Breaking up is hard to do

Accelerators are changing the way servers are being put together, putting the focus on fast interconnects. By **Chris Edwards**

In mid-March, graphics processor maker NVIDIA decided to outbid Intel by close to a reported $1bn in the pursuit of networking chipmaker Mellanox. For the NVIDIA of several years ago, it would have been an acquisition that made little sense, but today it is no longer serving just framerate-hungry players of Call of Duty.

CEO Jen-Hsun Huang sees the future of the company as being bound increasingly to data-centre servers.

The Mellanox purchase is part of a trend that is remaking the server from the inside out, driven half by the inability of designers of general-purpose processors to squeeze much more performance out of their architectures or Moore's Law and half by the massive growth in demand for machine learning among the likes of Baidu, Facebook, Google and Microsoft.

In this new model, data-centre servers begin to look more like supercomputers where local memory is mainly just cache. What used to be closely coupled DRAM moves into dedicated storage subsystems connected by a predominantly serial interconnect matrix, often represented by PCIe inside the shelves and blades. At the rack level and above, Ethernet is the carrier.

Huang told analysts on a conference call to explain the planned acquisition: "The dynamic that is happening here is that, in the future, it won't just be server-scale computing that people do, but it will be data centre-scale computing, where the network becomes an extension of the computing fabric."

Days later, the data-centre owners and component suppliers gathered at a conference 10km southeast of NVIDIA's HQ in downtown San Jose to demonstrate how they are breaking up the server to remake it.

A decade ago, it would have been big-iron vendors of the likes of Hewlett-Packard and IBM who dominated discussions about server architecture. Today, it is the data-centre owners themselves who are designing and building the hardware that goes into their racks, all the way from chips to enclosures, albeit with a great deal of reliance on contract manufacturers.

The rise of machine learning in the cloud has done much to make the designers rethink how their systems are put together. In his keynote at the OCP Summit, Facebook's director of technology strategy, Vijay Rao, claimed the half the company's data warehouse feeds into machine-learning algorithms that handle translations and numerous other services. He claimed some six billion translations requests a day go through Facebook's AI.

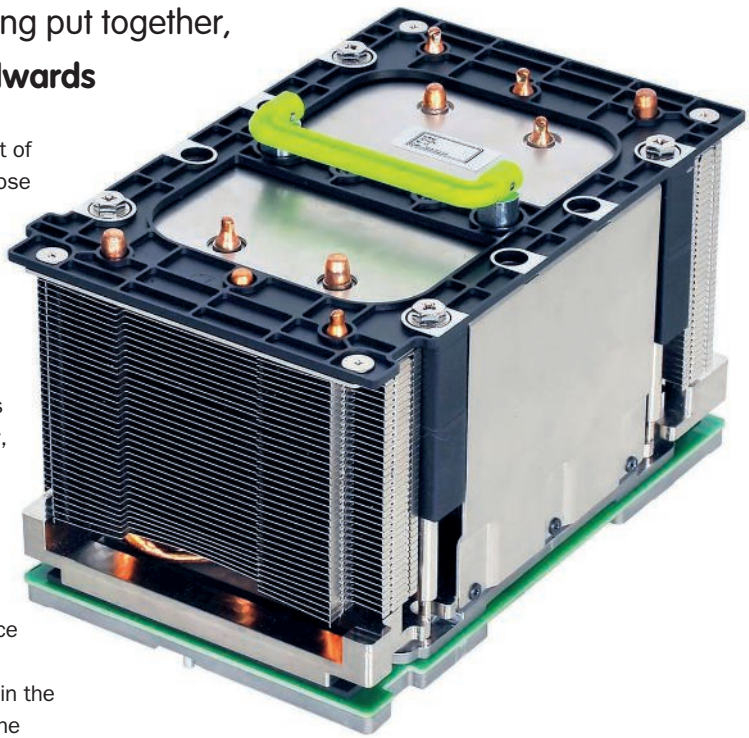

Figure 1: The Open Accelerator Module occupies the space of a small lunch box

Domain specific acceleration

General-purpose processors would collapse under that level of demand. So, Facebook and others have embraced domain-specific acceleration as the way to access the huge number of matrix multiplies needed to handle deep-learning models. But raw compute throughput is only half the story. Some of the models are enormous, demanding memory footprints of up to 2TB, according to Rao. No single accelerator can deal with such a large model. The processing needs to be spread across multiple chips, putting much greater emphasis on the interconnect between processing engines than ever before.

"Engineers designing a server used to make trade-offs between four basic elements: compute, storage, memory, and the network. Accelerators have now joined this group and trade-offs are between compute, memory, accelerators and buffers," Rao said.

Rao described how Facebook's main server design is now "tricked out with accelerators" that are assembled onto carrier cards that act as slot-in replacements for Intel Xeon processor cards, all connected by PCIe.

PCIe is likely to be supplanted by one of the growing number of communications standards vying for a place in high-performance computing. They include CCIX, Gen-Z, NVIDIA's proprietary nvLink and OpenCAPI, all of which are intended to provide high-speed paths between processors and memory. Just last month, Intel and a group of allies put another option in front of system designers: Compute

Express Link (CXL). A key distinction between these protocols and PCIe is that they support various forms of cache coherency between processor subsystems and the memory they use. This makes it possible to make accelerators peers of the processors that manage and, more importantly, avoid the need for software to explicitly shuttle data in and out of each accelerator's buffers. The software can simply provide a link to the incoming data and the coherency protocol will ensure the most up-to-date elements are mirrored into the accelerator's scratchpad.

Gaurav Singh, corporate vice president of silicon architecture and verification at Xilinx, says the kinds of cache coherency supported in CCIX are likely to be important for data-centre designers: "NVIDIA has made a case for how nvLink can benefit the machine-learning training use-case by allowing many GPUs to work on the same data set. CCIX enables the same functionality. With machine-learning inference, as datasets increase in size and with the move towards higher resolution images, CCIX can allow multiple accelerators to share data."

There are, however, potential downsides to the approach. caused by increased inter-memory traffic that may lead to unwanted high energy consumption. For this reason, Open Domain-Specific Architecture (ODSA), an OCP group working on ways to wire up accelerator chiplets inside a single package expect there to continue to be a mixture of non-coherent and coherent protocols in use in these systems. Facebook's Zion server design uses a coherent interconnect to link general-purpose processors to accelerators. But the accelerators have their own non-coherent fabric.

The question facing both groups like the ODSA and the server builders is which high-speed protocols and interfaces to support. According to Bapi Vinnakota, director of silicon-architecture programme management at network-processor specialist Netronome and a member of the ODSA workshop, one way to avoid having to make that decision as a group is to back PCIe's digital interface to physical-layer modules, known as PIPE. Chipmakers will be able to decide which electrical-layer links to support underneath and also pick and choose between various protocols that can sit on top of PIPE.

Fewer differences

In practice, there are fewer differences between the various protocols being put forward at the electrical level than at first appears. CCIX and CXL, for example, are both built on top of PCIe. CXL happens to focus on the as-yet uncomplete version 5.0 of PCIe. Anthony Torza, distinguished engineer at Xilinx, points out that the company can use the existing 32Gb/s serdes found in its 16nm UltraScale+ products to support PCIe version 5.0.

"Accelerators have now joined traditional elements and trade-offs are between compute, memory, accelerators and buffers." Vijay Rao

Gen-Z focuses on longer-distance memory transfers and is focused more on physical-layer interfaces similar to those used for high-speed internet. According to Dell senior architect Greg Casey, Gen-Z will support the same kind of 112Gb/s PAM4 interface as that needed for the 400G Ethernet that will interconnect server racks. That will let Gen-Z most likely serve as a longer-distance fabric that joins together CCIX or CXL-linked subsystems.

Work at Facebook to move beyond its existing PCIe-based mezzanine accelerators is taking a similarly interface-agnostic path to that of ODSA. Developed in conjunction with Baidu and Microsoft, the Open Accelerator Module (OAM) is a design that, when packaged with heatsink able to deal with nearly 1kW of power, occupies the space of a small lunchbox. These modules will slide into docks from the front of a rack to allow easy upgrades.

"We are proposing to build a universal baseboard [to host OAM daughterboards] that supports different interconnect topologies," Siamak Tavallaei, principal architect at Microsoft, said at the OCP Summit.

By building switches into the daughterboards, the baseboard should be able to handle many of the topologies used by large ML workloads, such as 3D mesh and torus connections. "Topologies have dependencies on the types of accelerators being used as well as the application," Tavallaei noted. As far as the baseboard itself is concerned: "Wires should be just wires."

Facebook hardware engineer Whitney Zhao said the team opted to use Molex's Mirror Mezzanine connector to link base and daughterboards. "It can support 56Gb/s NRZ or 112Gb/s PAM4 signalling. We don't know what speed we will need to support on OAM but we know, whatever it is, the connector will be able to handle it."

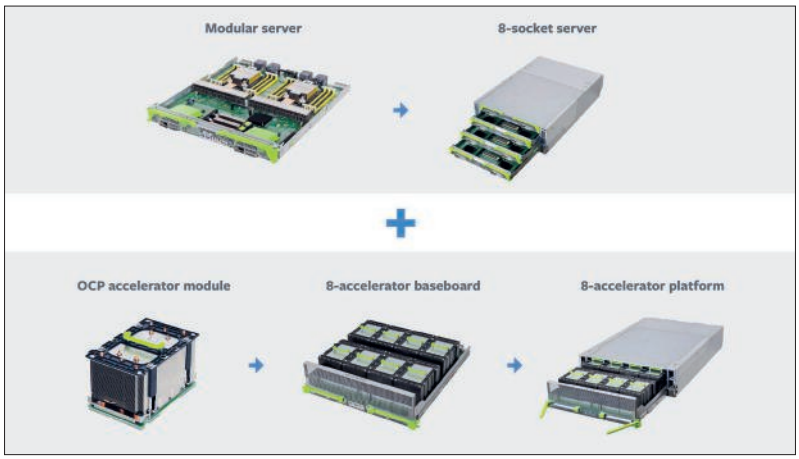

Figure 2: Modules will be able to slide into docks to allow for easy upgrades