Execution and Programming Models — Extreme-Scale and Beyond

Stéphane Zuckerman

July 18, 2019

Why Are We Gathered Here?

- Computing systems have undergone a fundamental transformation such that exploiting parallelism has become the only means possible for meeting increasing performance demands (e.g. speed, energy, and efficiency).
- However, to achieve scalable parallelism, we must address the challenges of the increasing performance demand (i.e. speed, energy efficiency, reliability, *etc.*) in this modern era of parallel computing – especially as the expected scale and complexity of future high-performance (HPC) systems is unprecedented.
- We have witnessed the community effort in exploring a **viable path** forward for the design of future HPC systems **over the next 10-15 years,** particularly when one considers that the limits of current semiconductor technology (the post-Moore's Law era).

Panel Organization

- Three questions were asked
- Panelists must answer the first one, and at least one of the remaining two questions
- Each panelist has five minutes to make their statement
  - Time will be strictly kept

Question 1: Program Execution Model (PXM) vs. Programming Model (PM)

What is the main distinction (as well as relation) between the concepts of PXM vs. PMs?
Please state your answer briefly and use your own words and intuition.

Question 2: System-level API and Fine-Grain Parallelism

There is a heated discussion and debate of the following vision:

> *In order to effectively and efficiently exploit the vast parallelism (both at coarse-grain and fine-grain levels) at extreme scale we need to break some traditional abstractions at both PXMs and PMs levels. This is essential in the design of a systems-level API for future extreme-scale parallel computing systems.*

What is your opinion on the above vision?

Question 3: Of the programmability of dataflow models

There has been significant concerns that

> *The dataflow/codelet community has always claimed that their model is more productive; however more recent work with task parallelism and the recent OCR project tried working with these types of models, and the scientific application community actually found them less productive.*

What is your observation/opinion?

Our Panelists

- Erik Altman, IBM
- Jean-Luc Gaudiot, University of California, Irvine
- Hironori Kasahara, Waseda University
- CJ Newburn, Nvidia
- Karthikeyan “Karu” Sankaralingam, University of Wisconsin

Erik Altman, IBM I

Erik Altman joined IBM Research in 1995. In 2014, he began an assignment with IBM's Corporate Technology Evaluation team, which coordinates studies reporting monthly to the CEO and other senior leaders in IBM about important technology directions and their impact on IBM business. During his time in IBM Research, Altman's research focused on binary translation and optimization, compilers, architectures, and micro-architectures. He has authored or co-authored more than 30 conference and journal papers, and has 25 patents and pending patent applications. He was one of the originators of IBM's DAISY binary translation project, that allowed VLIW architectures to have high performance and achieve 100% binary compatibility with PowerPC. He was also one of the original architects of the Cell processor chip that is to appear in the forthcoming Sony Playstation 3 game consoles. He has been the program chair and general chair of the PACT and P=ac2 Conferences and has served on numerous program committees. He has served as guest editor of IEEE Computer, the ACM Journal of

Erik Altman, IBM II

Instruction Level Parallelism (JILP), and the IBM Journal of Research and Development. He is past Chair of ACM SIGMICRO and is currently Chair of the ACM SIG Governing Board and member of ACM Council and Executive Council. He also currently serves as Editor-in-Chief of IEEE Micro.

Jean-Luc Gaudiot, UC Irvine I

Professor Jean-Luc Gaudiot received the Diplme d'Ingnieur from the cole Suprieure d'Ingnieurs en Electronique et Electrotechnique, Paris, France in 1976 and the M.S. and Ph.D. degrees in Computer Science from the University of California, Los Angeles in 1977 and 1982, respectively.

He is currently a Professor in the Electrical Engineering and Computer Science Department at the University of California, Irvine. He was Chair of the Department from 2003 to 2009. During his tenure, the department underwent significant changes. These include the hiring of twelve new faculty members (three senior professors) and the remarkable rise in the US News and World Report rankings of the Computer Engineering program from 42 to 28 (46 to 36 for the Electrical Engineering program).

Prior to joining UCI in January 2002, he was a Professor of Electrical Engineering at the University of Southern California since 1982, where he served as Director of the Computer Engineering Division for three years. He has also designed distributed

Jean-Luc Gaudiot, UC Irvine II

microprocessor systems at Teledyne Controls, Santa Monica, California (1979-1980) and performed research in innovative architectures at the TRW Technology Research Center, El Segundo, California (1980-1982). He frequently acts as consultant to companies that design high-performance computer architectures, and has served as an expert witness in patent infringement and product liability cases. His research interests include multithreaded architectures, fault-tolerant multiprocessors, and implementation of reconfigurable architectures. He has published over 200 journal and conference papers. His research has been sponsored by NSF, DoE, and DARPA, as well as a number of industrial organizations. From 2006 to 2009, he was the first Editor-in-Chief of the IEEE Computer Architecture Letters, a new publication of the IEEE Computer Society, which he helped found to the end of facilitating short, fast turnaround of fundamental ideas in the Computer Architecture domain. From 1999 to 2002, he was the Editor-in-Chief of the IEEE Transactions on Computers. In June

Jean-Luc Gaudiot, UC Irvine III

2001, he was elected chair of the IEEE Technical Committee on Computer Architecture, and re-elected in June 2003 for a second two-year term. In 2009, he was elected to the Board of Governors of the IEEE Computer Society for a 3-year-term. He was the Chair of the IEEE Computer Society Publications Board Transactions Operations Committee (2010-2011), the Chair of the IEEE Computer Society Publications Board Magazines Operations Committee in 2012, the IEEE Computer Society vice President, Educational Activities Board in 2013, and 2014-2015 IEEE Computer Society vice President, Publications Board. He is now the 2017 IEEE Computer Society President.

Dr. Gaudiot is a member of AAAS, ACM, and IEEE. He has also chaired the IFIP Working Group 10.3 (Concurrent Systems). He was co-General Chairman of the 1992 International Symposium on Computer Architecture, Program Committee Chairman of the 1993 IFIP Working Conference on Architectures and Compilation Techniques for Fine and Medium Grain Parallelism, the 1993 IEEE

Jean-Luc Gaudiot, UC Irvine IV

Symposium on Parallel and Distributed Processing (Systems Track), the 1995 Parallel Architectures and Compilation Techniques Conference (PACT 95), the High Performance Computer Architecture conference in 1999 (HPCA-5), and the 2005 International Parallel and Distributed Processing Symposium. In 1999, he became a Fellow of the IEEE, “For Contributions to the Programmability and Reliability of Dataflow Architectures. He was elevated to the rank of AAAS Fellow in 2007, For Distinguished Contributions to the Design and Analysis of Highly Efficient Multiprocessor and Memory System Architectures.”

Hironori Kasahara, Waseda University I

Hironori Kasahara is a professor in the Department of Computer Science and Engineering at Waseda University. He is an IEEE Fellow, an IPSJ Fellow, a Golden Core Member of the IEEE Computer Society, a professional member of the IEEE Eta Kappa Nu, and a member of the Engineering Academy of Japan and the Science Council of Japan. He received a PhD in 1985 from Waseda University, joined its faculty in 1986, and has been a professor of computer science since 1997 and a director of the Advanced Multicore Research Institute since 2004. He was a visiting scholar at the University of California, Berkeley, and the University of Illinois at UrbanaChampaigns Center for Supercomputing R&D. He has served as a chair or member of 250 society and government committees, including a member of the CS Board of Governors and Executive Committee; chair of CS Planing Committee, Constitution & Bylaws Committee, Multicore STC, and CS Japan chapter; associate editor of IEEE Transactions on Computers; vice PC chair of the 1996 ENIAC 50th Anniversary International

Hironori Kasahara, Waseda University II

Conference on Supercomputing; general chair of LCPC; PC member of SC, PACT, and ASPLOS; board member of IEEE Tokyo section; and member of the Earth Simulator and K supercomputer committees. Kasahara received the CS Golden Core Member Award, IFAC World Congress Young Author Prize, Sakai Special Research Award, and the Japanese Ministers Science and Technology Prize. He led Japanese national projects on parallelizing compilers and embedded multicores, and has presented 215 papers, 155 invited talks, and 30 patents. His research has appeared in 557 newspaper and web articles.

CJ Newburn, Nvidia

hris J. Newburn (CJ) is the Principal Architect in NVIDIA Compute Software for HPC strategy and the SW product roadmap, with a special focus on systems and programming models for scale. He has contributed to a combination of hardware and software technologies over the last twenty years and has over 80 patents. He is a community builder with a passion for extending the core capabilities of hardware and software platforms from HPC into AI, data science and visualization. He wrote a binary-optimizing, multi-grained, parallelizing compiler as part of his Ph.D. at Carnegie Mellon University. Before grad school, in the 80s, he did stints at a couple of start-ups, working on a voice recognizer and a VLIW supercomputer. He's delighted to have worked on volume products that his Mom used.