E-Guide

Is the GPU the new CPU?

February 2022

TechTarget | Data Science Central
A COMMUNITY FOR BIG DATA PRACTITIONERS

In this e-guide:

Is the GPU the new CPU?

KURT CAGLE, COMMUNITY EDITOR FOR DATA SCIENCE CENTRAL

I bought my daughter a new computer recently for school. We took the laptop home, then watched, increasingly horrified as out of the box the computer lagged so bad that what should have taken a few minutes for setup took nearly an hour. Back to the store we went, and this time she pointed out a new gaming computer – one that had interesting colored lights on the keyboard, but more to the point also supported many of her favorite games, including Minecraft.

A couple of days later, she called me into her bedroom, quickly built a Minecraft castle that was mostly in the dark, then added fire in a fire pit. On her last laptop, the same castle would have become a brighter orange-yellow, reflected from the dominant color of the fire. With her new laptop (and its mid-range Nvidia RTX graphics card) the fire cast multiple interacting shadows on the wall, including the shadow of me standing before the fire projecting on the walls. Yet it was when she put the fire out and then stood in awe as the room around her character dimmed realistically in the waning firelight that I realized that the GPU had finally overshadowed its older CPU sibling.

In this e-guide:

Is the GPU the new CPU?

From Bit Player to Starring Role

A Graphical Processing Unit (GPU) started out with the CPU as the part of the processor responsible for rendering the screen in early GUI environments. Because so many of those operations involved transformations (mathematical operations largely built around matrix manipulation), it was an obvious candidate for splitting off from the core so that the more expensive operations could be done in parallel to the CPU's general coordination role.

As computer software became more sophisticated (especially in the realm of gaming) the GPU also became more powerful, as you need significant compute cycles to manage 3D rendering and compositing. The GPU architecture, already built around parallelization of processes, took on other roles as spreadsheets led to data analytics led to machine learning. Similarly, bitcoin mining, which involved solving things like large prime numbers to act as keys to provide scarcity (and hence value) as proof of work algorithms.

For Nvidia in particular, this indicated that more and more of the regular operations that a computer performed involved complex computational capabilities. Already by the 2010s, Nvidia and similar GPU manufacturers were selling units for high-performance computing (HPC) to cloud service providers. This necessitated the introduction of a software layer called CUDA that emulated the instruction sets of the CPU while taking advantage of highly parallel processing.

In this e-guide:

Is the GPU the new CPU?

The CUDA layer, written in C++, was then extended across multiple languages and platforms, including Python, Java, and more recently Nodejs/Javascript. This has meant that high volume data-center calculations could be performed regardless of operating system or language.

However, with CUDA, the flip-side is becoming increasingly true as well. As more applications (and operating systems or containers such as Kubernetes) move to the Cloud, the expectation is that data-centric hardware is available to do the heavy lifting of not only graphics but also queries. Indeed, increasingly the distinction of what is a query has moved beyond the “simple” text query to instead encompass the ability to pull in contextually related graphs from disparate datasets, the ability to query into a machine learning module for classification (or classification training), and the ability to look at clouds of sensor data as tensor fields that can be queried for specific configuration states.

Graphical Computing and GPUs

Not surprisingly, all three of these cases involve high-speed optimization of graph path traversal, from the recursive graphs that form neural networks to the hypergraphs that constitute knowledge networks and out to the interconnected sensor nodes that represent the sensor networks. These are all operations that require both massive parallelism and high-speed computation, and as they become more central to every aspect of computation, so too does the need for hardware that can bring these capabilities to the network.

In this e-guide:

Is the GPU the new CPU?

One can argue that the metaverse ultimately falls into this realm as well, especially if you look at the extended reality (XR) aspects of augmented or virtual reality as being a mix of Spatio-temporal networks and associated clouds of networked metadata as again driving the hardware best suited for network traversal and computation. The gaming industry is, as I have argued elsewhere, the precursor for the Metaverse, and the GPU has largely evolved with the gaming industry as the "place" where these calculations were most necessary.

Meanwhile, the role of the CPU as a stand-alone processor is shifting into a hind-brain one: it handles the bootstrapping process of bringing up the cerebral cortex of the GPU cloud, managing virtualization (when that too isn't being managed by GPUs), and coordinating with dedicated digital signal processors (DSPs) to manage the acquisition and transmission of "sensory" signals into the larger context of that same GPU environment.

Are DPUs the Next Iteration of GPUs?

It's not a far leap to see the integrative aspect of such sensory data forcing yet another evolution on both the CPU and the GPU. Processing and aggregating signals (in the very broad sense) has often been an integrative function (and one typically done manually at great cost and complexity).

In this e-guide:

Is the GPU the new CPU?

However, the idea of an autonomous data processing unit (DPU) has been one that has been floated for a while now that serves to aggregate and transform signals into queryable stores.

Right now, most of these DPUs are handled independently, but as the stack becomes more codified, so too does the possibility that DPUs will end up themselves becoming etched in silicon, utilizing existing knowledge and soft prototypes to then enable deep data processing in a consistent hardware-oriented approach. This may be handled by GPUs as well, especially as modern GPUs can readily carve off portions of themselves to dedicate to specific but similar tasks, though it is also possible that the requirements of DPUs may be addressable by chips that have different architectures altogether.

One of the more intriguing ideas coming from the W3C is the notion of data pods (also known as Solid Pods). Pods are essentially virtual graph databases controlled and mediated through GPUs and communicating over a standard protocol, an innovation that could alter the landscape of data processing profoundly. Because such pods are likely to be integral to a strategy for digital twins and IoT integration (likely working with DPUs, which themselves are specializations of GPUs), much of what we think about data storage is about to be radically rewritten.

In this e-guide:

Is the GPU the new CPU?

GPU Networks and the Future

There has been an ongoing move towards networkification, the process of building out networks of functional units across different scales and latencies in order to achieve processing in a way that can't be solved by single units alone. Not surprisingly, GPUs, as graph processors (and just graphics processors), are adapting well to this usage. Clusters of GPUs (with relatively low latency coordinated connections) are now replacing both general-purpose supercomputers and distributed CPU clusters, typically with GPUs, in turn, being segmented and specialized for tasks such as rendering, deep learning, graph query and coordination.

This is taking place even as graph databases (some running on GPUs such as Amazon's Neptune) and the imminent emergence of the W3C Solid standard make storing state and metadata between such devices over distributed networks feasible. By 2035, it is likely that the notion of a stand-alone processor will seem as archaic as the notion of a standalone database. Instead, cloud computing will likely have become a sea of GPUs in dynamic, configurable networks, with data stored in graph-based nodes mediated by GPU controllers, with the CPU relegated primarily to the role of booting up devices.