The Future of Hardware Development - A Perspective from Systems Researchers

Yiying Zhang

University of California, San Diego

ACM Reference Format:

Yiying Zhang. 2021. The Future of Hardware Development - A Perspective from Systems Researchers. In *Workshop on Hot Topics in Operating Systems (HotOS '21)*, May 31–June 2, 2021, Ann Arbor, MI, USA. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3458336.3465298>

At HotOS 2011, Mogul *et al.* published a paper calling for “Reconnecting Architecture and OS Research” [9]. Ten years have passed. We now live in a post-x86 world where servers are no longer homogeneous. Accelerators like GPU and FPGA are dominating important workloads like machine learning and search in data centers [3, 4]. Many data centers have launched large scales of customized ASICs (*e.g.*, Google TPU [8], AWS Nitro [1]). At the same time, HBM and NVM are promised to change the landscape of memory and storage; various programmable networking devices like SmartNIC and programmable switches are making their ways into data centers. These exciting new hardware trends have driven systems researchers to innovate on software to fit new hardware technologies. What about the hardware?

Let us revisit the question raised in Mogul *et al.*’s HotOS’11 paper: “*should systems researchers take whatever computer architects have dished out?*” Driven both by the need and by recent advance in hardware programmability, significantly more systems researchers started to change or build their own hardware in recent years. In the last five years of SOSP/OSDI¹, there have been about 10% submissions with the topic “computer architecture”, with many being accepted (Figure 1). We believe that now is the time for systems researchers to gather and discuss the future directions of hardware development, from a systems perspective.

¹There was no topic related to hardware in SOSP’15, and we do not have access to statistics in prior years.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

HotOS ’21, May 31–June 2, 2021, Ann Arbor, MI, USA

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8438-4/21/05.

<https://doi.org/10.1145/3458336.3465298>

Figure 1: Papers with the “Computer Architecture” Topic in the Past Five Years.

Potential Topics

• **When should systems researchers investigate hardware design?** There could be many reasons to move from software to hardware, for example, to accelerate application or OS functionalities [2], to save CPU energy cost, to connect hardware units in non-traditional ways [6, 10], and to enhance security [11]. At the same time, not everything fits a hardware design (*e.g.*, complex data structures and/or rare operations) and hardware development is much more time consuming than software development, especially for software people. How can we make the right decision of what to move to hardware and what to leave in software? How can we estimate the potential benefits of a hardware solution without building the actual hardware?

• **How should systems researchers build hardware?** There are many ways to “build hardware”, ranging from simulation and emulation to FPGA and ASIC prototyping. How should systems researchers choose between the trade-off of development efforts and realistic implementation? What hardware platforms are affordable (in terms of monetary cost and development cost)? What programming languages and toolsets should systems researchers use? How to co-design software with hardware? How to scale out a hardware platform and evaluate its distributed setting?

• **What is the role of open-source hardware?** New open hardware standards like RISC-V [7] and p4 [5] are growing rapidly. Backed by both commercial and open-source implementations, they have fueled many recent research works and real products. Open-source hardware works only if there is the corresponding open-source software. Going forward, how can we build an open-source ecosystem with both hardware and software? What can we learn from the success of open-source software? Can open-source hardware be as successful?

• **How to connect different communities and academia and industry?** Now that systems researchers have investigated more hardware designs than 10 years ago, how should the systems community and the architecture community work together going forward? What about other related communities like networking and programming language? More practically, where should we publish hardware-related papers and how to write them in an approachable way that could be appreciated by software people? Finally, is it possible for hardware-based research to have real impact? What is a viable way to transfer academia hardware research into real products?

Potential Format

We plan to organize the panel with three parts. First, we will have a panel of academia and industry researchers discussing the past, current, and future of systems-hardware research (~45 minutes). It will be followed by an Ask-Me-Anything session where audience can ask any questions to the panelists (~20 minutes). Finally, we will organize breakout discussion groups and/or student panels (~25 minutes). Our current panelists include Aditya Akella, Theo Benson, Christos Kozyrakis, Timothy Roscoe, Chris Rossbach, Steven Swanson, and Michael Swift. We will continue reaching out to a broader community (e.g., industry, students).

REFERENCES

- [1] Amazon Web Services, Inc. AWS Nitro System. <https://aws.amazon.com/ec2/nitro/>.
- [2] Marco Spaziani Brunella, Giacomo Belocchi, Marco Bonola, Salvatore Pontarelli, Giuseppe Siracusano, Giuseppe Bianchi, Aniello Cammarano, Alessandro Palumbo, Luca Petrucci, and Roberto Bifulco. hXDP: Efficient Software Packet Processing on FPGA NICs. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI '20)*, November 2020.
- [3] A. M. Caulfield, E. S. Chung, A. Putnam, H. Angepat, J. Fowers, M. Haselman, S. Heil, M. Humphrey, P. Kaur, J. Kim, D. Lo, T. Massengill, K. Ovtcharov, M. Papamichael, L. Woods, S. Lanka, D. Chiou, and D. Burger. A cloud-scale acceleration architecture. In *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO '16)*, Taipei, Taiwan, December 2016.
- [4] Eric Chung, Jeremy Fowers, Kalin Ovtcharov, Michael Papamichael, Adrian Caulfield, Todd Massengill, Ming Liu, Daniel Lo, Shlomi Alkalay, Michael Haselman, Maleen Abeydeera, Logan Adams, Hari Angepat, Christian Boehn, Derek Chiou, Oren Firestein, Alessandro Forin, Kang Su Gatlin, Mahdi Ghandi, Stephen Heil, Kyle Holohan, Ahmad El Hussein, Tamas Juhasz, Kara Kagi, Ratna K. Kovvuri, Sitaram Lanka, Friedel van Megen, Dima Mukhortov, Prerak Patel, Brandon Perez, Amanda Rapsang, Steven Reinhardt, Bitu Rouhani, Adam Sapek, Raja Seera, Sangeetha Shekar, Balaji Sridharan, Gabriel Weisz, Lisa Woods, Phillip Yi Xiao, Dan Zhang, Ritchie Zhao, and Doug Burger. Serving DNNs in Real Time at Datacenter Scale with Project Brainwave. *IEEE Micro*, 38(2):8–20, 2018.
- [5] P4 Language Consortium. P4. <https://p4.org/>.
- [6] Stephen Ibanez, Alex Mallery, Serhat Arslan, Theo Jepsen, Muhammad Shabbaz, Nick McKeown, and Changhoon Kim. The nanoPU: Redesigning the CPU-Network Interface to Minimize RPC Tail Latency, 2020. arXiv:2010.12114.
- [7] RISC-V International. RISC-V. <https://riscv.org/>.
- [8] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snellman, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon. In-datacenter performance analysis of a tensor processing unit. In *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, 2017.
- [9] Mogul, Jeffrey C. and Baumann, Andrew and Roscoe, Timothy and Soares, Livio. Mind the Gap: Reconnecting Architecture and OS Research. In *Proceedings of the 13th USENIX Conference on Hot Topics in Operating Systems (HotOS '11)*, Napa, California, May 2011.
- [10] Yizhou Shan, Yutong Huang, Yilun Chen, and Yiying Zhang. Legoo: A disseminated, distributed OS for hardware resource disaggregation. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI '18)*, Carlsbad, CA, October 2018.
- [11] Stavros Volos, Kapil Vaswani, and Rodrigo Bruno. Graviton: Trusted Execution Environments on GPUs. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI '18)*, Carlsbad, CA, October 2018.