Panel 2

Industrial Perspectives Panel

Panel Moderator: Parthasarathy Ranganathan
Hewlett Packard Labs
partha.ranganathan@hp.com

Abstract
This year, we have a different format for HPCA's industrial session. We will have an "industrial perspectives panel" with practitioners from industry presenting their perspectives on interesting future technical challenges and opportunities for research. We have panelists from IBM, Microsoft, NVIDIA, and Sun exploring a spectrum of interesting areas from consumer to enterprise markets including "hot" topics around multicores, graphics accelerators, and solid-state memory.

Categories & Subject Descriptors: C.4.3 [**Performance of Systems**]: Reliability, availability and serviceability. B.3.2. [**Memory Structures**]: Design Styles. C.1.2. [**Processor Architectures**]: Multiple Data Stream Architectures (Multiprocessors).

General Terms: Management, Performance, Design, Economics, Reliability, Experimentation.

Challenges & Opportunities in Massively Parallel GPU Computing

David Luebke, NVIDIA Research

Modern GPUs (graphics processing units) provide a level of massively parallel computation that was once the preserve of supercomputers like the MasPar and Connection Machine. For example, NVIDIA's Tesla GPU is a heavily multithreaded chip providing up to 240 processing units, 30,720 concurrent in-flight threads, and a trillion FLOPS. The CUDA platform provides a scalable parallel programming model consisting of minimal but expressive changes to the familiar C/C++ language. Using this platform, researchers across science and engineering are accelerating applications in their discipline by up to two orders of magnitude. In this panel, I will motivate GPU computing and explore the transition it represents in massively parallel computing: from the domain of supercomputers to that of commodity "manycore" hardware available to all. I will discuss the goals, implications, and key abstractions of the CUDA programming model. With an installed base of over 100 million units, GPUs represent a tremendous opportunity for innovative, impactful research in parallel programming models and languages, numeric computing, and fault-tolerant computing.

Impact of Technology on Memory Systems of 2015

Moinuddin K Qureshi, IBM Research

Memory systems consisting of DRAM chips are starting to hit cost and power limits. Meanwhile, the realignment of two technologies, Phase Change Memory (PCM) and Flash, in the memory hierarchy promises tremendous growth in capacity, performance, and power efficiency for memory systems during the next decade. In this panel, I will talk about how technology will shape the memory systems of the future. The main-memory system, which has been made of DRAM for the previous three decades, can see significant growth in capacity using a more dense technology, PCM. DRAM will continue to play an important role in future systems by acting as a cache to reduce latency and write traffic to PCM. The faster latency of Flash in comparison to disk makes it a strong candidate for becoming the mainstream technology for File systems. Hard drives will still be used in large storage systems and for backups. Both PCM and Flash come with limited write endurance, which calls for both hardware and software innovations.

MultiCore: Experience, Directions, and Limitations

Rick Hetherington, Sun Microsystems

Sun has been a leader in multicore and multithread processors for some time. This talk will briefly share some of what we have experienced and what we can expect to see going forward. We will also touch on the challenges we face with mulitcore in the coming years.

Clouds and Data Centers

Dan Reed, Microsoft Research

Scientific discovery, business practice and social interactions are moving rapidly from a world of homogeneous and local systems to a world of distributed software, virtual organizations and cloud computing infrastructure. Let's step back and think about the longer-term future. Where is the technology going and what are the implications? What architectures are appropriate for cloud computing? How to we manage power, scale and time to market? What are the right size building blocks? How do we come to grips with the fact that our data centers are now bigger than the Internet was just a few years ago? How do we develop and support malleable software? What is the ecosystem of components in which distributed, data rich applications will operate? How do we optimize performance and reliability?

PPOPP'09, February 14-18, 2009, Raleigh, North Carolina, USA.
ACM 978-1-60558-397-6/09/02.