Networking in the Hadoop Cluster

Hadoop and other distributed systems are increasingly the solution of choice for next generation data volumes. A high capacity, any to any, easily manageable networking layer is critical for peak Hadoop performance.

Data analytics has become a key element of the business decision process over the last decade, and the ability to process unprecedented volumes of data a consequent deliverable and differentiator in the information economy. Classic systems based on relational databases and expensive hardware, while still useful for some applications, are increasingly unattractive compared to the scalability, economics, processing power and availability offered by today's network driven distributed solutions. The perhaps most popular of these next generation systems is Hadoop, a software framework that drives the compute engines in data centers from IBM to Facebook. .

Hadoop and the related Hadoop Distributed File System (HDFS) form an open source framework that allows clusters of commodity hardware servers to run parallelized, data intensive workloads. Actual clusters include shoe string research analytics to thirty petabyte data warehouses, and applications range from the most advanced machine learning algorithms to distributed databases and transcoding farms. Given sufficient storage, processing, and networking resources the possibilities are nearly endless.

HDFS

The Hadoop Distributed File System (HDFS) stores multiple copies of data in 64MB chunks throughout the system for fault tolerance and improved availability. File location is tracked by the Hadoop NameNode. Replication is increased relative to frequency of use, and a number of other tunable parameters and features such as RAID can be used depending on the application. Because replication is accomplished node to node rather than top down, a well architected Hadoop cluster needs to be able to handle significant any to any traffic loads.

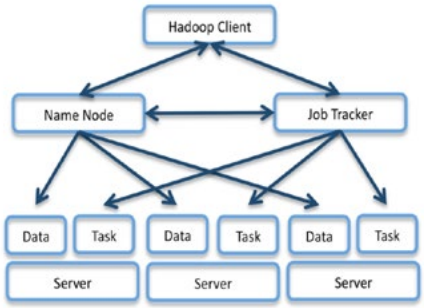

Figure 1: Hadoop Architecture Distributes Storage and Computation to the Cluster

Parallelization and Pushing The Computation to the Data

The Hadoop JobTracker breaks down large problems into smaller computational tasks assigned to servers in the cluster. In order to handle large data sets, servers are given tasks relevant to the data already present in their directly attached storage (DAS). This is often referred to as pushing the computation to the data, and is a critical part of processing petabytes - even with 100 GbE, a badly allocated workload could take weeks to simply read in all the data necessary! Finally, rack awareness allows the JobTracker to assign servers close to the data in the network topology if no directly attached server is available.

How the MapReduce Algorithm Works

MapReduce is the algorithm originally used in Google's massively parallel web ranking systems and forms the cornerstone of the Hadoop system. It is composed of two steps: Map and Reduce.

Map

Mapping refers to the process of breaking a large file into manageable chunks that can be processed in parallel. In data warehousing applications where many types of analysis are conducted on the same data set, these chunks may have already been formed and distributed across the cluster. However, for many processes involving changing data or one time analyses, the entire multi-terabyte to multi-petabyte workload must be efficiently transferred from storage to the cluster members on a per case basis - Facebook's larger clusters often intake 2PB per day. In these situations a high capacity network is critical to time-sensitive analytics.

Once the data has been distributed throughout the cluster, each of the servers processes the data into an intermediate format paired to a "key" which determines where it will be sent next for processing. A very simple example of this might be mapping each of the works of Monty Python to a separate server which will count how many times any word appears in that particular text.

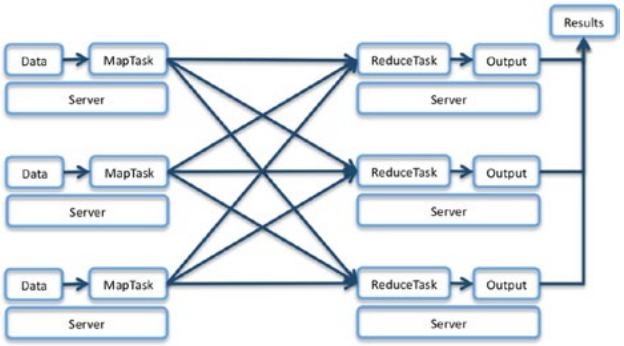

Figure 2: Data flows without persistence from Map to Reduce and requires complete, any to any network topologies.

Reduce

When the Mapping Servers have completed their tasks, they send the intermediate data to the appropriate Reduce Server based on the data key. While many tasks have significant compression after the Mapping calculations are completed, other analyses such as the sorting used in descriptive statistics require almost the entire data set to be re-allocated, or "shuffled" to the Reduce Servers. At this point the data network is the critical path, and its performance and latency directly impact the shuffle phase of a data set reduction. High-speed, non-blocking network switches ensure that the Hadoop cluster is running at peak efficiency. .

The Reduce Servers integrate the data received from Map servers and create an aggregate result per key that can be either reported directly or used for further analysis. To continue with our previous example, each Map Server would have by now sent the intermediate results of frequency keyed by word to the appropriate Reduce Servers. The Reduce Servers can thus perform a number of analytics such as calculating the aggregate sum of any word that Python wrote, or perhaps classifying which work had the greatest ratio of 'Spam' to 'Cheese'. Reductions are the final step before useful information can be extracted, but in many cases also generate entirely new data sets that can be fed back into the Hadoop cluster for further insight.

Impact of Network Designs on Hadoop Cluster Performance

A network designed for Hadoop applications, rather than standard enterprise applications, can make a big difference in the performance of the cluster.

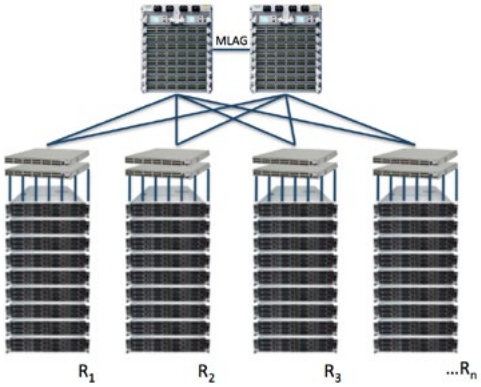

Figure 3: A high performance, any to any network architecture is critical to optimal Hadoop cluster performance.

High Capacity, Any-to-Any Topology, and Incremental Scalability

Getting data into the cluster can be the first bottleneck, and whether replicating or shuffling data, Hadoop requires significant any to any node traffic to get its job done. In order to efficiently access stored results or simply calculate new ones, a well-provisioned network with full any to any capacity, low latency, and high bandwidth can significantly improve Hadoop cluster performance. Finally, as workloads grow, it is important that the network can sustain the inclusion of additional servers in an incremental fashion - Hadoop only scales in proportion to the compute resources networked together at any time.

High Availability and Fault Tolerance

While Hadoop has self-contained fault tolerance in any single node, a failure in network connectivity in any but the largest clusters will halt both HDFS data accessibility and any currently running jobs in their tracks. Highly available, fault-tolerant networking equipment can make sure that the Hadoop cluster stays running, and furthermore assist in a quick re-provisioning of any failed server nodes during execution.

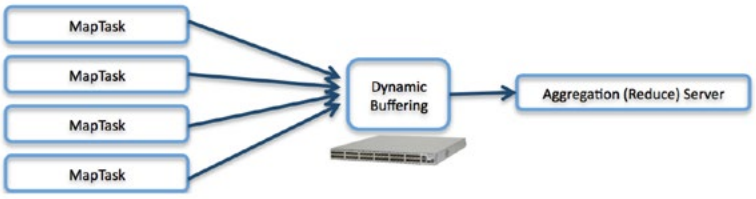

Figure 4: Hadoop's data shuffle between Map and Reduce creates unavoidable fan in when multiple Map servers must stream results to a single Reduce server.

Dynamic Buffers and Visibility

Traffic fan in is an unavoidable fact of aggregation. Networks employing dynamically allocated buffers can shift resources to congested ports in real time for superior adaptability in the face of rapidly changing traffic workloads. If dynamically allocated network buffers are employed, even the most oversubscribed Reduce server can receive all its intermediate data without lost packets and the consequent network overload and inefficiency that would otherwise occur. Finally, tools such as Arista's Latency Analyzer allow network introspection and cluster reconfiguration to eliminate bottlenecks and create workload dependent cluster optimization.

Management And Extensibility

Getting the most out of any scaled solution requires proper management tools and a framework for customized application needs. Because the EOS Extensible Operating System is based on Linux, EOS provides the perfect foundation for leveraging open source tools and creating user defined functionality.

Admins gain immediate productivity with their preferred binaries and scripts without needing to learn and relearn proprietary operating systems. Cluster management systems such as perfSONAR, gmond, Nagios or Ganglia can be run directly on the switch to detect and proactively address any unexpected data center conditions - possible responses range from email and SMS alerts to actions immediately shifting topology and configuration. EOS allows the full power of open source to be leveraged for a smoothly running cluster.

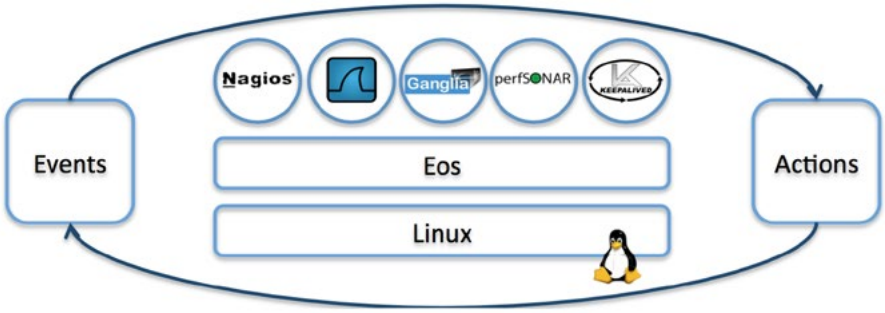

Figure 5: Linux and open source tools can be combined for event driven reactivity and visibility in the next generation network.

Furthermore, anyone who has managed large installations knows that data center class automation begins the moment hardware meets rack. Arista's Zero Touch Provisioning allows bare metal switches to be instantly configured and operational in the network, and EOS can even use PXE to provision servers parametizably by VLAN or optionId. Finally, dynamic topology detection can automate modification to the Hadoop XML files which control location aware programming - a daunting task when separate teams install hardware and write software, and one which is critical to efficiently pushing computation to the correct server. Ultimately, Arista's powerful tools for automation significantly reduce the overhead of large cluster management, allowing IT staff to focus on meeting and exceeding actual business deliverables.

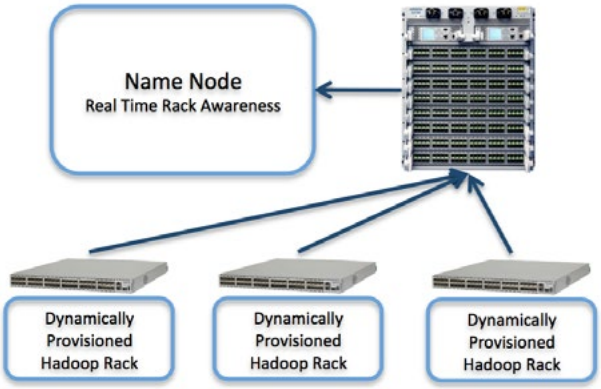

Figure 6: Arista's Zero Touch Provisioning and extensibility create dynamically provisioned deployment and real time rack awareness in Hadoop clusters.

Key Take Aways

Hadoop and other distributed solutions are managing data sets of unprecedented and still growing scale. Business intelligence and other forms of analysis will increasingly rely on frameworks such as these whose peak performance is achieved with high capacity, any to any, easily manageable network technologies. Arista Networks is committed to delivering stable, high-performance solutions from the silicon up for Hadoop and any other demanding workloads of the next generation data center.

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