URL : https://www.itbusinessedge.com/blogs/infrastructure/chip-design-starts-to-target-ai.html

Print Article

Chip Design Starts to Target AI

Artificial intelligence (AI) is clearly at the top of the enterprise's priority list for the coming year, but as we mentioned last week, it may be a little while before the revolutionary promises of this particular technology bear real fruit.

Part of the reason for that is the fact that traditional data infrastructure is simply not up to the task of crunching numbers at the scale needed to produce meaningful results from intelligent systems and applications. And unlike technology shifts in the past, supporting AI workloads is not merely a matter of deploying new hardware or increasing performance, although these are important, but in crafting entirely new ways to gather, interpret and disseminate digital information.

On a fundamental level, this will require changes to core processing infrastructure, but when it comes to enterprise-class workloads, it seems the work in this area is still in a fairly nascent stage.

Part of this effort is to push the core count to new extremes. Researchers at Washington State University and Carnegie Mellon University are working on a **datacenter-on-chip (DoC) architecture** that everyone from Google and Amazon to the U.S. military can use to kick AI performance into high gear. The idea is to combine perhaps thousands of GPUs and CPUs on a single chip, utilizing both wired and wireless communications to boost their interactivity. Details are expected to be released next month at the Big Data for HLS expo in Israel.

Meanwhile, the European Commission has instituted the **Mont-Blanc 2020 project** to devise an ARM-based SoC for exascale workloads. The effort has drawn organizations like ARM Ltd., Atos/Bull, Kalray Corp. and SemiDynamics to work out a solution that combines a manycore architecture with low-power ASIC and FPGA implementations in support of applications ranging from storage and networking to autonomous vehicles. The goal here is not to produce a commercial product but to develop a basic blueprint for attributes like vector length, network and memory bandwidth and a workable power envelope.

This is not to say that AI-ready systems have yet to hit the channel. At the moment, however, the necessary integration is taking place in hardware, not silicon. One example is **IBM's new Power System AC922 server**, which packs a pair of the company's Power9 CPUs with up to six Nvidia Tesla V100 GPUs and support for no less than three memory interfaces (PCIe 4.0, NVLink 2.0 and OpenCAPI) to nearly quadruple performance over a standard x86 machine. The company says the design is suitable for AI applications like neural networks and deep learning, as well as more traditional HPC functions, both of which are typically hampered by bandwidth limitations between cores rather than the processing power of the cores themselves. (Disclosure: I provide content services to IBM.)

But in an age when hyperscale data companies are crafting their own infrastructure solutions, it should come as no surprise that these efforts are beginning to target AI. **Alibaba recently turned to NXP Semiconductors** to integrate the company's IoT operating system into NXP's various application processors, microcontroller chips and multicore devices, with an eye toward powering smart cars, smart retail ecosystems and smart homes. Although this solution is aimed more at the edge than core data center workloads, it nevertheless represents the need for increasingly diverse processing architectures to accommodate the growing complexity of data infrastructure.

The fact that emerging AI platforms are being built around multiple cores and high-speed interconnects points up the fact that processing alone is not enough for a successful application. To a significant degree, AI will require high levels of orchestration between disparate resources, at least if developers hope to craft a truly unique user experience.

The foundation of good AI will be in silicon, but the real magic will come through a top-to-bottom reimagining of the data ecosystem.

Arthur Cole writes about infrastructure for IT Business Edge. Cole has been covering the high-tech media and computing industries for more than 20 years, having served as editor of TV Technology, Video Technology News, Internet News and Multimedia Weekly. His contributions have appeared in Communications Today and Enterprise Networking Planet and as web content for numerous high-tech clients like TwinStrata and Carpathia. Follow Art on Twitter @acole602.

Print Article