Technology to reduce network switches in cluster supercomputers by 40 percent
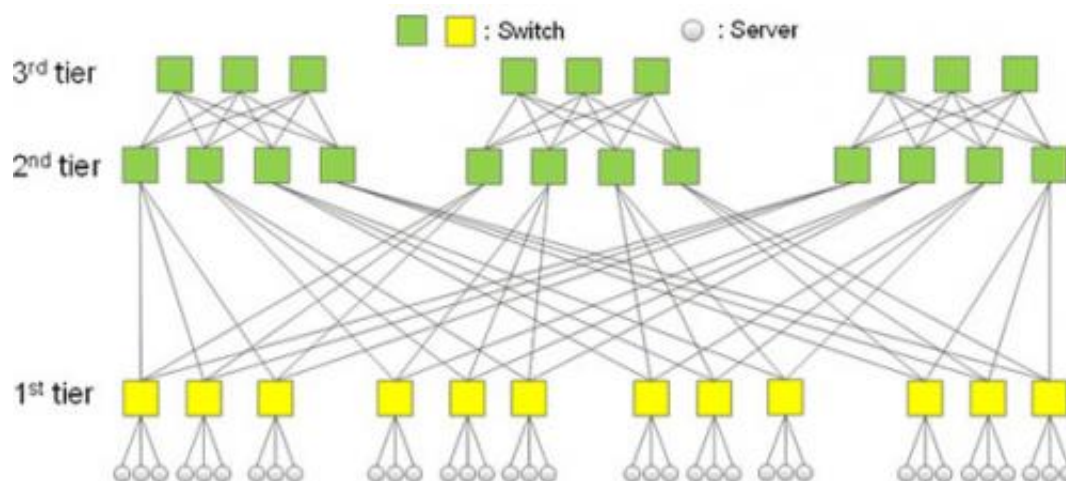
July 16 2014

Figure 1: A fat-tree network topology

Fujitsu Laboratories today announced that it has developed a technology that reduces the number of network switches used in a cluster supercomputer system comprised of several thousand units by 40% while maintaining the same level of network performance. Existing cluster supercomputers typically use a "fat tree" network topology, in which, for example, 6,000 servers would require about 800 switches, or possibly more than 2,000 switches, with network performance that needs redundancy and other features. Networks account for up to about 20% of the power consumed by a supercomputer system, which means there are high expectations for a new network technology that can maintain

good network performance with fewer switches. Fujitsu Laboratories has used a multi-layer full mesh topology in combination with a newly developed communications algorithm that controls transmission sequences to avoid data collisions. This means that, even in all-to-all communications, which are prone to bottlenecks during application execution, performance stays on par with existing technology while using roughly 40% fewer switches, saving energy without sacrificing performance.

Details of this technology are being presented at the Summer United Workshops on Parallel, Distributed and Cooperative Processing 2014 (SWoPP 2014), opening July 28 in Niigata City, Japan.

Background

Cluster supercomputers have been widely used in the fields of manufacturing, such as for the design of mobile phones, cars, and airplanes, as well as scientific technology computing. Increasingly, though, they are being used in new areas, such as in in silico drug discovery and medicine, and to analyze earthquakes and weather phenomena, and these applications require even more powerful supercomputers. To realize increased supercomputing performance, multiple servers are connected by networks. These servers are equipped with high-performance computation units consisting of accelerators that are typically many-core processors which have multiple CPUs or GPGPUs.

Technological Issues

In order for the supercomputer's computing performance to be useful to a wide range of applications, the network joining the servers needs to have higher performance. In the fat-tree network topology (Figure 1),

tiers are set based on the extent of the servers being connected, and the redundancy of paths in the tree-like network topology that connects the switches results in fast network performance. For example, a system with 6,000 servers would require 800 switches, each with 36 ports, to connect them. Thanks to the redundancy of routes in the fat-tree topology, when running a fast Fourier transform, for example, as part of an analysis on a cluster supercomputer, all-to-all communications among the servers shows good network performance. Meanwhile, many-core processors in individual servers or accelerators such as GPGPUs produce dramatic jumps in performance. Network performance needs to be improved so that it stays balanced with computational performance, and this requires many more switches, but increasing the number of switches entails the problem of higher costs for materials, electric power, and installed space.
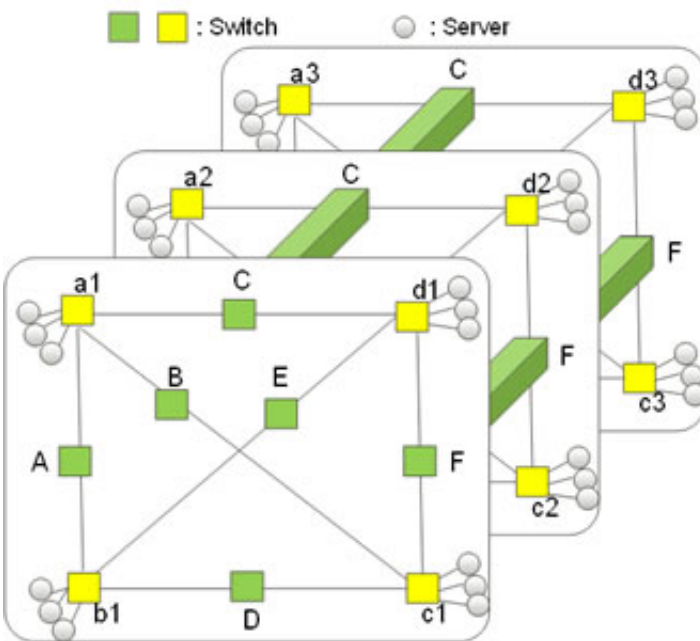

Figure 2: A multi-layer full mesh network topology (mock-3D representation)

About the Technology

What Fujitsu Laboratories has done is to develop a technology that can accommodate a large number of servers with relatively few switches by considering what would be an optimized data-exchange process, then connecting the cluster in a new way. This reduces the number of switches needed to connect a given number of nodes by roughly 40% compared to a fat-tree network topology while maintaining equivalent performance levels under the maximum-load communication pattern of all-to-all communications. Key features of the technology are as follows.

1. Multi-layer full-mesh network topology (Figures 2, 3)

Fujitsu Laboratories developed a structure where switches for indirect connections are arrayed around the periphery of a full-mesh framework that connects all switches directly, and multiple full-mesh structures are connected to each other. Compared to a three-layer fat-tree network topology (Figure 1), this eliminates an entire layer of switches, with switch ports being used more efficiently and a smaller number of switches in use.

2. Data-exchange process avoids path contention

In all-to-all communications, where each server is exchanging data with every other server, reducing the number of switches also reduces the number of paths between servers, which is likely to result in collisions. Fujitsu Laboratories was able to achieve all-to-all communications performance on par with a fat-tree topology by taking advantage of the multi-layer full mesh network topology in the process of transferring data between servers. By using scheduling, servers connected to the various apex switches (A through F) will divert to a different apex, and

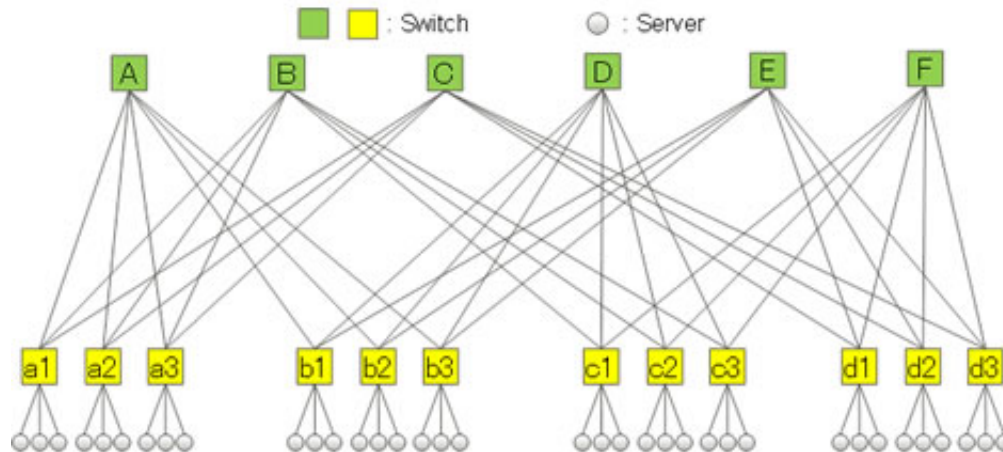also by avoid collisions within paths that traverse different layers (a1 through d3).

Figure 3: A multi-layer full mesh network topology (flat representation)

Results

This technology makes it possible to maintain the performance of large-scale cluster supercomputers that are needed for such applications as drug discovery and medicine, and to analyze earthquakes and weather phenomena, while lowering facility costs and power costs. This thereby enables the provision of supercomputers that achieve high performance while conserving energy.

Future Plans

Fujitsu Laboratories plans to have a practical implementation of this technology during fiscal 2015. It also plans to continue research into topologies for large-scale computing systems that do not depend on increasing numbers of switches.