# Hadoop Network Design

## Network Design Considerations for Hadoop 'Big Data Clusters' and the Hadoop File System

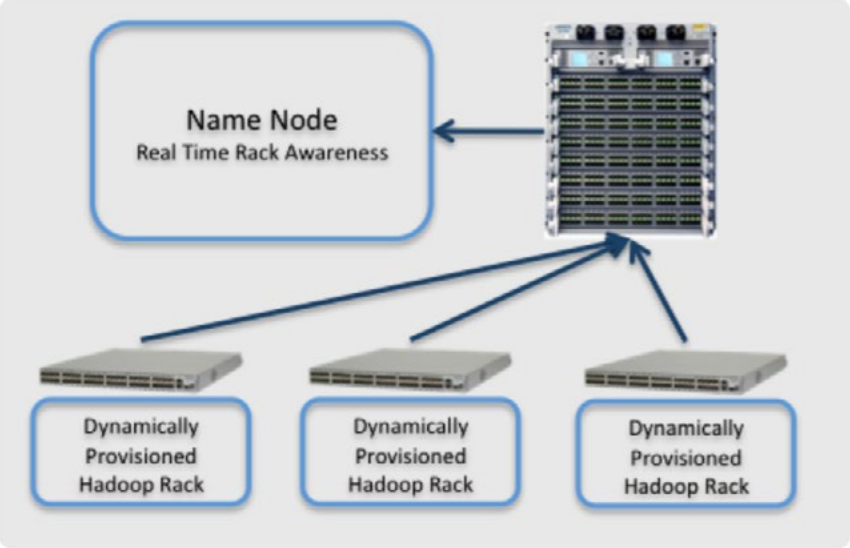
Hadoop is unique in that it has a 'rack aware' file system - it actually understands the relationship between which servers are in which cabinet and which switch supports them. With this information it is able to better distribute data and ensure that a copy of each set of data is distributed across different servers connected to different switches. This prevents any single switch failure from causing data loss.

Arista has designed the network infrastructures and implemented some of the largest and most mission critical Hadoop clusters around the world with applications ranging from:

- Web Analytics Data Mining
- Ad Serving and Targeting
- Pharmaceutical Research
- National Intelligence
- Network Security and Pattern Matching
- Retail Merchandising

### Hadoop Introduction

When the JobTracker distributes workload/computation to the servers that are storing data it tries to put the workload on the server co- located with the data to be mined. If that server is already being utilized then it sends the computation to another server in the same cabinet as the primary server with the data. This ensures that the network backbone is avoided for all bulk data movement except during data ingestion. The ability to moderately oversubscribe the network backbone goes up - so rather than wire-speed network bandwidth you can use 3:1 like the 7050 supports or 5:1 with the 7124SX, enabling a more cost effective deployment of a scale-out storage architecture.



Hadoop is also a Layer 3 aware file system - it uses IP for node addressing - this means it is routable. There is no requirement, nor benefit, at all, to building a large and flat Layer-2 network for Hadoop. You can stick with routing, building a scalable, stable, and easily supported ECMP network based on OSPF in the smaller deployments, and BGP in the larger ones and it will be stable and contain broadcasts and faults to each cabinet. Operationally these are simpler networks to troubleshoot and maintain with full toolsets available in every host stack and network element: Traceroute, Ping, Arping, fping, etc are available for L3 network day-today troubleshooting without requiring customer tool development or locking yourself into a single-vendor proprietary architecture.

**Redundancy:** Two switches at the top of each cabinet is a common enterprise recommendation - its certainly ensures that any switch failure doesn't bring any server down. In a traditional enterprise data center it is a common recommendation. However in a Hadoop cluster customers have a choice - something every other vendor, will not usually recommend: go with a single ToR switch for all cabinets except for the main cabinet that keeps the NameNode and JobTracker servers.

If the NameNode and/or JobTracker fail the job stops and the cluster fails, these two servers need to be well protected. However, once you exceed ten cabinets any single switch failure will reduce processing capacity by less than 10% and a declining percentage as the cluster scales out. The decision to use network redundancy at the leaf/access-layer becomes an operational decision around network upgrade process more than it becomes about data integrity and data availability as we increasingly trust the filesystem and application tiers to handle this responsibility.

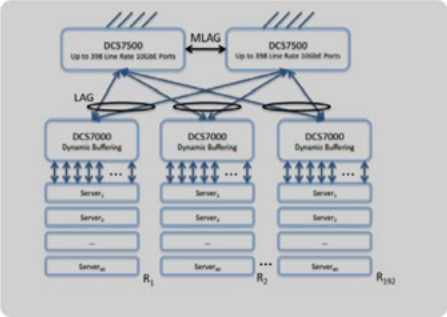
Data Integrity and Optimization: Because Hadoop is network-aware and the organization of the data structures is based, in part, on the network topology ensuring that we have an accurate mapping of network topology to servers is of paramount importance. With EOS extensibility Arista customers can load extensions that let them automatically update the Hadoop Rack Awareness configuration. This ensures several important things:

1. Data is distributed properly across servers so no single point of failure exists. Misconfiguration, or a lack of configuration, could inadvertently enable the NameNode to 'distribute' the data to three separate storage nodes that are all connected to the same switch. A switch failure then causes data loss and the data mining job stops or worse has invalid results.
2. As the Hadoop cluster scales no human has to maintain the Rack Awareness section, it is automatically updated.
3. As nodes age and are replaced or upgraded the topology self-constructs and data is automatically distributed properly.
4. Performance is improved and deterministic because no data is ever more than one network 'hop' (single MAC/IP lookup, no proprietary fabric semantics or tunneling games) from the computation that depends on that data. Jobs get distributed to the right place.

**Performance:** Hadoop performs best with a wire-speed Rack switch. Because it is L3 aware Arista recommends a switch that is wirespeed for L2 and L3 operations with no performance penalty. This helps with data ingestion which is the largest bulk data move the network has to absorb because of the Hadoop Rack Awareness architecture, but more importantly during all operational runtime it eliminates worry and simplifies trouble-shooting. If the switch is wirespeed you only have to worry about the uplinks being congested, not the switch fabric itself causing drops. This is easier to measure and report on. Avoid devices with internal oversubscription and 'router on a stick' L3 architectures that are often hard to troubleshoot and do not accurately display where and when congestion occurs.

The amount of network capacity between the leaf and spine can be decreased below 1:1 because of the Rack Awareness. This enables us to build larger networks for less cost and maps to the 7050s 3:1 (48:16) design point very well. It can also map well to 5:1 (20:4 on a 7124SX). The 7124SX is well suited for Hadoop - most of the servers are 2RU or larger servers with a high density of SAS or SATA disk in them. 1RU and Blades are almost never used - if your server vendor is recommending a bladed approach look very closely at the storage capacity per blade and the efficiency of processing your computation against the data set at scale. Many larger Hadoop clusters >1000 nodes often see CPU utilization of less than 5-10%. This is because the JobTracker cannot aggregate results and distribute computations fast enough to stress the local CPUs on the storage nodes. Increasing the storage density per compute node balances this out and delivers increased overall efficiency.

***Switching Oversubscription:*** *Often when vendors build oversubscribed products they do this for cost reasons and cut the counters that are necessary to detect congestion and packet loss. This eliminates your ability to determine if you have a network performance problem that is affecting critical data mining results.*




The spine switches should again be wirespeed to simplify troubleshooting. Decent sized buffers here are critical for stable operation and maximum application performance as they reduce retransmission during bulk data ingestion and reduce the likelihood of a drop during job distribution and computational result reporting. They are not as critical as in some other applications where the large data sets are moved to the computation and a single drop can cause a retransmission to notably slow an applications responsiveness or performance. Also as TCP is used for most of the operations involving large data movement backoff works for congestion management.

Arista recommends the Arista 7050 or the Arista 7500 for the spine/backbone of the Hadoop cluster. Because every platform supports L3 routing we can simply expand the width of the network backbone to cost- effectively horizontally scale with Arista 7050s or for a larger cluster the Arista 7500 simplifies the topology by easily supporting 4000 node and larger Hadoop clusters while providing deep buffers so packet loss and TCP retransmission do not negatively affect application performance.

**Provisioning:** Because these networks scale out quickly as customers unlock the potential of Hadoop for data mining and online storage of large data sets such as web analytics, NetFlow/sFlow analysis, RFID scanner data, clinical trials, and intelligence gathering/pattern recognition the ability to rapidly and deterministically scale out a Hadoop Cluster is of paramount importance. Arista's Zero Touch Provisioning and Zero Touch Replacement provide the mechanism to deploy a rack in less than half-an-hour, with no human intervention after cabling and powering the cabinet. The use of Chef, Puppet, or another PXE boot tool enables the switches to then auto-provision the attached servers. Arista CloudVision provides a single point of CLI management that, when coupled with ZTP/ZTR, enables us to upgrade 50% of the switches in less than five minutes without causing noticeable packet loss in an MLAG topology (when used with Hitless MLAG and MLAG ISSU). This of course requires the servers to dual-home to the switches. Once the network is operational there are some advantages to using LANZ for tracking of congestion events as they build up and determining the root cause of the congestion so it can be properly managed and ameliorated with either increased network capacity or reallocation of workload.

## Summary

Hadoop is a very interesting application that crosses lines between network, storage, and application - but this is one of the key reasons the Map Reduce function is incredibly efficient. It is also why it is incredibly disruptive to traditional enterprise vendors who are trying to re-position their legacy products to fit into this new world order.

Arista is committed to supporting Hadoop the way it was designed to operate. With EOS, the world's most advanced network operating system, we are able to integrate natively with the Hadoop operating model and ensure the network is an active participant the construction of the Hadoop topology, the monitoring of application

performance, and the efficient operation and provisioning of a large scale data mining operation.

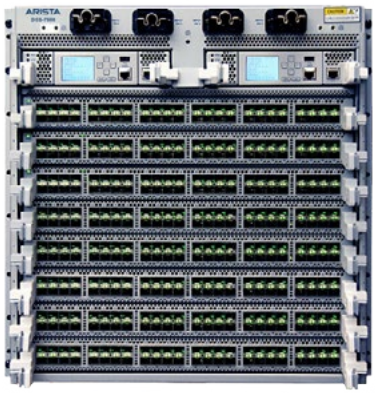
## Common Equipment Utilized

**Arista 7050 Series**

Wirespeed multi-layer data center class 52 and 64-port switches with flexible options for 10Gb SFP and triple- speed 10GBASE-T support.

**Arista 7100 Series**

Ultra-low Latency and 24-port wirespeed L2/L3 data center switching make this an optimum rack switch for up to 20 servers - a common configuration in Hadoop cluster designs. LANZ support enables real-time network performance monitoring.



**Arista 7500 Series**

Wirespeed and deep buffered multi-layer spine switch for the most demanding applications and traffic patterns. Up to 384 ports with the lowest power draw per 10Gb port in its class enables maximum compute density. Commonly used in the largest Hadoop clusters in the world scaling to hundreds of cabinets and over 10 petabytes of online storage.

**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