SGI® Hadoop Solutions

Solutions for Data Analysis on Large Clusters

Hadoop is a framework for building and deploying data storage and data analysis systems, using large distributed clusters. Hadoop is ideal for large amounts of data that can be easily decomposed. SGI has deployed tens of thousands of Hadoop servers on several SGI system architectures including its Rackable™, CloudRack™ C2, and ICE® servers. SGI supplies systems and services to optimize Hadoop clusters for performance, density, power utilization and cost. SGI assists customers to architect Hadoop solutions based on their unique data and analytics requirements as well as delivering pre-configured reference implementations to speed customers' time to production.

Introduction

Hadoop implements a computational paradigm named MapReduce where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. The Hadoop framework transparently provides applications both reliability and data motion. In addition, it provides a distributed file system that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster.

MapReduce

MapReduce is a framework for processing highly distributable problems across huge datasets using a large number of computers (nodes). Computational processing can occur on either unstructured or structured data. MapReduce consists of two steps:

A "map" step: The master node takes the input, partitions it into smaller tasks, and assigns them to distributed compute nodes. An assigned node may do this again in turn, leading to a multi-level structure. Each node processes its assigned tasks, and passes the answer back to its assigning node.

A "reduce" step: The master node then collects the answers to all the tasks and combines them in some way to form the output – the answer to the problem it was originally trying to solve.

MapReduce directly allows for distributed processing. Provided each mapping operation is independent of the others, all maps can be performed in parallel – though in practice it is limited by the number of independent data sources and/or the number of CPUs near each source.

Similarly, a set of 'reducers' can perform the reduction phase - provided all outputs of the map operation that share the same key are presented to the same reducer at the same time. While this process can often appear inefficient compared to algorithms that are more sequential, MapReduce can be applied to significantly larger datasets than standard servers can handle – a large cluster can use MapReduce to sort a petabyte or more of data in only a few hours. The parallelism also provides for recovery from partial failure of servers or storage during the operation: if one mapper or reducer fails, the work can be rescheduled – assuming the input data is still available.

Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS) stores large files across multiple machines. It achieves reliability by replicating the data across multiple hosts and does not require RAID storage on hosts. With the default replication value of 3, data is stored on three nodes: two on the same rack, and one on a different rack.

The filesystem is built from a cluster of data nodes, each of which serves up blocks of data over the network using a block protocol specific to HDFS. They also serve the data over HTTP, allowing access to all content from a web browser or other client. Data nodes can talk to each other to rebalance data, to move copies around, and to implement the replication of data.

SGI® Hadoop Solutions

Hadoop Resiliency

Both MapReduce and the distributed file system are designed so that node failures are automatically handled by the framework. With a large number of servers in a Hadoop cluster, it is assumed that individual nodes will fail. The Hadoop infrastructure recognizes this and the overall design accommodates individual node failures.

The filesystem requires one unique server, the name node. This is a single point of failure for an HDFS installation. If the name node goes down, the filesystem is offline. When it comes back up, the name node must replay all outstanding operations. This replay process can take over half an hour for a large cluster. To address the exposure of this single point of failure, an HDFS filesystem generally includes what is called a secondary name node, which regularly builds snapshots of the name node's directory information. These snapshots can be used to restart the name node after a failure without having to replay the entire journal of the filesystem.

For additional information on Hadoop, a great source is Wikipedia, http://en.wikipedia.org/wiki/Hadoop on which much of the information in the previous sections is based.

Solution Components

The most important components in a Hadoop solution are the servers, their local storage, the networking infrastructure and the software: middleware and analytics applications. Hadoop clusters can be large in both size and complexity so the design choices made can significantly impact performance and cost. SGI has many customers who have deployed SGI Hadoop clusters over many years, and has significant experience in terms of optimal motherboard and CPU selection, power supplies and fans, storage selection, and the ideal rack and cooling solutions for a given set of customer requirements.

SGI Cluster Systems for Hadoop

The number, type and configuration of the servers in a Hadoop cluster is an important decision. There are a number of considerations and here the main focus is on the data nodes.

Hadoop applications can vary in terms of their needs for I/O, memory and CPU resources. Because of this, the guidelines below can only be general, with specific adjustments needed for each application.

SGI has deployed Hadoop clusters on several architectures including its Rackable™, CloudRack™ C2, and ICE servers. We review each of these and the unique capabilities that each brings to a Hadoop cluster.

Rackable servers are rack mount servers that provide industry-leading density, power-savings and flexibility. They can be deployed in either back-to-back or flow-through configurations, using either traditional AC or rack-level DC power. With the ability to draw from a wide variety of commodity components, the Rackable line allows a customer multiple configuration axes, not limited only to number and size of memory DIMMs, CPUs, and disk drives, but also allowing the choice of different motherboard layouts and capabilities. Different motherboard options can give a Hadoop cluster different optimization points in terms of performance versus power savings and other parameters such as manageability.

The CloudRack C2 server extends the level of flexibility and power savings available in the Rackable line by implementing a tray design that creates a 'breadboard' over which different combinations and layouts of motherboards and disk drives can be arranged. CloudRack allows for the largest disk to motherboard ratios available in the industry today. Cooling is provided at the rack level with large, efficient fans. While the DC variant provides rack level rectification, the new AC version provides power supplies at the tray level.

The SGI ICE blade-based system is designed around tightly integrated Infiniband in one or two planes: two completely independent networks with QDR data speeds. With an Infiniband backplane, each SGI ICE blade enclosure minimizes cable complexity. As with CloudRack, larger, more efficient fans and power supplies replace smaller less efficient ones local to the node. By using one Infiniband plane for storage traffic, the system eliminates the need for local disk drives that add cost, space and reliability challenges in favor of one or more fast and reliable external storage arrays. SGI ICE is chosen for Hadoop among those customers who want to deliver high performance in analytics applications, that are network bandwidth limited, or customers that have existing data stores in Infiniband network infrastructures.

SGI® Hadoop Solutions

Server Configurations

Ideal server configurations for Hadoop cluster nodes vary widely but, generally, it is recommended that at least six, and preferably more, drives be deployed with a PCI-based HBA for performance. SGI delivers local storage drives with a very wide range of parameters: form factor, protocol, transfer rate, rotational speed, and size. Different sizes and capacities of SATA drives, for example, can be mixed and matched for ideal combinations of performance, capacity, cost and power savings. Generally 7200 RPM SATA drive are the best choice for price/performance.

In terms of memory capacity, it is recommended that at least 24 to 48 GB per server be deployed, with the number of DIMMs and sizes correctly balanced for either the four memory channel case with AMD or the three channel case with Intel.

Initial Hadoop configurations were deployed using two socket servers, 12 to 24 GB of memory per server, and 4 500GB SATA drives. As Hadoop cluster options have matured, the trend is now toward larger memory configurations based on the application workload and disk drive configuration.

SGI has installed Hadoop clusters with both AMD and Intel processors. The new AMD Opteron 6100 processor, with up to 12 cores per socket can be an ideal option for a Hadoop server. Below is a recommended configuration using the Rackable C2005 half-depth chassis which, as a 2U design, optimizes airflow, but when used in a back-to-back configuration has similar densities to competing 1U designs. The design using the Tyan 8230 twin-socket motherboard (SGI "G1") which has an optimized layout allowing for a good I/O and DIMM combination:

Component	Quantity	Selection
Chassis	1	Rackable C2005
Motherboard	1	Tyan 8230 (G1)
CPUs	2 x 1.8 GHz, 65W	AMD Opteron 6124 HE
DIMMs	8 x 8 GB	DDR3 registered, 1333 Mhz
HDD	2 x 300 GB	SATA, 100000 RPM 2.5"
HDD	5 x 2 TB	SATA, 7200 RPM, 3.5"
RAID Card	1	Megaraid 8708EM2

For better cooling, the Tyan 8236 (SGI "G5") motherboard with a rear placement processor layout could be used, and for even better density this board plus the Rackable C1001 chassis makes a good choice.

On the Intel side, the C2005 configuration below is a recommended balanced configuration. It is based on the Tyan S7012 (SGI "TY7" designation) twin-socket motherboard. The S7012 has an ideal memory layout for the three memory channel design of the Intel Xeon 5600 processor.

Component	Quantity	Selection
Chassis	1	Rackable C2005
Motherboard	1	Tyan 7012 (TY7) SATA
CPUs	2 x 2.66 GHz, 80W	Intel Xeon 5640
DIMMs	9 x 8 GB	DDR3 registered, 1333 Mhz
HDD	10 x 1 TB	SATA, 7200 RPM 2.5"
RAID Card	1	Megaraid 8708EM2

SGI's first Hadoop Cluster Reference Implementation uses the Intel 5500WB board (SGI "TY6" designation), giving access to Intel Power Node Manager (IPNM) capabilities that are important in managing power. Similar configurations have been installed for a number of SGI customers. The IPNM technology when combined with the SGI Management Center Power Option gives these customers the ability to manage system power consumption at a very fine level, within a data center power envelope, and/or allows them to tune Hadoop-based analytics application performance versus power consumption to arrive at the maximum operations per watt for a given application. The TY6 board balances cost, power usage and features. It has a slimmer memory footprint than many other boards, reducing cost and improving airflow at the cost of memory capacity and performance. All of the C2005 configurations support the addition of multiple PCI cards for specific Hadoop nodes (name, data, etc.) such as 10GigE, Quad NID, and Fibre Channel.

SGI ICE blades are available with either Intel or AMD processors. Available DIMM slots vary between 12 for the Intel blades and 16 for the AMD blades, which are the proper number for ideal balancing of the three or four channel memory footprints. Some minimum storage may be available local to the node, but most Hadoop installations with SGI ICE use higher capacity, higher-performance storage that can be directly accessed via the many Infiniband ports available in an SGI ICE system.

SGI® Hadoop Solutions

Networking

When local drives reside on each node, as in Rackable or CloudRack Hadoop installations, Gigabit Ethernet tends to provide sufficient network bandwidth and latency. If two NICs are available per board, they can be bonded together to increase bandwidth. Some installations have chosen 10GigE as the networking interconnect for higher bandwidth over GigE.

For SGI ICE installations, sufficient bandwidth is supplied via the Infiniband backbone. SGI tests have shown that similar results are attained for applications accessing high performance disks via an Infiniband network versus lower performance disks local to the node. Administrative traffic is separated from application traffic via an additional GigE-based administrative network.

Software

SGI has delivered Hadoop clusters for use with customer installed software and complete factory installed, pre-integrated Hadoop clusters.

Hadoop is available as a set of open-source software components, downloadable from the http://hadoop.apache.org website. The 0.21.0 version or later of Hadoop is recommended with various feature enhancements and bug fixes over previous versions.

Also, a number of organizations and vendors provide distributions of the Hadoop software, and support and services. Cloudera Inc., is the leading provider of Apache Hadoop-based data management software and services. SGI is a member of the Cloudera Connect Partner Program and resells and provides support for Cloudera software and services. SGI distributes Cloudera software factory-installed on SGI® Hadoop Clusters. The relationship enables the two companies to jointly build, sell and deploy complete, end-to-end solutions for enterprises deploying Apache Hadoop in performance-intensive environments and/or that need minimal time to production. SGI's Hadoop Cluster Reference Implementation is a Cloudera Certified Technology.

SGI Management Center™ is the recommended cluster manager for Hadoop installations, with an industry-leading set of features. SGI Management Center Power Option optimizes power consumption for specific Hadoop installations using Intel Intelligent Power Node Manager firmware.

Based on SGI's relationships with key business intelligence (BI) software vendors, SGI's Hadoop Cluster Reference Implementation also delivers a tested, optimized, ready-to-run Hadoop system with an ecosystem of analytical solutions which allow developers to more easily construct best-in-class, BI solutions. SGI partners with Kitenga, Datameer, Pentaho and Quantum4D to deliver these differentiated analytical capabilities to customers in federal, financial, social media, telecommunications, and other key markets.

SGI and Kitenga offer a new generation of Big Data insight engine with integrated search, information modeling and visualization capabilities. SGI's and Datameer's relationship provides a business intelligence platform for Hadoop with data integration, spreadsheet UI for analytics, and data visualization. This enables business users to access, analyze and visualize massive amounts data on dashboards. SGI also works with Pentaho, providing Pentaho Business Analytics, leveraging a graphical ETL environment for creating and managing Hadoop MapReduce jobs. This easily integrates data from other sources and provides end-to-end business analytics on Hadoop, including reporting, ad hoc query, interactive analysis and data integration. With Quantum4D, SGI offers data modeling and interactive data visualization capabilities on Hadoop for actionable insight.

SGI Business Intelligence Solutions

Hadoop is ideal for large amounts of data that can be easily decomposed. When the data is not easily decomposed, a large shared memory server like the SGI® UV can be used as part of a Hadoop compute complex to copy large volumes of data into memory at one time. With a very low latency interconnect, the UV can work on large datasets in real-time, and is used for applications such as fraud detection and security analytics.

Rely on the Experts

SGI Professional Services has deployed many large Hadoop clusters, including one on the Top 500 (http://top500.org/). SGI's worldwide team of consultants averages more than 20 years of technical computing experience and can integrate, configure and test multi-component Hadoop solutions at the customer site or in the factory for in-factory acceptances. Fixed-cost or customized Hadoop Implementation services are available worldwide.

In addition, SGI Customer Services offers flexible Support Programs with worldwide reach. Mix and match coverage and response times to customize the right amount of protection. The SGI service organization is committed to excellence and consistently attains high ratings from customer satisfaction surveys.

SGI Hadoop Solutions

With the flexibility of many different server options across multiple product lines, SGI offers the most extensive Hadoop solutions in the industry, backed up by world-class support and services to keep your Hadoop solution running in real-time.

©2011 SGI. SGI and OpenFOAM are registered trademarks of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries. All other trademarks are property of their respective holders. Specifications subject to change without notice. 27112011 4307