CASE STUDY

High Performance Computing (HPC)
with Intel® Omni-Path Architecture

Tokyo Institute of Technology Chooses Intel® Omni-Path Architecture for Tsubame 3

Price/performance, thermal stability, and adaptive routing are key features for enabling #1 on Green 500 list

Hewlett Packard Enterprise

Tsubame at a glance

- Tsubame 3, the second-generation large, production cluster based on heterogeneous computing at Tokyo Institute of Technology (Tokyo Tech); #61 on June 2017 Top 500 list and #1 on June 2017 Green 500 list
- The system based upon HPE Apollo* 8600 blades, which are smaller than a 1U server, contains 540 very fat nodes with four GPUs (2160 total), two 14-core Intel® Xeon® processor E5-2680 v4 (15,120 cores total), 2 TB of Intel® SSD DC Product Family for NVMe, and two dual-port Intel® Omni-Path Architecture (Intel® OPA) Host Fabric Adapters (HFAs)
- Interconnects a key aspect of the design: NVLink* interconnect between GPUs, fully switched PCIe fabric per node, single Intel OPA port for each GPU (2160 ports total)
- Intel OPA chosen for its price, low power and high thermal stability, and end-to-end adaptive routing

Challenge

How do you make a good thing better? Professor Satoshi Matsuoka of the Tokyo Institute of Technology (Tokyo Tech) has been designing and building high-performance computing (HPC) clusters for 20 years. Among the systems he and his team at Tokyo Tech have architected, Tsubame 1 (2006) and Tsubame 2 (2010) have shown him the importance of heterogeneous HPC systems for scientific research, analytics, and artificial intelligence (AI). Tsubame 2, built on Intel® Xeon® processors and Nvidia* GPUs with InfiniBand* QDR, was Japan's first peta-scale HPC production system that achieved #4 on the Top500, was the #1 Green 500 production supercomputer, and was the fastest supercomputer in Japan at the time.

東京工業大学
Tokyo Institute of Technology

For Matsuoka, the next-generation machine needed to take all the goodness of Tsubame 2, enhance it with new technologies to not only advance all the current and latest generations of simulation codes, but also drive the latest application targets—which included deep learning/machine learning, AI, and very big data analytics—and make it more efficient that its predecessor. "It was natural that while designing Tsubame 3, we would enhance our heterogeneous design efforts of Tsubame 2, not only for performance, but for efficiency as well," said Matsuoka.

Solution

"Tsubame 3 is the second-generation large-scale production heterogeneous supercomputer at Tokyo Tech," added Matsuoka. "When you look at the machine based on HPE Apollo* 8600 blades, it doesn't appear that big. There are only 540 nodes. But they are very different, very fat nodes with four Nvidia* Tesla* GP100 GPUs, two 14-core Intel Xeon processors E5-2680 v4, two dual-port Intel OPA host fabric adapters (HFAs), and 2 TB of NVMe storage, all in a non-traditional server design that takes up only 1U of space."

Interconnect is Key

Key to the design is the consideration of the multiple interconnects. "Tsubame 3 has a large number of many-core processors coupled with the Intel Xeon processors," commented Matsuoka. "It is designed for both analytics and AI workloads to co-exist with simulation workloads and run synergistically. So, we need to be able to efficiently manage, use, and move data within the node and across the system without bottlenecks. We have vast interconnects; we need bandwidth everywhere."

To alleviate bottlenecks while computing with massive data sets, Tsubame 3 has three switched interconnects running at full bandwidths across the machine. In

each node, the GPUs have their very own, proprietary NVLink* interconnect between them running at 20 GB/sec. A switched PCIe network interconnects the CPUs, storage, Intel OPA HFAs, and GPUs to the rest of the server and other components in the system at 16 GB/sec. The inter-node communication is through 100 gigabit Intel OPA (12.5 GB/sec), with an Intel OPA port for each GPU. "There's less than a 2:1 difference between the bandwidth of these links," stated Matsuoka. "If you look at the machine within the node and across the nodes, any two components communicating with each other have a connection path at a minimum of 12.5 GB/sec."

Why Intel® OPA

For choosing the fabric, Matsuoka benchmarked both Intel OPA and 100 gigabit InfiniBand, which were nearly equal in performance, but there were other critical factors to consider. "We were going to have this very extensive fabric with a very high number of injection ports and many optical cables to accommodate a full bi-section network," said Matsuoka. "Three factors favored Intel OPA over EDR. Intel OPA was lower cost. It has lower power and thermal stability. And Intel OPA adaptive routing works well, from the testing we've done and the observations we made on large machines, like Oakforest-PACS at University of Tokyo." With such dense nodes in Tsubame 3, thermal stability was a key concern, considering the heat generated within the chassis—even with the liquid cooling to the GPUs and CPUs.

According to Matsuoka, one of the bigger challenges with Tsubame 2 was the lack of proper adaptive routing with InfiniBand QDR, especially to deal with the degenerative effects of optical cable aging. "Over time AOCs die. And there is some delay between detecting a bad cable and actually replacing it. These can significantly impact a job running on the machine," commented Matsuoka. "So, with these key features, including Intel OPA's end-to-end adaptive routing, we chose the Intel fabric for Tsubame 3."

Results

When Tsubame 2 was built, AI was not in the mainstream that it is today. But over the last few years, as new data sciences have been maturing, machine learning/deep learning workloads have been running on Tsubame 2. Tsubame 3 was designed to support both existing and emerging AI workloads. "We believe that a lot of applications running on Tsubame 2 will carry over to Tsubame 3, because the high injection rates of Intel OPA will support the network bandwidth-sensitive codes, such as atmospheric weather, computational fluid dynamics, material science, and others. But also, these high injection rates will allow us to accelerate big data and scalable learning."

Matsuoka's research with machine learning has shown that for small-scale learning, network bandwidth doesn't really matter that much, because each chip has a lot of capability by itself. But when multiple many-core processors are employed in large-scale machine learning, or scalable learning, then network injection bandwidth in the nodes plays a significant role in sustaining scalability. "Having this excess amount of injection bandwidth will allow these machine learning and deep learning workloads to scale and allow us to accelerate those applications," he concluded.

The high bandwidth of all the interconnects in the system also plays an important role in supporting the co-existence of both analytics/machine learning and simulation. "We have a partnership with a group that does weather forecasting," described Matsuoka. "They want to run the weather simulations in real time, while they collect sophisticated, streaming sensor data, like from phase array radar that generates a terabyte per second. They couple the two workloads so that they can continuously correct the trajectory of the simulation based on the analysis of the incoming data. The only way to do this is to co-locate the workloads on the same machine with enough bandwidth across the system." Tsubame 3 was built to support these kinds of usages, of which there are many, according to Matsuoka.

At the time of measurement for the June 2017 TOP500 rankings, Tokyo Tech only ran on a small subset of the full configuration of Tsubame 3 and achieved #61 on the Top 500 and #1 on the Green 500 with 14.11 gigaflops/watt, an RMax of just under 2 petaflops, and a theoretical peak of over 3 petaflops. Tsubame 3 became fully operational August 1, 2017 with its full 12.1 petaflops configuration, and the hope is for even better rankings on the November benchmark lists including the Top500 and the Green500.

Solution Summary

Tsubame 3 is the second-generation, large-scale production cluster at Tokyo Institute of Technology that is built on HPE Apollo 8600 blades with a heterogeneous computing architecture incorporating Nvidia GPUs, Intel Xeon processor E5-2680 v4, Intel SSD DC Product Family NVME storage devices, and dual port Intel OPA HFAs in a 1U chassis. It was designed to run co-located big data analytics and machine learning/deep learning with simulation workloads. Three interconnect technologies in the system—NVLink for the GPUs, PCIe for the nodes, and Intel OPA for the system fabric—enable large-scale machine learning and powerful simulation. Key to the Intel OPA selection was its price, low power and high thermal stability, and end-to-end adaptive routing capability.

Where to Get More Information

Learn more about Tsubame 3 at http://www.gsic.Tokyo Tech.ac.jp/en.

Learn more about Intel Omni-Path Architecture at https://www.intel.com/hpcfabrics.

102017/RJMJ/RJ/PDF

Please Recycle 336635-001US