REPORT REPRINT

Think outside the CPU: the emergence of the GPU-powered database

JAMES CURTIS

12 JAN 2017

GPUs, or graphics processing units, are known to drive the graphics of popular video games, but we are starting to see GPUs power databases. With a slew of new vendors emerging, is this a fad or a legitimate market?

451 Research®

Since early 2015, we have witnessed the emergence of a growing trend within the database market: a slew of emerging vendors looking to leverage GPUs, or graphics processing units, to drive databases for data processing. Advances or movements within the broader database market are neither new nor unexpected, given that the relational model has been in existence for the better part of 40 years. But GPU-database vendors continue to intrigue us, in the sense that many of them leverage SQL but have been built specifically around GPUs – instead of CPUs – specifically targeting fast analytics.

THE 451 TAKE

While the overall database market is rather large, the GPU-database market is still in its early days. The fact that we have covered no less than five vendors recently would suggest there is something here. The performance metrics and the capabilities of GPUs are truly impressive. GPUs are also much more compact – there are many more cores per GPU compared with CPUs, leading to an overall smaller footprint in the datacenter. However, we would like to see the market progress more in terms of GPU adoption. Cloud options are improving, but are nowhere near what is available for CPU-based machines. And if history has taught us anything, it's that performance never seems to go out of style. Mix that with an enterprise's need for better analytics, and we have an ideal match – all of which points to a bright future for GPUs.

The overall database market is quite robust. At 451 Research, we segment the database market into a number of categories. For operational databases, which include vendors driving operational business applications, including relational, NoSQL and NewSQL data models, we have estimated the potential CAGR to be 13%, according to our December 2016 Market Monitor report. For our analytical database category, in which we would include the emerging GPU database vendors, we have estimated the potential CAGR to be 10%. More noteworthy is that, for vendors with revenue of less than $1bn, we have estimated a CAGR of 39%. The fact that we see new and emerging vendors pushing the database boundaries is quite encouraging. After all, there's still a good amount of money to be made.

Sometimes CPUs and GPUs are discussed in terms of one being better than the other. In reality, the architectures are quite different, leading to an apples-to-oranges comparison. At a high level, a CPU is a very good general processor, handling a variety of complex tasks well. A GPU, on the other hand, is more specialized in nature and can do certain tasks extremely well, but not all tasks. Generally speaking, CPUs operate serially, in that tasks are performed in some kind of order. Conversely, GPUs process data in parallel, in that many tasks can be performed simultaneously. That being the case, it is often asked why GPUs aren't the preferred processing method over the ubiquitous CPUs? The simple answer is that GPUs simply aren't good at some tasks; in fact, sometimes they are worse from a performance standpoint. It's more of an issue of leveraging the GPUs for what they do well. For instance, GPUs can be orders of magnitude faster at joins, and much slower at updates and deletes.

WHAT'S DRIVING THE MARKET?

What's driving the recent emergence of several new GPU database vendors? It's likely not a single event, but a confluence of several events. In August 1999, NVIDIA announced its GeForce 256 GPU chip, which is often referred to as the first graphics processing unit chip, even though graphics-processing technology has been around for decades. NVIDIA's announcement, however, essentially codified the term GPU along with the company's own definition of what constituted a GPU chip. We can also point to the movement of general-purpose processing on GPUs, which is estimated to have begun around 2001, as well as 2007 when the CUDA programming language was released. General-purpose computing on GPUs enables developers to programmatically leverage GPUs to perform some tasks that were normally handled by the CPU. We can also reference the so-called big-data movement and the rise in the amount of data that organizations can collect, store and analyze. GPUs are particularly skilled at certain analytical workloads and, given their processing capability, are able to handle extremely large datasets.

WHAT DOES THE FUTURE HOLD?

While the market is showing signs of GPU growth, we also see a few hurdles. One is the ability for organizations to consume GPU technology. Many vendors often lead with an on-premises strategy. It's not that these vendors don't have a cloud offering (BlazingDB has a cloud-first approach), but it's often the limited options for the cloud. Amazon AWS offer its GPU G2 instances and recently released its beefed up P2 instances. IBM has GPU instances on SoftLayer (branded Bluemix), and Google Cloud Platform recently announced GPU instances. Microsoft Azure has GPU instances available for preview. There is also the fact that GPUs requires specific programming. Applications and algorithms need to be programmed specifically to leverage the GPUs, which require special skills with CUDA and OpenGL (GPU programming languages).

LOOKING AT THE EMERGING PLAYERS

While the market is still nascent, we have covered a number of these vendors and identified some of their differences. We do, however, expect others to emerge.

Kinetica traces its origins to 2009, when Amit Vij and Nima Negahban were driving an IT consultancy firm at the US government. The company's first contract was with the US Army Intelligence division to handle hundreds of real-time data feeds, where analytics could then be run on the data in real time. Kinetica often positions itself as providing a 'speed layer,' where its offering can be integrated into an existing environment, such a Hadoop. The company takes a hybrid approach by combining the benefits of GPUs with CPUs – its offering runs in CPU memory, but leverages GPUs for certain workloads.

MapD was founded in 2013 by Todd Mostak, who based the core MapD technology on his research at MIT. MapD offers two core products: a GPU database and a visualization front-end tool. The company's pitch is that, while GPUs offer great performance advantages with analytics, visualization is equally important. MapD is differentiated in that it is optimized to run in GPU memory, although the company claims it can run in CPU or a hybrid CPU and GPU setup.

BlazingDB was founded in 2015 by brothers Rodrigo and Felipe Aramburu. The brothers were operating an IT consultancy firm in Peru and working on a project that involved carrying out multi-table, multi-column joins on approximately 14 different legacy database systems. The fix turned out to be a GPU-based table joiner that eventually served as the basis for the company's current offering. BlazingDB positions itself as targeting enterprises that have the largest of large datasets – BlazingDB runs on disk as opposed to in-memory.

SQream Technologies was founded in 2010 by Ami Gal and Kostya Varakin, and is based in Ramat Gan, Israel, which is east of Tel Aviv. Similarly to the other vendors, SQream drives a performance message. The company is differentiated by the fact that it runs a NoSQL database underneath while it drives certain workloads to the GPUs for processing.

Others worth noting include brytlyt, which offers a massively parallel processing database based on a PostgreSQL clone that pushes number-crunching workloads to GPUs. Another is Blazegraph, which drives a graph database but uses GPUs for processing. Databricks, the company that commercially supports Apache Spark, recently announced that it was adding clusters with GPU support, primarily meant to drive deep learning workloads.