vSMP Foundation for Big Data and Analytics

The "Big Data" term is generally used to describe datasets that are too large or complex to be analyzed with standard database management systems. When a dataset is considered to be a "Big Data" is a moving target, since the amount of data created each year grows, as do the tools (software) and hardware to make sense of the information. Many use the terms volume (amount of data), velocity (speed of data in and out) and variety of data to describe "Big Data".

vSMP Foundation™ from ScaleMP can help to address some of the challenges for organizations that need to deal with these very large amounts of data. Large datasets can be analyzed and interpreted in two ways:

- **Distributed Processing** – use many separate (thin) computers, where each system analyzes a portion of the data. This method is sometimes called **scale-out** or **horizontal scaling**.
- **Shared Memory Processing** – use large systems with enough resources to analyze huge amounts of the data. This method is sometimes called **scale-up** or **vertical scaling**.

Distributed processing

Depending on the type of data to be analyzed and the desired outcome will determine where a distributed system would be the best fit for an organization. Simple searches through a list of records could be easily distributed to a set of systems, with the results from each server then collected. An example of this could be searching through all driver license records for those with blue eyes. A portion of the driver license records could be kept on each server without issues such as overlap or dependency on other server results. Apache Hadoop is an open source software framework that supports data-intensive distributed applications. It enables applications to work with hundreds to thousands of computational independent computers and petabytes of data. Hadoop was derived from Google's MapReduce and Google File System (GFS) papers.

Advantages and disadvantages of distributed processing

The main advantage of distributed processing is its ability to scale just by adding "one more node". On the other hand it requires certain skillsets and management capabilities to manage Hadoop cluster which is needed to set up the software on multiple systems, and keep it tuned and running. It also worth noting that Hadoop is suitable for cases where data interdependency is low and requires small (if any) data replication.

Shared Memory processing

If the amount of data is complex, unstructured, or where multiple algorithms are required to be used on the data, a large shared-memory system would be best. Much more of the data could be held in the memory of the system, and different processes could all operate on the same data, while the data resides in memory. For instance, monitoring thousands of video feeds to determine any correlation between the images would benefit from keeping all the feeds in main memory and having multiple applications all work with significant sections of the data.

Advantages and disadvantages of shared-memory processing

While it is much easier to manage single large-scale system and host all the data and processing on one machine, such systems tend to be quite expensive. The reduction in OPEX as result of single system to manage and a reduction in DBA complexity come at the cost of the hardware.

ScaleMP™
Creating the Power of One

The High-end Virtualization Company

ScaleMP vSMP Foundation

vSMP Foundation from ScaleMP creates a virtual shared-memory system from a distributed infrastructure, providing the best of both worlds for big-data and analytics problems. On one hand, it allows scale just by adding "one more node" but still keeps the OPEX advantage of a shared-memory system. It provides benefits for small Hadoop deployments where the OPEX costs are high, and can handle big-data cases where data cannot be easily distributed by providing a shared-memory processing environment.

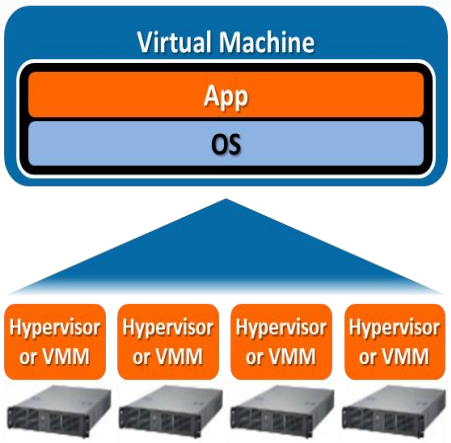

Figure 1: ScaleMP vSMP Foundation

How does it work?

vSMP Foundation creates a single virtual machine with CPUs, RAM and I/O aggregated from several smaller systems. This allows for more data to be held in memory, directly accessible by any of the CPUs in the aggregated system.

Advantages and disadvantages of vSMP Foundation

With complex datasets that require a number of steps of processing and significant computations, the ability to hold larger datasets in memory, without having to swap to disk, greatly reduces the time to understanding trends within the data.

Smaller, more distributed types of analyses can also benefit from vSMP Foundation. For example, if a distributed algorithm is used, a certain number of servers must be maintained and administered. By using vSMP Foundation on a small to medium sized cluster, this greatly simplifies the administration cost, while the application will run at the same performance. By reducing the administration costs associated with a small cluster running a distributed algorithm, an increased ROI can be achieved, while continuing to run familiar big data analysis applications.

Summary

With the increasing business requirements to analyze and understand significant volumes of data, it is important to create an IT system than can respond to those needs. By using vSMP Foundation to combine the low cost of scale-out systems with the advantages of scale-up systems, cost savings can be realized while maintaining a highly responsive system to make sense of Big Data information.

	Distributed Processing	Shared Memory Processing	vSMP Foundation
Advantages	Low cost infrastructure (CAPEX) with pay as you grow characteristics	Single system to manage (OPEX)	• Low cost infrastructure (CAPEX) with pay as you grow characteristics • Single system to manage (OPEX)
Disadvantages	Management cost (OPEX)	Platform cost (CAPEX)	

Want More Info? Want to Test vSMP Foundation?

Need additional technical information, system requirements or want to know more about testing and implementing vSMP Foundation? Visit our site www.scalemp.com or mail info@scalemp.com

07/2012