Distributed Core Architecture Using the Z9000 Core Switching System

A Dell Technical White Paper

September 2011

Contents

Figures

Automating the Network

As data centers scale to support thousands of servers, IT managers are seeking better ways to network those servers while reducing costs and power consumption. Moreover, in large-scale data center cluster environments inter-node communication bandwidth is increasingly becoming the main bottleneck. For these environments, applications need to exchange information with remote nodes for execution of their local computation. For example, web search engines require parallel communication with every node in the cluster to provide the most relevant results, and web servers may require interaction with hundreds of sub-services running on remote nodes.

Compute nodes located across different physical switches may not have full bandwidth in a conventional hierarchical network design of interconnected switches. In these designs, overall bandwidth is limited by the bandwidth available at the root of a hierarchical communication tree. The solution is a distributed core architecture based on low-cost, high-capacity switches. This paper describes the use of Dell Force10's Z9000™ core switching system in a distributed core architecture to address these issues.

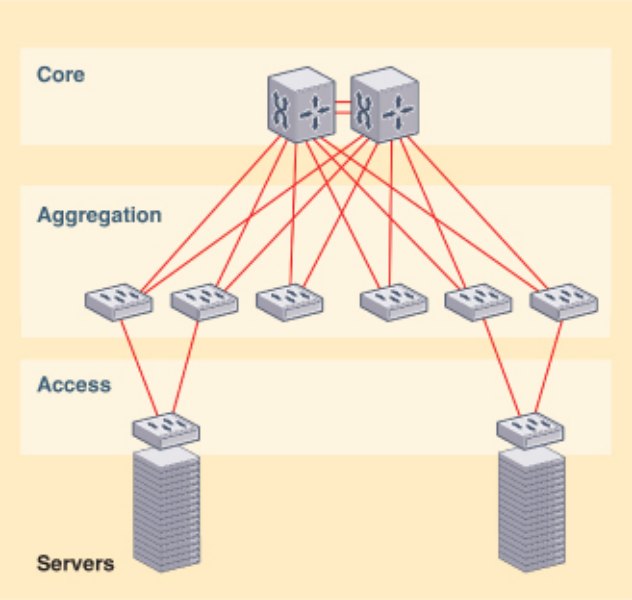
The 'Big Data' Challenge

A growing number of websites, portals, search engines, and analytical applications are dealing with extremely large data sets, otherwise known as 'big data.' Big data reflects data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes can range from a few dozen terabytes to many petabytes of data in a single data set. Big data requires high-performance systems to efficiently process large quantities of data in real time or near real time. Technologies being applied to big data include massively parallel processing (MPP) databases, data-mining grids, Apache Hadoop Framework, distributed file systems, distributed databases, MapReduce algorithms, cloud computing platforms, the Internet, and archival storage systems.

Big data typically employs large compute clusters and advanced techniques and algorithms to reduce data sets and control how data moves to and from servers. It also requires new networking architectures that connect computers in a very fast, high-performance way.

Networking vendors are reacting to these demands with new networking fabric configurations that support large compute clusters. Traditional hierarchical network designs are useful for certain types of data centers, but they're not well suited for big data compute-cluster applications because of the predominant east-west traffic patterns (Figure 1).

Figure 1. Conventional 3-Tier Data Center Architecture: Core, Aggregation, and Access

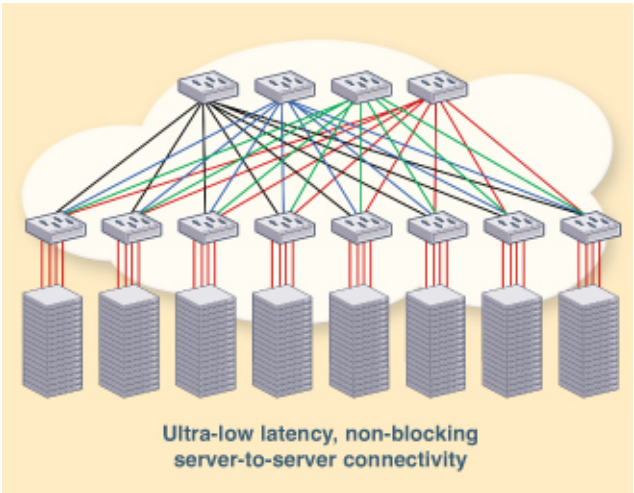


Traditional data center architectures excel when traffic patterns are predominantly north and south, in other words in and out of the data center. But when traffic is predominantly east-west in nature, as is the case with compute clusters, a distributed core architecture is best suited for the task.

Distributed Core Networks

Distributed core networks are purpose-built, high-performance networking fabrics capable of scaling to 160+ Tbps. (Figure 2). Commonly referred to as a 'leaf-spine architecture,' this architecture employs two types of nodes: one that connects servers or top-of-rack elements (leaf node) and the second that connects switches (spine node). When configured properly, a 3-stage Clos network can be derived from the leaf-spine system providing extremely low-latency, non-blocking performance between any two ports of the fabric.

Figure 2. Open Distributed Core Data Center Architecture



There are several advantages to a distributed core architecture:

- Cost-Effective: The core can be massively scaled through the use of multiple, low-cost Ethernet switches vs. traditional and expensive chassis-based systems
- High-Performance: There is full bisectional bandwidth for any-to-any communication
- Optimized for Clusters: Any host can communicate with any other host in the network at the full bandwidth of its network interface
- Ultra-Resilient: Nodes can be restarted or replaced without losing the entire switching fabric
- Control Plane Flexibility: A distributed core fabric can use standards-based Ethernet (TRILL) or IP protocols (OSPF, BGP)

Dell Force10 Z9000 Distributed Core Switch

Most core switches, typically chassis systems, are not suited to a distributed core design because they are simply too large and expensive to use in leaf-and-spine configurations at scale. The Dell Force10 Z9000 core switching system, however, is purpose-built for leaf-spine networks. The Z9000 is a 2-rack unit, 800-watt device incorporating 32 40 GbE ports (128 10GbE ports) and is available for a fraction of the cost of competing chassis-based products. The Z9000 is extremely cost-effective, especially when building very large fabrics. At scale, the Z9000 can support up to 64 spine nodes and 128 leaf nodes to create a massive 160 Tbps core in an extremely small footprint and low power budget. Network design and dimensioning are derived from these calculations:

- Number of devices in fabric: 3N/2
- Number of fabics ports: N2/2
- N = Number of switch ports per fabric node

At 128-ports of 10GbE per Z9000, maximum fabric dimensioning is as follows:

- Maximum number of devices in the fabric: 192
- Maxiumum number of fabric ports: 8192

Figure 3. Performance – The Z9000 delivers 2.5 Tbps of switching capacity and can scale to 160 Tbps in leaf-andspine architecture.

- Power – At 800 watts, the Z9000 draws 1/20 the power of competing core switches, enabling data center owners to stay well within tight power budgets even with a massively scaled core.
- Packaging – At 2RU, the Z9000 is one-tenth the size of competing core switches, making it possible to scale massively with an efficient use of space.

The Z9000 employs standards-based layer 2 and layer 3 control plane technologies. At layer 3 OSPF and BGP are used for the control plane, using ECMP for load distribution across the leaf-and-spine architecture. BGP multi-pathing can be enabled for load distribution between leaf and spine nodes. A multi-area OSPF design can be constructed to limit flooding domains and achieve routing efficiencies. Alternatively, if the fabric is architected at Layer 2, TRILL can be used for the control plane to enable multi-path support across the fabric. In either case, layer 2 or layer 3, Z9000-based distributed core architectures offer complete flexibility and control at scale.

Making Fabrics a Reality

With the Z9000, fabrics can be constructed that scale up to 160 Tbps and support as many as 24,000 10 GbE servers (assumes 3:1 over-subscription). Ultimately, however, it's the economics behind the solution that sets Z9000-based fabric solutions apart from the competition. At a fraction the cost, power consumption, and space of chassis-based systems, the Z9000 and its fabric solutions fundamentally redefine data center fabric economics. In the end, this makes fabric solutions available to a broader set of customers as the solutions become more financially viable.

Figure 4. Redefining Fabric Economics with Z9000—Scale-Out Power Consumption Comparison

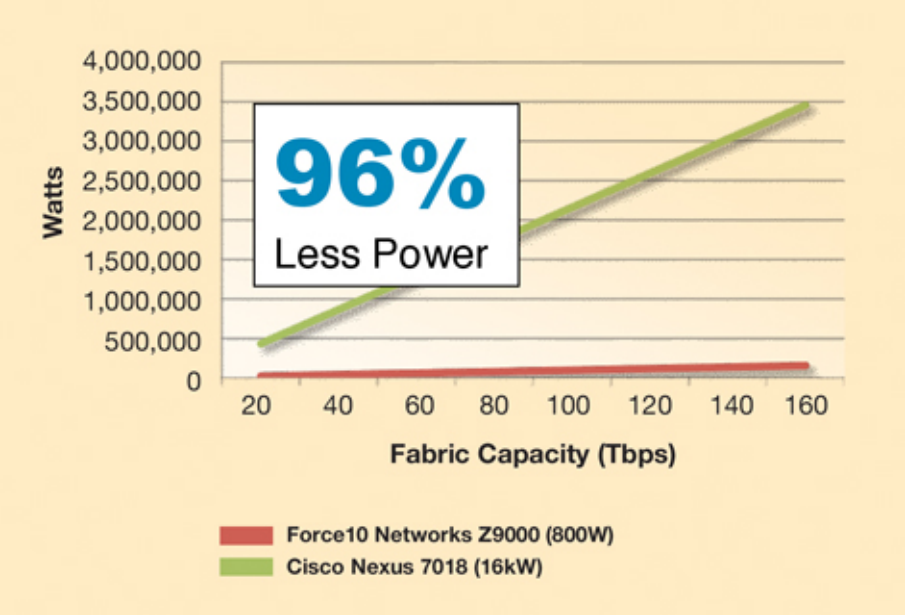
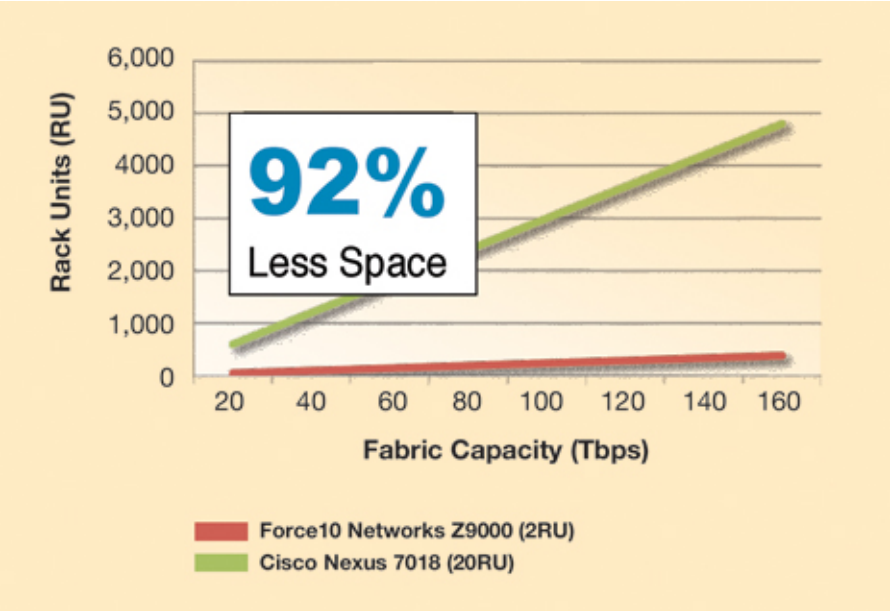


Figure 5. Redefining Fabric Economics with Z9000—Scale-Out Footprint Comparison



Summary

Distributed core architectures provide greater scalability, higher bandwidth, and higher resiliency as networking foundation for data centers that handle massive amounts of data and employ large-scale compute clusters. Dell Force10's Z9000 is the only core switch that is purpose-built for distributed core architectures and cost-optimized for scale-out fabric solutions of any size. See for yourself what the Z9000 and distributed core architecture can do for you.