RETROSPECTIVE: ASIC Clouds: Specializing the Datacenter

Michael B Taylor
University of Washington

I. Introduction

This paper, published in 2016, made the case for why large hyperscaler companies should design ASIC Clouds—data centers full of specialized ASIC-based servers—in order to optimize total cost of ownership (TCO) for scale-out applications. Each system would contain multiple ASICs that were optimized for a particular application, reducing the cost of hardware and the energy consumption; each ASIC would be filled with on-chip network connected accelerators as well as off-chip connections to DRAM, external I/O and other ASICs.

Intellectually, the paper built on prior work by the authors that established the role of specialization to combat the end of Dennard Scaling [O1][O2], and prior work that examined the progression of Bitcoin mining, from CPUs to GPUs to FPGA and finally to datacenters full of SHA-256 ASICs [O18].

The paper grounded itself by building detailed calibrated models for Bitcoin mining server design, and then extended these models for two more interesting cases: machine learning (ML), and video transcoding. In our quest for predicting what workloads would justify design of ASIC clouds, we had settled on those two as being promising candidates with potential for exponential growth and promising economics for hyperscalers. Prior to our paper submission, most academics did not think that datacenters full of ASICs made economic sense.

Our prediction of hyperscalers building both ML and video-transcoding ASIC clouds proved incredibly prescient. Our paper (which we developed during the 2013-2015 time period) predated Google's announcement of the TPU [1], and also the Google ISCA 2017 TPU paper [2]. Post-TPU, Google ramped up development of a video transcoding ASIC cloud, another salient prediction of the paper, and as a result of our paper, several authors of this paper were invited to join that effort. The Google VPU is detailed in ISCA 2021 [3]. Follow-on multi-chip TPU systems were examined by Google in CACM 2020 [4]. Perhaps one of our greatest surprises versus our early paper was that Google was using the TPU not just for scale-out workloads like websearch and image recognition, but for scale-up training of neural networks to drive their ML R&D.

Our follow-on work looked at the impact of non-recurring engineering cost (NRE) on optimizing ASIC cloud server design for TCO [5]. This paper was also selected for IEEE Micro Top Picks for which a higher-level version of the article was written [6], and also for Communications of the ACM Research highlights [7], which offered a more detailed retrospective.

References

[1] I. Magaki, M. Khazraee, L. Vega, and M. Taylor, "ASIC Clouds: Specializing the Datacenter," in *UCSD CSE Technical Report*, May 8, 2016. [Online]. Available: https://escholarship.org/uc/item/4bf9f938

[2] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-l. 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. Snelham, 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," *ISCA*, 2017.

[3] P. Ranganathan, D. Stodolsky, J. Calow, J. Dorfman, M. Guevara, C. W. Smullen IV, A. Kuusela, R. Balasubramanian, S. Bhatia, P. Chauhan, A. Cheung, I. S. Chong, N. Dasharathi, J. Feng, B. Fosco, S. Foss, B. Gelb, S. J. Gwin, Y. Hase, D.-k. He, C. R. Ho, R. W. Huffman Jr., E. Indupalli, I. Jayaram, P. Kongetira, C. M. Kyaw, A. Laursen, Y. Li, F. Lou, K. A. Lucke, J. Maaninen, R. Macias, M. Mahony, D. A. Munday, S. Muroor, N. Penukonda, E. Perkins-Argueta, D. Persaud, A. Ramirez, V.-M. Rautio, Y. Ripley, A. Salek, S. Sekar, S. N. Sokolov, R. Springer, D. Stark, M. Tan, M. S. Wachsler, A. C. Walton, D. A. Wickeraad, A. Wijaya, and H. K. Wu, "Warehouse-scale video acceleration: Co-design and deployment in the wild," in *ASPLOS*, 2021.

[4] N. P. Jouppi, D. H. Yoon, G. Kurian, S. Li, N. Patil, J. Laudon, C. Young, and D. Patterson, "A domain-specific supercomputer for training deep neural networks," *Communications of the ACM*, p. 67–78, jun 2020.

[5] M. Khazraee, L. Zhang, L. Vega, and M. Taylor, "Moonwalk: NRE Optimization in ASIC Clouds or, accelerators will use old silicon," in *Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2017.

[6] M. Khazraee, L. Vega, I. Magaki, and M. Taylor, "Specializing a Planet's Computation: ASIC Clouds," *IEEE Micro*, May 2017.

[7] M. B. Taylor, L. Vega, M. Khazraee, I. Magaki, S. Davidson, and D. Richmond, "ASIC clouds: Specializing the datacenter for planet-scale applications," *CACM*, pp. 103–109, 2020.