Transforming Infrastructure for the New Era of Big Data and High Performance Computing

IBM Software-Defined Infrastructure (SDI) Solutions Enhance Business Agility, Productivity, Efficiency and Quality

IBM

Contents

Executive summary

Maximize business value with faster, deeper and higher quality insights by transforming IT infrastructure into one that can reliably and efficiently handle both Big Data Analytics (BDA) and High Performance Computing (HPC). This paper discusses the benefits of evolving to a Software-Defined Infrastructure (SDI) from traditionally discrete compute environments. An SDI is a single, more efficient and productive shared infrastructure for both HPC and BDA workloads as well as a new generation of born-in-the-cloud workloads.

Integrating High Performance Computing and Big Data Analytics

Driven by the need for faster and higher quality results, the lines between compute-intensive and data-intensive workloads are blurring. As data gathering techniques improve and simulation becomes more sophisticated, larger datasets are ingested, generated and retained at each stage of the analytic pipeline (Figure 1) - from data input, preparation and simulation to downstream analytics, visualization and interpretation. Across all industries, organizations are seeking to obtain maximum value from their data, and this demands faster, more scalable and more cost-effective IT infrastructure.

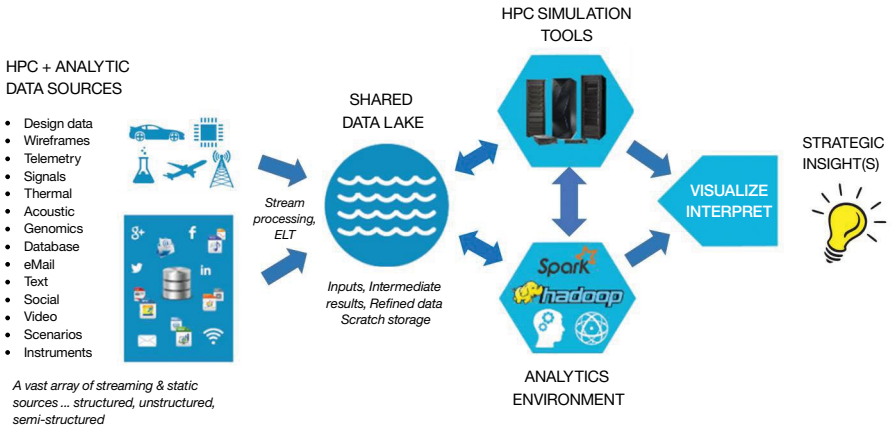

Figure 1. Intersection of HPC and Big Data Workflows

As we gather, process and store larger datasets from all sources, such as sensors, instruments, log files and so on, HPC workloads increasingly look like big data workloads. Big data workloads are becoming more compute-intensive in both performance and scale, and looking more like HPC workloads, particularly in the areas of cyber-security, fraud detection, and social data analytics. Both types of workloads place increasingly similar demands on IT infrastructures to the point where the same infrastructure can support both.

This New Generation of High-Performance and Data-Intensive Workloads

In the Automotive industry, engineers use HPC software to simulate vehicle collisions, and then conduct crash tests, collecting data from tens of thousands of sensors for further analysis. Increasingly they are analyzing big data such as field defect data, service and warranty data, and real-time telemetry from vehicles-in-use. By augmenting HPC with big data analytics, manufacturers can gain deeper analysis from petabytes of data to build better products.

In *Healthcare and Life Sciences*, genomic medicine pipelines are large and sophisticated workflows with dozens of compute- and data-intensive tasks that span across Next Generation Sequencing (NGS), Translational Medicine and Personalized Healthcare. To develop new treatments, institutions rely on HPC and increasingly big data technologies such as Apache Spark to execute hundreds of thousands of jobs to analyze petabytes of data including text and images that are often distributed across tens of thousands of files.

Financial Services firms are looking to maximize the value from their existing businesses while creating new revenue streams. Firms such as Fannie Mae are increasingly analyzing both structured and unstructured data including email and PDF files to improve profits and investment outcomes, as well as to find patterns and trends in their clients' and/or employees' activities that may signal investment opportunities or fraud.

Need for a Common Compute and Data Infrastructure

To support these more demanding compute- and data-intensive workloads and achieve higher quality results faster, organizations are demanding a faster, more scalable, and more capable infrastructure. Adding more hardware is not always possible or sustainable due to costs, complexities, and the risk of cluster and data sprawl. Building a single common infrastructure for both types of workloads is both desirable and feasible.

In a joint paper published by Indiana University and Rutgers University, researchers concluded that HPC and Big Data Analytics have many similarities to support the use of a common, unified infrastructure stack. They are:

- "Pleasingly parallel local structure is often seen in both simulations and big data"
- Models in the big data area "often need HPC hardware and software enhancements to get good performances"
- Simulations "nearly always involve a mix of point-to-point messaging and collective operations like broadcast, gather, scatter and reduction," sharing a commonality with big data problems, which often involve collective operations
- HPC simulations and big data tend to be loosely synchronous and iterative

The researchers concluded that HPC and big data workloads can share the same infrastructure, eliminating the need for separate IT silos. Sharing can also significantly reduce data costs, allowing data to be stored once and shared among different workloads. Consolidating compute and storage siloes simplifies system administration to further reduce costs and improve efficiency.

There are also challenges in combining multiple workloads onto a single infrastructure including: managing service level agreements (SLAs); harmonizing workload and resource managers; and supporting different hardware and file systems on-premises and in the cloud. So, how can you evolve to this shared, common infrastructure across different hardware platforms, on premises, and in public and hybrid cloud?

Introducing the Software-Defined Infrastructure Approach

Over the past few years, IT organizations have recognized the limitations of traditional IT architectures. A silo approach tends to foster inefficient use of and access to compute resources, resulting in artificial capacity shortages even if overall capacity is sufficient. The rapid adoption of big data frameworks like Hadoop MapReduce and Apache Spark, which benefit from maximum usage of resources in parallel, has amplified the need for a more unified approach to IT.

In response, firms are now looking to move toward a Software-Defined Infrastructure (SDI), which is a dynamic resource-, workload- and data-aware environment that adapts automatically to real-time computing needs. A SDI optimizes the placement and execution of workloads, orchestrating infrastructure resources as needed to meet SLAs. It is platform agnostic, supporting the widest range of hardware, frameworks, and APIs.

Evolving to an SDI enables your organization to handle HPC and big data applications as well as a newer generation of born-in-the-cloud frameworks on a single, more efficient, faster, and more agile infrastructure.

Benefits of SDI

A SDI supports both compute- and data-intensive workflows better than siloed IT architectures by:

- *Supporting multi-tenancy*, enabling different businesses and applications to share infrastructure in a well-behaved fashion. Sharing resources reduces costs and allows IT to adapt resources to support new business and compute demands faster. SLAs govern resource use ensuring fairness to all.
- *Optimizing use of compute resource of all kinds* for up to 150x faster time-to-results
- *Scaling to handle massive file* and job counts, and extreme I/O. In some industries such as life sciences, a single workflow may create and access up to one million files.[1] SDI supports a wide variety of storage architectures and devices to keep pace with massive I/O demands

IBM Spectrum Computing and IBM Spectrum Storage Solutions

Many clients are already realizing the benefits of a SDI—The full portfolio of IBM® Software Defined Infrastructure (Figure 2) solutions—IBM Spectrum™ Computing and IBM Spectrum Storage™ solutions—are proven to run the most demanding workloads.

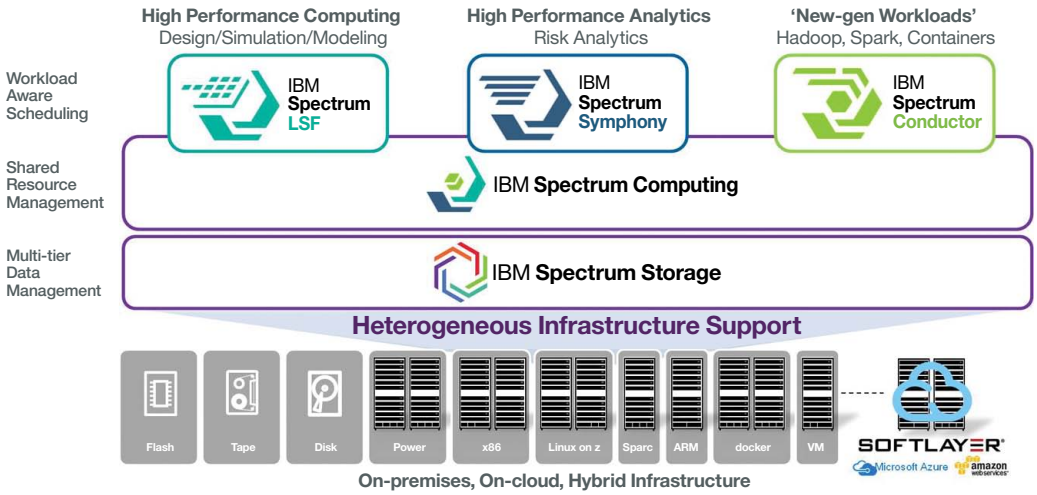

Figure 2. The Complete Software Defined Infrastructure Portfolio

The IBM Spectrum Computing portfolio (formerly IBM Platform Computing™) provides sophisticated and proven workload and resource management solutions—IBM Spectrum LSF, IBM Spectrum Symphony™ and IBM Spectrum Conductor™—for many of the world's leading organizations including Red Bull Racing, Wellcome Sanger Trust, Cypress Semiconductor, and the CME Group IBM Spectrum Storage is the first software defined storage portfolio designed to simplify and speed storage management.

IBM Software Defined Infrastructure has helped Forbes 2000 clients such as Citigroup accelerate their analytics up to 100 times, while lowering infrastructure costs—on premises and in the cloud—while meeting changing business demands faster.

Conclusions

By adopting Software Defined Infrastructure (SDI), you can scale and accelerate analytics, even as the volume, velocity and variety of data continue to grow. Clients across many industries are using IBM Software Defined Infrastructure portfolio of solutions—IBM Spectrum Computing and IBM Spectrum Storage software—for greater IT agility, productivity, and efficiency, and for faster and better insights into data of all types.

For more information

To learn more about IBM Spectrum Computing, please contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/systems/spectrum-computing/big-data-and-hpc

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing

New Orchard Road
Armonk, NY 10504
U.S.A.

Produced in the United States of America
October 2016

[1] Talk by Patricia Kovatch describing Mt Sinai Minerva systems at Usenix conference. https://www.usenix.org/sites/default/files/conference/protected-files/lisa15_slides_kovatch.pdf

Please Recycle

DCW03068-USEN-00