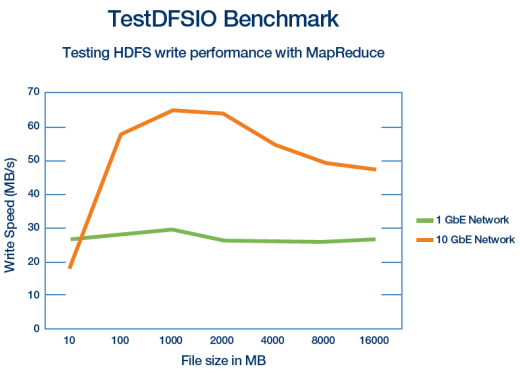
Big Data

Big Data Becoming a Common Problem

Inside

IDC projects that the digital universe will reach 40 zetabytes (ZB) by 2020, an amount that exceeds previous forecasts by 5 ZBs, resulting in a 50-fold growth from the beginning of 2010. With an ever-increasing amount of this data being unstructured it is changing the fundamental ways in which we manage and extract value from data. The term unstructured data refers to information that either does not have a pre-defined data model and/or does not fit well into relational tables. Generally it is text-heavy, but may contain other data such as dates, numbers, etc. This type of data does not fit neatly into the fields of a relational database management system (RDBMS). This data comprises what is more commonly known as big data. Gartner defines big data as high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. In the past the ability to process big data was proprietary and expensive, with few people who knew how to deal with it. The current investment in RDBMS does not meet the need or scale with the massive amount of unstructured data. Mobility, social networking and search data all comes in as unstructured and needs some form of big data analytics to help increase its value. For many this means using big data analytics on the front-end and putting the data once sorted and processed into traditional relational databases on the backend – but without some preprocessing this is not possible and large amounts of relevant information are lost.



The Problem

Big data clusters traditionally are built with commodity server hardware. In the past, 1GbE connectivity was complementary to the servers being deployed. As Moore's law predicts, the hardware processing power keeps increasing and the speed of storage I/O continues to climb while the cost of network I/O is dropping rapidly. When building a cluster you have to keep this in mind when choosing your network I/O speed. With newer switching platforms like the Arista DCS-7050T redefining the price point for 10Gb Ethernet, and 10GBASET integration on the server motherboard, 10GbE is now dramatically less on a dollar per Gb basis than 1GbE. When building out any solution remember to look at all the components to ensure they are complementary to ensure maximum performance. This is more critical now than ever as compute, storage and network I/O are all aligning from a price/performance perspective.

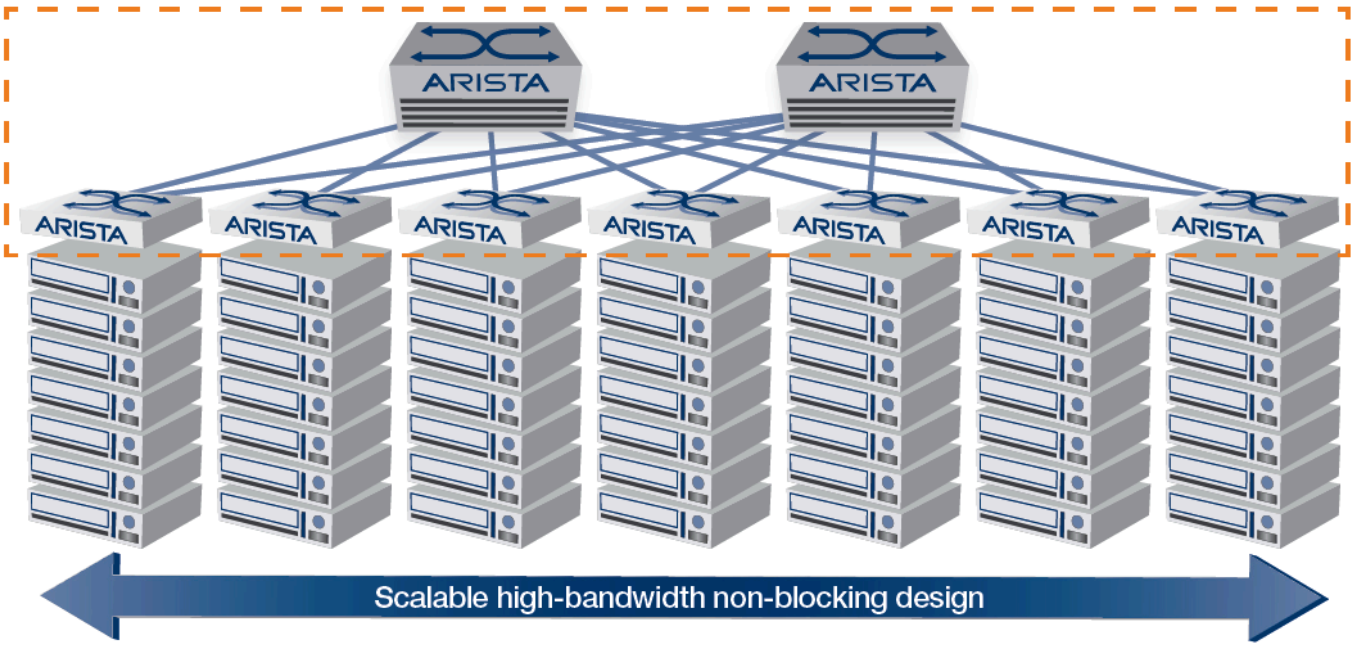
When processing a data set, moving the data and distributing the data evenly across the nodes in the cluster is the first step. This is commonly referred to as data ingression and egression, and is the most common bottleneck as data moves from the source systems for processing.

The replication process of data sets across racks results in a high level of East-West traffic. This process is constrained by the performance of the switching infrastructure and the bandwidth of the inter-switch links. For smaller deployments with one or two racks this does not present a challenge as a single switch or pair of switches can be connected back-to-back. This design works well for most deployments without having to be concerned with too much traffic overwhelming the infrastructure. As the size of the data sets being pushed into the big data clusters continues to climb, bandwidth can become a scarce resource. When transmitting multi-gigabyte data sets, nodes, even on the same switch, can quickly saturate 1GbE link capacity. This problem is exacerbated when the traffic is between racks, as the available bandwidth is generally less, and in some cases transfer requests may be delayed or dropped.

As a cluster scales, the number of jobs grows and the size of the data sets increases. Many find that 1GbE node connectivity requires large buffering at the switch or alternatively 10GBASET provides the necessary headroom to ensure optimal performance. In short, it is critical to ensure you factor in processor time, memory, disk I/O subsystem and network bandwidth when building a cluster. If any of these are not in balance, the overall performance of the cluster will suffer.

Building the Optimal Gig Data Network Infrastructure - The Two-Tier Model

Today's large-scale big data clusters perform optimally with 10Gbps wire speed top of rack switches, and single hop connections from the top of rack switch to the aggregation layer. Arista refers to this a leaf/spine design, or a two tier networking design. Spines forward traffic along optimal paths between nodes at Layer 2 or Layer 3 while leafs control the flow of traffic between servers. Cross sectional interconnect bandwidth can be improved though adding additional connections between the leafspine layers taking advantage of L3 ECMP.



This two-tiered Leaf-Spine architecture allows connections scaling to 50,000+ end user ports nodes providing maximum throughput for your data sets. At the spine, routing between the leaf switches that have the highest traffic exchange is desired. At the leaf, line rate performance enabling scale-out application deployments is highly desirable. This scaled out high bandwidth design allows for optimizing the performance of a job by minimizing ingestion/egestion and RPC retransmissions.

The Spine Layer

The diagram below provides an architectural view of how to deploy Apple TV, with a centralized mobility controller, and with SDN enabled switches. The tunneling of the set-up services (control) is shown in purple, the determination of the forwarding path is shown in dotted green lines, and the actual forwarding path is shown in solid green lines.

Recommended Platforms and Modules

When building a large-scale cluster, for designs of greater than 1000 compute nodes the Arista 7500 is recommended. The line rate switching capacity, unmatched switch fabric bandwidth, large per port buffers, and industry leading 10 / 40GbE density make the Arista 7500 the perfect choice to provide maximum scale at the spine layer. In designs that call for less than 1000 compute nodes, the Arista 7050 provides line rate performance in a compact low power footprint.

- The Arista 7500 will allow big data clusters to scale beyond 55,000 compute nodes.
- Big data clusters indicate peaks in buffer utilization in excess of 40MB

Making a design decision requires factoring in multiple components. Arista leverages our experience with the smallest to the largest big data clusters in the world when determining the optimal design and oversubscription. The table below shows platform scale based upon this experience.

	Arista 7508E	Arista 7050
Maximum Scale	55,0000+ compute nodes	2,000 compute nodes
Per port buffering	128MB	5MB (dynamic)
Fabric capacity	30Tbps	1.28Tbps
Forwarding capacity	5.7Bpps	960Mpps
Line-rate 10GbE ports	1152	64
Size	11RU	1RU
Power consumption per 10GbE port	3.5W	2W

If maximum performance and scale is the end goal then the Arista 7500 series is the clear recommendation. If minimal footprint, power and cooling are the driving factors then the Arista 7050 should be the platform of choice.

The Arista 7050T is a wire rate, low power consumption, 1RU wire speed 10GBASETswitching platform. This platform supports 100M/1G/10G over most existing cable plants and is available with 36, 52 or 64 ports, providing flexibility based on the connectivity requirements of the rack. The 7050T provides a cost-effective, high performance switch and is rapidly becoming the de facto standard when building large big data clusters.

Buffering and Performance

The design of the network fabric has a major effect on application performance, and switch buffers can have a major impact in this regard for big data clusters. When looking at buffering requirements there are three key concepts one must consider.

- Throughput – How much bandwidth is available per switch and per server
- Fairness – How throughput is allocated to multiple flows at same level of service
- Good-put – Useful data carried, excluding packet drops and retransmissions

Packets will drop when buffers are exhausted, causing retransmissions and making throughput suffer, thus turning 'good-put' into 'bad-put'. Observations made in realworld large-scale (defined as greater than 1K nodes) big data clusters indicate peaks in buffer utilization in excess of 40MB. As a rule of thumb the network buffer per attached server should equal 1-2MB, calculated by the formula (number of flows per server) * (KB per flow).

Arista LANZ can provide export queuing information in real-time with event driven notification on a per-port basis defined by a configurable queue-depth.

Big data can benefit from deep buffering at critical points within the switching infrastructure. Since big data clusters can often oversubscribe links, performance can dramatically improve with deep buffering rather than dropping and retransmitting packets. As the number of flows increases, buffering becomes a critical design point. The large buffers of the Arista 7500 have been designed for fairness and optimal handling of high volume traffic flows.

Summary

Big data provides a very interesting application problem that crosses the lines between network, storage, and application. Big data can provide tremendous insight and tremendous business value. A properly architected big data cluster can provide disruptive and important knowledge to traditional enterprise customers. However, applying legacy architectural principles with massive oversubscription and minimal buffering will degrade the overall big data cluster and in turn reduce the cluster's effectiveness.

Arista is committed to supporting big data clusters in the way they were designed to operate with a non-blocking, deep buffered, high-speed data center network. This coupled with Arista's EOS, the world's most advanced network operating system, allows best-in-class native integration with popular big data distributions such as Hadoop.

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