Retrospective: A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services

Andrew Putnam Adrian M. Caulfield Eric S. Chung
Logan Adams Kypros Constantinides John Demme Daniel Firestone Stephen Heil
Matt Humphrey Daniel Lo Todd Massengill Michael Papamichael Yi Yuan
Sitaram Lanka Steve Reinhardt Derek Chiou Doug Burger

Microsoft

I. Historical Context

Hardware specialization can improve performance and energy efficiency by several orders of magnitude over conventional CPUs [7]. However, the wide variety of cloud applications, their rapid rate of change, and the need to support multiple geographical regions and generations simultaneously make scalable custom hardware for the cloud challenging.

The Catapult program began in 2010, one year after the launch of Microsoft Bing and two years after the launch of Red Dog, the predecessor to Azure. Bing approached Microsoft Research (MSR) to find ways to give Bing a competitive edge. Many at Microsoft and in the architecture research community envisioned 1000 core "manycore" multiprocessors as the future for application acceleration, but our team looked to hardware specialization as a better approach for Bing. The short timeline of the request (as soon as possible) combined with the low budget quickly ruled out custom hardware solutions. We looked at both GPUs and FPGAs and decided that FPGAs could cover a broader range of workloads, and the SIMD execution model of GPUs did not match the latency-sensitive Bing workload which made batching requests impractical. Accordingly, MSR developed an accelerator platform based on FPGAs, with the vision of eventually creating a fully-custom solution.

Under the codename Project Catapult, the 2014 paper focused on the architecture and deployment of FPGA hardware in Microsoft's cloud and the hardware/software co-design that doubled Bing's ranking throughput and reduced latency by 30% at true production scale.

It is worth noting that the 1,632 servers in the paper was not a random number. This was the smallest number number of servers that could run one Bing instance. With workload performance measured in 99%+ tail latencies and heavily dependent on IO performance, engineering and deploying scale systems is the only way to do realistic evaluations.

Supporting specialized hardware at scale was more challenging than initially envisioned. We went through three design iterations. First, we designed a "mega-board" with six large FPGAs, four of which were placed in a special server per rack. However, datacenters prefer homogeneous racks to simplify power/cooling and to limit the blast radius of failures. In addition, network designs that concentrate traffic at one node result in severe network congestion and bad tail latencies, and tail latencies are a more important metric for cloud workloads like Bing than average performance.

Next, we designed the torus network described in the 2014 paper, whichenabled homogeneous racks, reduced the blast radius of hardware failures, and alleviated network congestion. However, commercial constraints changed just as quickly. While our initial efforts had focused on Bing, Azure was growing at 2x each year. At the time we wrote the 2014 paper, we knew that we would have to shift to the bump-in-the-wire topology introduced later in [1] to satisfy both Azure and Bing.

Since 2015 Microsoft has deployed specialized FPGA hardware in nearly every cloud server. Catapult is the largest public deployment of FPGAs in a reconfigurable computing role. Many millions have been deployed across 6 continents and 60+ regions. The FPGA has been used for compute offload, networking and storage acceleration, AI offload, and has evolved multiple times in each use case in a way that could only be done in reprogrammable silicon. This paper demonstrated that reconfigurable computing could have high value at scale in datacenters and the cloud, which previously had not been shown, and that subsequently the two main FPGA companies (Altera and Xilinx), which had been independent for decades, were acquired by the two largest datacenter CPU vendors (Intel and AMD respectively).

II. Applications Scope

The class of problems that benefit from hardware acceleration is vast, as are the kinds of architectures that could address certain applications. We have found FPGAs to be relevant to a wider variety of domains and workload classes than any other commercial cloud accelerator architecture, though they have shortcomings in each space, especially compared to a dedicated accelerator for only one application space.

For example, Catapult began as a general-purpose compute accelerator platform well before the rush of architectural specialization for AI, yet it still provided a competitive platform for early AI workloads comprised of boosted decision tree forests and DNNs like early CNNs [5] and LSTMs, and BERT-style transformer models, though not LLMs. It also worked very well as a platform for SmartNICs – specialized accelerators for networking, storage, and security.

Development in one application space benefits efforts in other spaces. Infrastructure applications provide scale to keep costs down and robustness to keep platforms highly-reliable.

To properly scope the impact of this work, we define three inter-related but distinct forks:

- *(1) Infrastructure & Data-Movement Acceleration*
- *(2) Application-specific Program Acceleration*
- *(3) AI Acceleration*

A. Infrastructure & Data-Movement Acceleration

Azure SmartNIC [2] is largest use of FPGAs in Microsoft's datacenters. The FPGA is used to offload the software-defined networking (SDN) stack, which virtualizes network traffic from virtual machines. SDN functionality includes checking access control lists, load balancing, network address translation, virtual network support, metering, etc.

Since the initial deployment, the SDN Role (GFT) has gone through three major rewrites, improving key performance metrics like packets-per-second by over 4x, adding support for encrypted VNETs and containers, enhanced packet capture and filtering, and simultaneously *reducing* overall resource utilization. In addition, storage offload capabilities in the form of NVMe and data processing accelerators were recently added, resulting in a 2.5x performance improvement in IOPS and throughput [6] for best-in-class performance in the 100Gb network generation.

B. Application-specific Acceleration

The Bing algorithm described in the original paper is a mix of application-specific program acceleration in the form of Feature Extraction (FE) and Free-Form Expression processing (FFE), which then fed into an AI accelerator, the Decision Tree Scorer (DTS). Between Pilot and Production, we had switched from a dedicated torus network to converged Ethernet and had cut all but the FE functionality.

There are numerous general-purpose use cases being investigated within Azure, but currentlys none of them are public.

C. AI Acceleration

While technically a subset of the above category, AI accelerators have been the most visible category of cloud accelerator architectures. Shortly after Bing went to production with FPGAs, they shifted from boosted decision trees to integrating DNNs into the ranking pipeline. While numerous DNNs were trained and evaluated to significantly improve search, most proposed DNNs could not deploy due to the lack of available latency and CPU budget for each query. The existing fleet of deployed FPGAs and the hardware microservices platform leveraging our custom FPGA-resident LTL (Lightweight Transport Layer) transport protocol provided a solution. The Brainwave framework [3] was deployed side-by-side with FE to great effect. The network allowed every CPU to remotely leverage multiple distinct accelerators for each query, using a shared pool of FPGAs. Brainwave's latency at the required batch size of one was significantly better than other available hardware, including the latest CPUs and GPUs.

Recently, the size, compute requirements, and ubiquity of the largest models make using FPGAs for AI less competitive. Workload stability and market size have allowed GPUs to flourish. While FPGAs could play a role in lightweight model inference where the largest ML models are unnecessary [4], we do not foresee a broader role for AI computation on FPGA while model sizes continue to grow. However, the role of application-specific algorithms like FE and LCS which prepare data to feed into AI engines remains relevant.

The fact that the same FPGAs hardware remained competitive from decision tree forests through a multiple DNNs demonstrated the platform's adaptability at scale. With less focus on AI, we envision more opportunity to focus advancing other categories of applications, particularly specialized program acceleration, where other commercial accelerators have not gained a significant foothold in the cloud.

III. Additional Lessons

Overall, Microsoft continues to benefit from the flexibility FPGAs enabled in the fleet. We 've rolled new features to large fractions of the fleet across multiple generations of FPGAs. While some features would emerge with software-programmable architectures, others (e.g. NVMe virtualization or DNN acceleration) would overwhelm core-based architectures. We've been successful at supporting increased server lifetimes, which has a positive impact on cost and sustainability.

One additional advantage of the FPGA approach is the impact FPGA accelerator development has on the quality of the pure software solution. During the first Bing development, the initial 625x improvement in DTS quickly reduced to 32x with software optimization before settling at 125x for just that portion. The software optimizations reduced the portion of time DTS took as part of the overall Bing stack, and Amdahl's Law necessitated finding new portions of Bing's ranking algorithm to justify deployment. As we found new portions of Bing to offload, pure software could be improved in similar ways. Hardware that enables software developers to do A/B comparisons in production helps software engineers identify true bottlenecks and think of new ways to overcome existing obstacles.

IV. Looking Forward

The evolution of cloud workloads is hard to predict, and changing commercial constraints make it even more difficult to know how to push specialization into the datacenter. 6+ year deployment lifetimes and supporting 4+ generations of specialized hardware simultaneously makes finding lasting solutions even more difficult.

While we have been successful at deploying and utilizing FPGAs at scale for a wide variety of use cases, on-boarding new applications is still dificult. We conjecture that generative AI can be applied to development, making programable accelerators more viable to deploy in cloud infrastructures, making reconfigurable computing (perhaps in more forms) even more important than they are today, but also for new "general for a domain" classes of accelerators to emerge.

REFERENCES

[1] A. M. Caulfield, E. S. Chung, A. Putnam, H. Angepat, J. Fowers, M. Haselman, S. Heil, M. Humphrey, P. Kaur, J.-Y. 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)*, 2016, pp. 1–13.

[2] D. Firestone, A. Putnam, S. Mundkur, D. Chiou, A. Dabagh, M. Andrewartha, H. Angepat, V. Bhanu, A. Caulfield, E. Chung, H. K. Chandrappa, S. Chaturmohta, M. Humphrey, J. Lavier, N. Lam, F. Liu, K. Ovtcharov, J. Padhye, G. Popuri, S. Raindel, T. Sapre, M. Shaw, G. Silva, M. Sivakumar, N. Srivastava, A. Verma, Q. Zuhair, D. Bansal, D. Burger, K. Vaid, D. A. Maltz, and A. Greenberg, "Azure Accelerated Networking: SmartNICs in the Public Cloud," in *NSDI*, 2018.

[3] J. Fowers, K. Ovtcharov, M. Papamichael, T. Massengill, M. Liu, D. Lo, S. Alkalay, M. Haselman, L. Adams, M. Ghandi, S. Heil, P. Patel, A. Sapek, G. Weisz, L. Woods, S. Lanka, S. K. Reinhardt, A. M. Caulfield, E. S. Chung, and D. Burger, "A configurable cloud-scale dnn processor for real-time ai," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, 2018, pp. 1–14.

[4] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro, J. Law, K. Lee, J. Lu, P. Noordhuis, M. Smelyanskiy, L. Xiong, and X. Wang, "Applied machine learning at facebook: A datacenter infrastructure perspective," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2018, pp. 620–629.

[5] K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, and E. S. Chung, "Accelerating deep convolutional neural networks using specialized hardware," 2015.

[6] P. Shan, "Increased remote storage performance with nvme-enabled ebsv5 vms now generally available," *Azure Compute Blog*, May 2023.

[7] N. Zhang and B. Brodersen, "The cost of flexibility in systems on a chip design for signal processing applications," *University of California, Berkeley, Tech. Rep*, 2002.