascale machines must be approximately three orders of magnitude more energy efficient than current machines, with a target of delivering an exa-op at 20 MW. This number implies the delivery of 50 giga operations per watt. In other words, the average energy consumption per operation must reach 20 pico-Joules (pJ). In comparison, Intel's Core Duo mobile processor (circa 2006), consumed more than 10,000 pJ per instruction average. The power issue also contributes to many other challenge areas. Power concerns have limited the increase in clock frequency and led to the multicore era of microprocessors, which is increasing the amount of parallelism in supercomputers at a faster rate than what has historically been seen. Issues with power and energy have also driven, at least in part, a shift to systems that cannot adequately balance compute capability with memory and interconnect performance. In addition to these four areas, increasing part counts is causing growing concerns in the area of resilience. Each of these areas is discussed in more detail below.

Power Large machines spend most of the energy transferring data from or to remote caches, memories, and disks. Minimizing data transport energy, rather than only looking at arithmetic logic unit (ALU) energy, is the critical challenge. Several evolutionary approaches to attaining more energy-efficient architectures exist. At the circuit level, these approaches involve designing circuits for energy and power efficiency, rather than for speed, as in most current approaches. Such designs include on-chip interconnection network circuits for low-swing, and new memory layouts and bank organizations that minimize the amount of capacitance switched per access. Future memory designs must minimize the energy spent charging and discharging lines, possibly through memory designs that include hierarchical bit-line organizations. At the microarchitecture level, evolutionary approaches involve simplifying the cores, making their pipelines shallower and their execution engines less speculative. Finally, at the machine architecture level, a popular approach is to augment the processing nodes with accelerators that are energy-efficient for some operations. Unfortunately, attaining three orders of magnitude higher efficiency in energy and power requires all of this and much more. In particular, it calls for the investment on several research areas that we list next.

Possible near terms areas of investment include: non-silicon memory (see memory section below) and silicon photonics (see interconnect section). Additionally, one of the most effective approaches for energy-efficient operation is to reduce the supply voltage. For example, one possible research area is to set Vdd to a value only slightly higher than the transistor threshold voltage (Vth). This is called Near-Threshold Voltage (NTV) operation, and can lead to a 10x increase in energy efficiency, at the cost of throughput. However, it can cause less reliable operation and research into how to mitigate this would also be needed.

Increasing Parallelism The path to exascale involves increasing the level of parallelism within a node in addition to increasing the number of nodes in a system. The level of each is currently a research question, but to some degree both will be major challenges. At the node level, parallelism will become increasingly hierarchical in the form of a significant increase in the number of computational domains, processing elements per domain, and hardware contexts per processing element. In addition, the width of the computation/vector units will continue to increase. This architectural complexity will be a challenge for the areas of algorithms, programming models, operating systems and compilers. At the system level, the number of nodes in a system will also increase substantially beyond our current platforms. Although this will be a challenge for many areas, resiliency and reliability will have to see significant improvements in order to achieve practical operational performance metrics, such as mean time to interrupt of an application and mean time to failure of the platform.

Memory Memory, not processing capability, is the fundamental challenge to building an exascale system. Projecting the current trends for today's universal memory technology (DRAM), exascale system parameters will push the boundaries of memory capacity, energy efficiency, and performance. Furthermore, disk storage, when considered as another level of the storage hierarchy, may also fail to provide adequate performance, reliability, and energy efficiency, by the end of this decade. As a result, future exascale systems may have a dramatically reduced memory capacity per core, which in turn, increases communication and lowers computational efficiencies.

Recent technological advances in memory technology and system integration, such as Spin-Torque-Transfer RAM (STT-RAM), phase-change RAM (PC-RAM), and Memristors (R-RAM), can address some of these issues by minimizing the energy required to store data persistently when compared to DRAM. Some of these technologies can also increase memory capacity. These improvements can also add new application and software capabilities, such as fault tolerance through node-distributed checkpointing, rather than flushing memory state to rotating disk. In addition, new hardware integration techniques, such as 3D stacking, can provide the opportunity for higher memory bandwidth to processing elements, while maintaining reasonable thermal densities.

Interconnect One critical problem is the growing disparity between the energy required for computation versus the energy required to move communication. At every level -- on opposite sides within a chip, between chips on a motherboard, and connecting racks in a datacenter – the energy dissipated in moving data around is not expected to improve significantly compared with the energy consumed locally in performing computation.

Further research into technological advances, such as 3D integration and silicon photonics, can address interconnect power problem by minimizing the energy required to traverse across the interconnect medium (For example, through-silicon vias result in minimal vertical distances; photonic links exhibit low channel loss for long distances). Energy-efficient transceiver circuits that actuate these novel electrical interfaces then become the dominant consumer in the link budget. Other areas include the reliability/resiliency of these new interconnect interfaces, and architecture co-design of control mechanisms that can optimize the link power for a required bandwidth

Resilience Resilience covers a broad scope ranging from hardware reliability to techniques employed to mitigate hardware and software failures. Traditionally, checkpoints have been used to minimize recalculation due to failure, but even if FIT rates remain constant, the shear number of components in an Exascale platform potentially reduces platform interrupt frequency to minutes rather than hours or days. Addressing resilience for exascale platforms requires a multi-facetted research approach. To achieve an acceptable interrupt frequency, research is necessary to increase the FIT rates for individual hardware components. Research should also be directed at new and novel approaches to application resilience including a move from bulk synchronous processing to models that are insensitive to individual component interrupts. Fault tolerant algorithms, for example, algorithms that tolerate soft errors or can specify portions of the algorithm that must produce correct results should also be investigated. Finally, advances in existing Reliability Availability and Serviceability (RAS) systems are necessary.

Conclusions

The path to exascale brings a host of new challenges, and magnifies many of the existing ones. Peak power and energy usage have become key drivers for all aspects of the system. Energy concerns are driving an accelerating pace in the increase in parallelism as we continue into the multi-/many-core era. System balance is also becoming harder to maintain as the cost of data movement begins to dominate energy usage. Successfully negotiating these challenges will require codesign across all levels of the system. However, because of long hardware design cycles, the opportunity to impact exascale component architectures is short and rapidly developing a codesign strategy with vendors is of utmost importance.