Hewlett Packard Enterprise

Hewlett Packard Labs

Heterogeneity: No Accelerator is an Island

Authors: Paolo Faraboschi with Dejan Milojicic

At the 2018 IEEE International Conference for Rebooting Computing, the keynote entitled “Computing in the Cambrian Era” used the term “Cambrian explosion” to describe the proliferation of accelerators (and corresponding startups) in the artificial intelligence/machine learning (AI/ML) space. The talk made the point that the computing world was on the path of becoming heterogeneous, and there was no turning back.

The term “Cambrian explosion” has since then become common. See, for example, the 2019 Forbes article “Compute Cambrian Explosion” by Tom Coughlin, where he describes several related trends across the industry.

Four years have passed since the ICRC keynote, and billions in funding have been poured into the space. Now it’s the time to take a step back and look at some of what we were saying back then and what has happened. In short, while the trend towards compute diversity and heterogeneity has continued, it has been much slower than anticipated. Very few new players have reached the point where they have products that can generate revenue, and overall the feeling in the community is that we may be on the verge of a different stage in this evolution process.

In August, side conversations during the HotChips conference gave rise to the term “Ordovician casualty” in reference to the geological era after the Cambrian explosion. What geologists discovered was that the Ordovician period was characterized by a harsh environment: a much colder climate, changes in the oxygenation of the ocean, and possibly frequent meteorite collisions, which caused a mass extinction that wiped out 85 percent of marine species. Similarly, five years into the “Cambrian explosion” era of AI/ML accelerators, we are starting to see trends that are “taking the oxygen away from the system.”

What are the challenges to the use of heterogeneity?

- Programmability
- Optimization of code
- Integration into the overall architecture

A new software stack has traditionally been the Achilles heel of every new hardware design. Despite the fact that AI/ML developers are converging to a few widely accepted frameworks like PyTorch and TensorFlow™, we are still far from the “make-and-run” approach of traditional software. Getting a model from zero to deployment remains a multi-month deployment effort and includes complex pipelines that have to deal with data selection and preparation, model design and optimization, training, neural architecture search, hyper-parameter optimization, validation, deployment, and finally, closing the loop on “trustworthy AI” metrics, such as bias, explainability, and robustness. In this process, any tool that deviates from known practices or introduces a new bug is a major impediment to productivity.

As Alexis Bjorlin of Meta recently said during her AI Hardware Summit keynote, "Innovation velocity always outweighs runtime efficiency." New hardware typically requires a new and unproven toolchain, different performance tuning recipes, and sometimes, when quantization and different precisions are involved, a different convergence process for training or accuracy for inference.

All of this requires data scientists and AI/ML engineers to spend more time and re-qualify tools and test their practices on them. It hurts productivity, at least in the initial stages. Of course, some other start-ups, like OctoML, are developing techniques to automate the mapping and optimization process, but we are still in the early stages of the journey.

What is the community doing wrong?

- Siloed approach
- Reinventing the wheel as opposed to focusing on novelty
- Thousand flowers blooming

Incumbents like NVIDIA® and Intel® are not standing still and are moving the competitive bar higher.

“For all these numerous start-ups out there who would like to eat our lunch, it’s a very rapidly moving lunch,” said NVIDIA’s Bill Dally in his keynote at ScaledML in 2019.

Indeed, since then, NVIDIA has already introduced two new generations of GPUs: Ampere in 2020 (up to 2.5x Volta performance) and Hopper in 2022 (up to 4.5x Ampere performance). So, accelerators whose design started in 2018 have seen their competition target move by up to 11x in performance on specific benchmarks. Google™ Tensor Processing Unit™ (TPU) has seen a similar progression.

Other large players like AMD and Qualcomm® have entered the market with massive investments. Some of the largest start-ups, like Cerebras, Graphcore®, and Sambanova, have raised several hundred million dollars in funding and already have a second generation of processors in production and sold to customers. Combined with the shift in macroeconomics—growing inflation, market uncertainty, and the increasing cost of money—this is likely to trigger a major “extinction” between 2023 and 2024 when some start-ups are going to reach the end of the last funding round.

Moore’s Law has slowed down, but it’s still very much alive. Optimizations are compensating for some of Moore’s deceleration in areas such as reduced precision and structured sparsity. Reduced precision uses fewer bits when not needed and better formats than integer and IEEE FP, like BFP16 and now BFP8. In structured sparsity, down to 10 percent of multiplications are needed; some have a zero element, some are duplicated, even more so at reduced precision, and several architectural approaches are popping up.

For accelerators that are just rearranging different combinations of digital logic and SRAM, the runway is becoming pretty short. Architectural innovations are a “one-time” play. Once you’ve revealed them, they don’t enable too many future improvement opportunities. Also, if you don’t take advantage of them right away, your competitors can and will start adopting them. This happened for tensor cores (matrix multiply engines) that have appeared in TPUs, then GPUs, then CPUs. It also happened with reduced precision, and it’s happening with structural sparsity.

To attempt addressing this shrinking differentiation gap, some companies have started to climb up the system ladder and build larger-scale offerings (not just processors, but boards, servers, or even entire clusters), blurring the boundaries between their core intellectual property and where system integrators normally play. In some cases, this has pushed them to reinvent the wheel in areas such as schedulers, networking, physical infrastructure, or storage that don’t add much novelty but stretch their engineering resources thin. More importantly, it has caused system integrators to perceive them as competition rather than partners, complicating conversations and slowing paths to market.

How market entrants reach working silicon?

- Reality hit hard with growing model sizes
- Market moved from Enterprise vs. Cloud providers market
- And price points moved too

Several accelerator start-ups have recently reached working silicon, typically with a one to two years delay compared to their initial estimates (partially due to COVID-19 side effects, but also because building a chip is never easy). They are now moving “from slides to reality” and starting to measure real performance on actual systems rather than simulated projections. This is adding a dose of reality to their projections with the additional difficulty that some of the AI/ML workloads have shifted since they started the design. For example, model size of state-of-the-art neural networks has outgrown any projection, and what looked like a large model five years ago is now insignificantly small when compared to some of the trillion-weight language transformers.

The market dynamics have also been shifting, and what we observed is that many hardware start-ups have shifted their focus to selling their products to the enterprise market, but realized that they had built it for a different market, the cloud. In the early days of the Cambrian explosion, everyone was targeting the large and growing cloud service provider (CSP) market as the success story to sell to their investors. So, they embraced design choices—physical form factors, integration points, software stack, management system, performance/power/cost envelope—that were a better match to CSPs. However, in the meantime, several CSPs ended up developing their own AI/ML engines, from the Google TPU to AWS's Inferentia and Trainium processors to the Alibaba Hanguang chip. Even end-users, like Tesla, have started building their own AI processors with the Dojo effort.

This meant that a large chunk of the initial target market had disappeared, so all the start-ups we talked to have begun to refocus their effort on the enterprise market, going through traditional system integrators like Dell, Atos, or HPE. Unfortunately, the enterprise market has different requirements, which require redesigns, changes to the software stack, and in general further delays. Finally, start-ups are now confronted with the problem of setting a price for their devices and facing the reality that they now have to compete with a far improved generation of their competitors, which, as we discussed above, have not been standing still. While absolute performance and performance-per-watt are important, performance-per-dollar of the end-to-end system is key, especially in the enterprise market.

What are the new areas of opportunity?

- Optimally matching software to hardware
- Providing unified frameworks to both hide and leverage heterogeneity
- Standardized interfaces

It takes two to tango. Just like RISC offloaded the complexity of hardware to software, there are still hidden opportunities in the infancy of AI for new algorithms and new software approaches that can motivate new hardware design. In other words, good old hardware-software co-design. Lots of innovation is still happening from the bottom up, from hardware architects, while there is quite a bit of opportunity for innovation at the intersection of HPC, AI, and data analytics. The more we think holistically, the higher our chances of success.

Transparency is costly, even when picking your poison. Heterogeneity was meant to be an answer to the deceleration of Moore's Law. But it introduced tremendous complexity in programming different instruction-set architectures, programming models, communication mechanisms, and different interconnects and fabrics. All attempts to hide this heterogeneity using layers of unifying frameworks address the latter problem but at tremendous performance costs. A careful exposure of heterogeneity to programmers in certain cases while hiding it in another is a non-trivial but absolutely required approach to take. Transparency is always costly.

No accelerator is an island. No matter how successful an accelerator architecture is, it never lives on its own but in conjunction with other processing units. To achieve effective balance, one needs to define standardized interfaces to maintain interoperability across components. In the history of computing, there are very few de facto standards that survived. Therefore, exposing interfaces, especially those critical for interoperability, is a win-win for accelerator vendors, system integrators, developers, and, ultimately, the user community.

"TAMO!" or, "Then, A Miracle Occurs", as our colleague Kirk Bresniker would say. The next step for many quantum computing scientists is based on TAMO. Given the amount of investment poured into quantum computing research and development, this is not an unreasonable assumption for quantum developers. However, for the rest of us in the community, we expect that some new breakthrough will happen, be it quantum or quantum-inspired, or quantum alternative, and start thinking about how to integrate this new citizen in the world of heterogeneity.

Heterogeneity is our friend and our frenemy. We need it, but we also need to overcome all the baggage that comes with it to leverage all its benefits.

Learn more at

hpe.com/us/en/hewlett-packard-labs

a00128625enw

Hewlett Packard Enterprise