HPC Deployment Scenarios

Private and public High Performance Computing systems are continually increasing in size, density, power requirements, storage, and performance. As these systems grow, the network becomes more and more important to allow full utilization of compute resources. Arista Networks has created purpose built switches designed for high performance needs, and continues to drive innovation, allowing full utilization of compute and storage resources across thousands of nodes and petabytes of storage. The following provides a guideline to the most common HPC deployment scenarios deployed in some of the largest HPC environments.

Purpose Built Hardware

The Arista 7500E Series delivers line rate non-blocking switching that enables faster and simpler network designs.

The 7500E Series offers two choices for the datacenter a 4-slot 7504 and the 8-slot 7508. The 7500E Series supports a range of interface speeds from 100Mbps up to 100Gbps Ethernet in a single system, ensuring broad choices without limiting system performance when scaling from 10G to 100G. Each 10Gb port provides over 100MB of buffer per port or over 1GB per 100G port.

The 7508E is an 11RU chassis with a 30Tbps fabric that supports up to 8 linecards and provides 1,152x10Gb ports, 288x40Gb ports, or 96x100Gb Ethernet ports in a single system - unparalleled density and performance in the industry.

The Arista 7504E provides room for 4 linecards in a compact 7RU chassis that delivers 15Tbps of bandwidth allowing up to 576x10Gb ports, 144x40Gb ports, and a massive 48x100Gb Ethernet ports.

A choice of high-density wire-speed 10Gb, 40Gb and 100Gb linecards is fully supported with the ability to mix and match any combination of modules. The 40Gb and 100Gb modules enable up to 144x10G ports per linecard. Each 40G interface can be used as either a single 40G or quad 10G Ethernet ports. The 100G interfaces can be a single port of 100Gb, three ports of 40Gb or 12 ports of 10Gb Ethernet.

With both 10Gb and 40Gb Ethernet NICs being deployed today in the storage and on the hosts, the Arista 7500E provides future proofing in the core to allow scaling for any size HPC cluster.

Figure 1: 7504E and 7508E with up to 1,152 10G ports

Business Drivers

HPC environments generally have a common set of business drivers. These drivers heavily influence the purchasing and consumption of hardware, including networking equipment. Understanding these business drivers are key to understanding the value of Arista networking equipment.

1. Efficient Compute – Every HPC environment wants to maximize their compute power with their given budget
2. Scalability – HPC environments need to scale to thousands of nodes and petabytes of data and generally require low latency node to node communication for MPI type applications
3. Large, reliable storage – Most HPC environments differ from Hadoop environments mainly in the distribution of storage and size of storage requirements. Multi-petabyte systems are common and batch processes to the nodes are common. This dictates large bandwidth requirements from nodes to storage

As these common business drivers drive the majority of purchases, networking needs are focused on how to maximize these requirements and minimize the financial impact. Arista's success in the HPC environment is mainly due to knowing these requirements and building networking equipment that increases the true compute power, allows wild scalability, and provides reliable, massive bandwidth to the storage systems.

Common Design Oversights

When designing HPC networks, customers frequently run into a few common design oversights.

1. HPC compute acquisitions are commonly based on theoretical peta-flop growth. This tends to lead to large numbers of CPU cores and minimal focus on inter-process communication (IPC) requirements that are demanded in large-scale HPC environments. While theoretical peta-flop math allows for linear growth with CPU acquisition, actual peta-flop growth depends heavily on the bandwidth between nodes and storage.
2. Many HPC environments tend to neglect managing the network. Most HPC environments spend most of their time either developing algorithms or managing the servers, and rightly so. As such, large flat layer 2 networks, undiagnosed network inefficiencies, and unknown network issues tend to creep in when the right tools and design is not setup.
3. Most HPC environments grow over time, and not in one large acquisition. However, most HPC clusters are designed with the initial acquisition in mind. Because of this, many HPC networks tend to have a well laid out initial network, then poor scale out of compute as time goes on.

Given these oversights, and many like them, Arista has found that when the best practices provided are followed, an HPC cluster can take advantage of Software Defined Networks to scale large cluster growth and low costs and drive high compute efficiencies.

Common Deployment Scenarios

Due to the major business drivers, a couple of common deployment scenarios tend to be deployed most often in HPC clusters.

The first common design deployment is a two-tier, spine-leaf architecture (Clos) that provides large scalability for the majority of HPC deployments. Many HPC networks prefer to utilize one flat layer 2 segment, or a minimal set of flat layer 2 segments with a few thousand nodes per segment. This is generally due to a historical price difference between layer 2 and layer 3 networks and known scalability limits of layer 2. Current best practices utilize the inherent multipath qualities, the broadcast limitations, and the enhanced troubleshooting tools of layer 3 in the spine and the leafs and utilize features like VXLAN to extend layer 2 when needed.

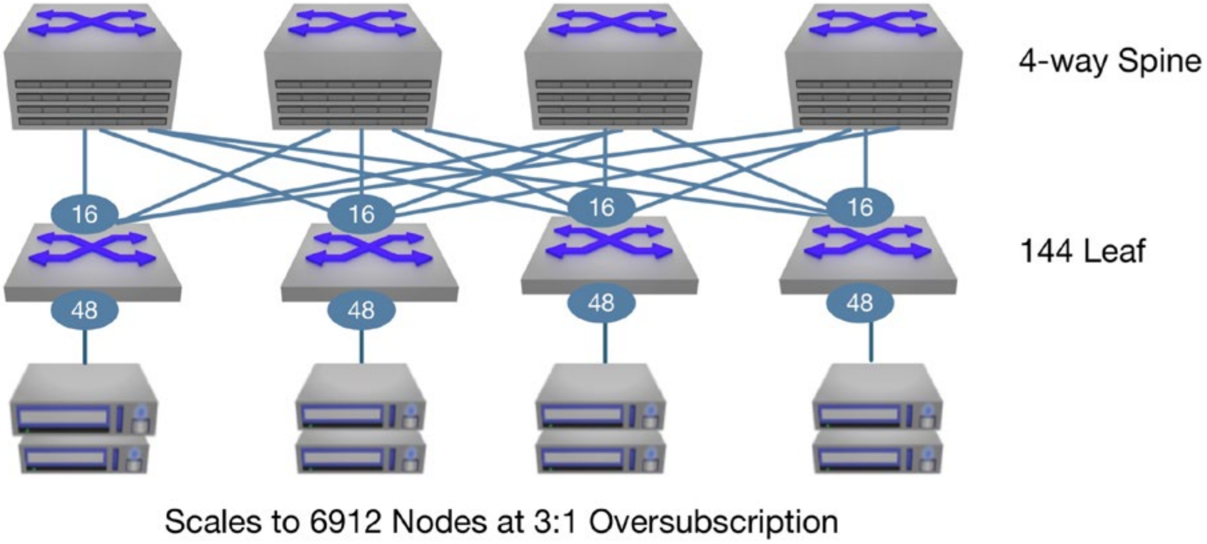

Figure 2: Two-Tier Layer 3 ECMP Example

The scale capabilities of a two-tier Clos design, along with the inter-rack bandwidth and low latency, allows for the best functionality and compute efficiency with a very low network cost per node. This structure also allows for variability and customization to the given workloads depending on the environment.

The second major design is a 3-tier, aggregation-pod architecture that allows for discrete, large financial acquisitions and scalability at the cost of limited inter-pod bandwidth. This type of design is popular where customers lease large quantities of nodes and storage once or twice a year and want to contain those purchases in a single pod.

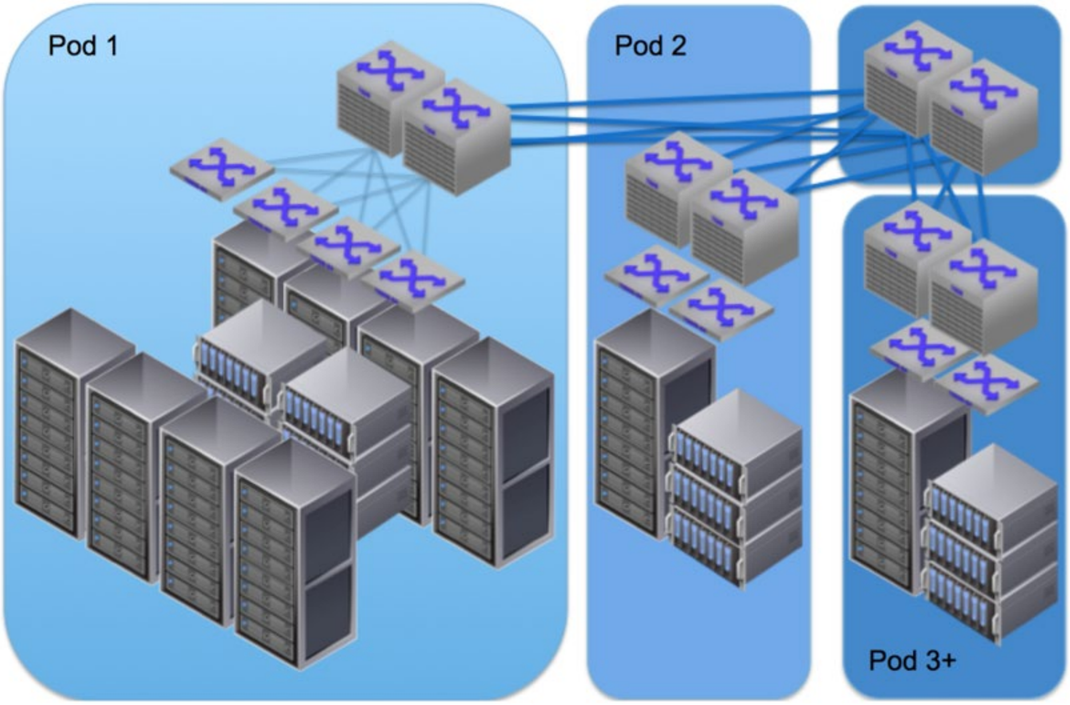

Figure 3: POD Design

This is a 3-tier design where layer 2 is either contained at the top of rack, or within a pod. The inter-pod communication is all managed with layer 3 routing using either OSPF or BGP equal cost multipathing. This allows for full utilization across multiple third tier switches, while creating high availability and redundancy. The main advantage in this design is aligning the datacenter with the business's growth pattern. When infrequent large purchases are made, or different business units want to stay separated within the DC, this design creates a well- defined separation.

HPC Key Features

Arista EOS provides key features that make an Arista switch key for low-latency, high bandwidth applications that are found in the HPC space.

CloudVision Management

Arista EOS allows for very elegant, multi-switch management using Cloudvision. Cloudvision utilizes XMPP to provide a reliable means of communication to each and every device on your network, while allowing the customer to segregate each device. While a customer HPC environment grows more and more over time, the management infrastructure and complexity stays flat and allows the customer to manage many devices at once.

ASU - Minimal Downtime for Single Attached Devices

Many HPC environments desire single connected compute nodes to minimize the cost impact of the network infrastructure. However, as such, switch downtime becomes a critical risk to mitigate and becomes a major pain point of the network infrastructure team. Arista's ASU provides a bridge to allow for switch upgrades and maintenance with very little downtime. This downtime is so small, in fact, that many HPCs can reload network devices without any noticeable impact to the hosts they're connected to. Arista's ASU provides reloads that can have layer two forwarding up in 10s of seconds and full layer 3 routing up in approximately 40 seconds. Compare this with other vendors' solutions that in the least take 5 minutes, and in the most require slow FPGA upgrades and can literally take hours to perform a 'normal' upgrade.

Deep Buffers

Deep buffers are a requirement in current HPC designs due to the massive scale required. Any time multiple hosts attempt to traverse an oversubscribed link, or they need to communicate with the same device at the same time, there is a potential for dropping packets. Deep buffers mitigate and can almost completely eliminate this problem. As seen in Figure 1, data taken from a live, large HPC environment consumed a large amount of buffer per port. Most switches on the market today use less than 9MB of buffer total, where as the Arista 7500E series has over 125MB per port available. In this specific case, a low buffer spine switch would have dropped packets on the majority of ports, causing needless network retransmits and compute slow down. Tests prove that while large buffers can cause performance issues in low speed, high latency environments, in HPC, deep buffers prove to be essential to line-rate, lossless performance.

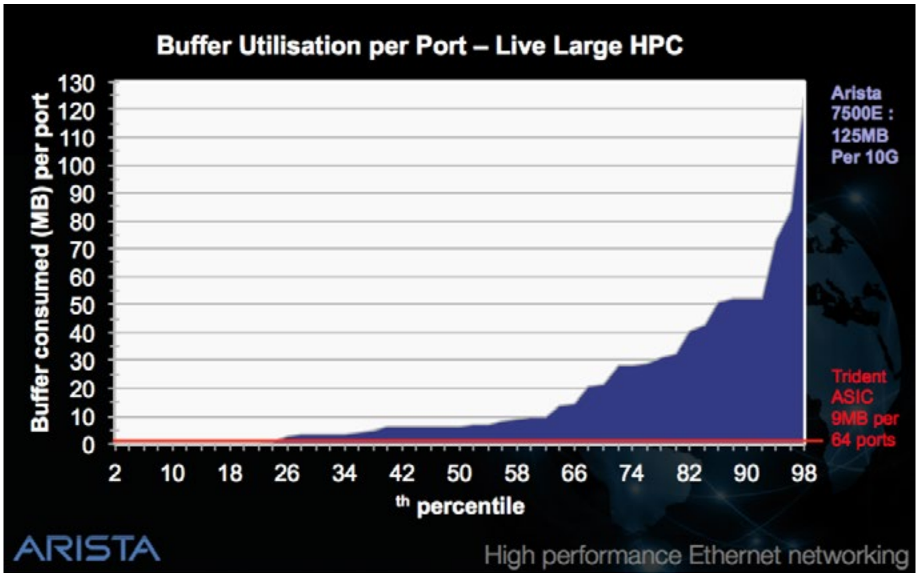

Figure 4: Buffer Utilization per Port for Live Large HPC

Conclusion

HPC environments are deployed and operate completely different from enterprise networks. Arista Networks empowers the HPC network operator to deploy high bandwidth, cost effective, scalable solutions with very high availability and designed specifically for their environment. Arista's portfolio can provide very high performance driven 10G and 40G networks while providing the tools, support, and open environment that is demanded in high performance computing networks.

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office
1390 Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A & B, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062