WHITE PAPER

BlueData Enables Virtualization of Enterprise Hadoop* and Spark* Workloads

Virtualization for Big Data
Intel® Xeon® Processor E5-2600 v3 Product Family

The BlueData EPIC* software platform offers data center operators the agility and cost performance of virtualized infrastructure for big data, with high manageability and flexibility when integrating into existing data center environments.

Even as virtualization has spread throughout the data center, Apache Hadoop continues to be deployed almost exclusively on bare-metal physical servers. Processing overhead and I/O latency typically associated with virtualization have prevented big data architects from virtualizing Hadoop implementations.

As a result, most Hadoop initiatives have been limited in terms of agility, with infrastructure changes such as provisioning a new server for Hadoop often taking weeks or even months. This infrastructure complexity continues to slow down adoption in enterprise deployments. Apache Spark is a relatively new big data technology, but interest is growing rapidly; many of these same deployment challenges apply to on-premises Spark implementations.

The BlueData EPIC (Elastic Private Instant Clusters) software platform addresses these limitations, enabling data center operators to accelerate Hadoop and Spark implementations on Intel® architecture-based servers.

Introduction to BlueData EPIC

The BlueData EPIC software platform reduces the complexity of big data infrastructure deployments, providing the ability for end users to quickly and easily deploy Hadoop or Spark clusters in a virtualized environment running on Docker containers. These clusters can deliver faster time-to-value for big data, providing the cloud-like experience of Hadoop-as-a-Service or Spark-as-a-Service in their own data centers.

The BlueData EPIC platform helps improve hardware utilization, reduces cluster sprawl, and minimizes the need to move data for big data analytics. BlueData EPIC also provides for simplified deployment and administration, while making virtual clusters look and feel like physical clusters for big data analytics.

Taking advantage of the power of containers and virtualization, BlueData's software helps deliver greater agility and cost-efficiency for on-premises big data infrastructure. The benefits of these capabilities include the following:

- **Business agility**. Virtual clusters can be spun up or down in minutes, providing elasticity for capacity spikes, as well as rapid response to emerging business needs.
- **Data protection**. Multiple virtual workloads can co-exist on the same multi-tenant physical cluster, while isolating data on each virtual cluster from the others.
- **Resource efficiency**. Multiple business units and user groups can share physical cluster resources, avoiding the cost and complexity of each having its own big data infrastructure.

To meet varying customer needs, the EPIC software platform is available in two editions. EPIC Lite is a community edition of the platform that is available for a single instance, free of charge; it is intended for evaluation purposes and for personal use. EPIC Enterprise is a fully supported, highly scalable commercial edition that is available on a subscription basis for up to hundreds of physical nodes. For a full comparison of the two product editions, see www.bluedata.com/product/comparison.

Hadoop-as-a-Service or Spark-as-a-Service in an On-Premises Deployment Model

The BlueData EPIC* software platform gives business users the ability to set up self-service virtual Hadoop* or Spark* clusters without having to submit requests for scarce IT resources and then wait for an environment to be set up for them. Within minutes, data scientists and analysts can deploy big data services and applications to meet their needs on demand.

The ability to explore, analyze, and draw insights from data allows users to seize business opportunities while they are still relevant.

- **Ad hoc analytics**. Identify emerging trends and relationships to enhance decision support.
- **"Fail-fast" experimentation**. Try out new approaches to big data challenges with minimal investment.
- **Rapid response**. Spin virtual clusters up and down fast, as changing needs and opportunities dictate.

BlueData EPIC Software Architecture

The core components of the EPIC platform—ElasticPlane*, IOBoost*, and DataTap*—are illustrated in Figure 1 and described below.

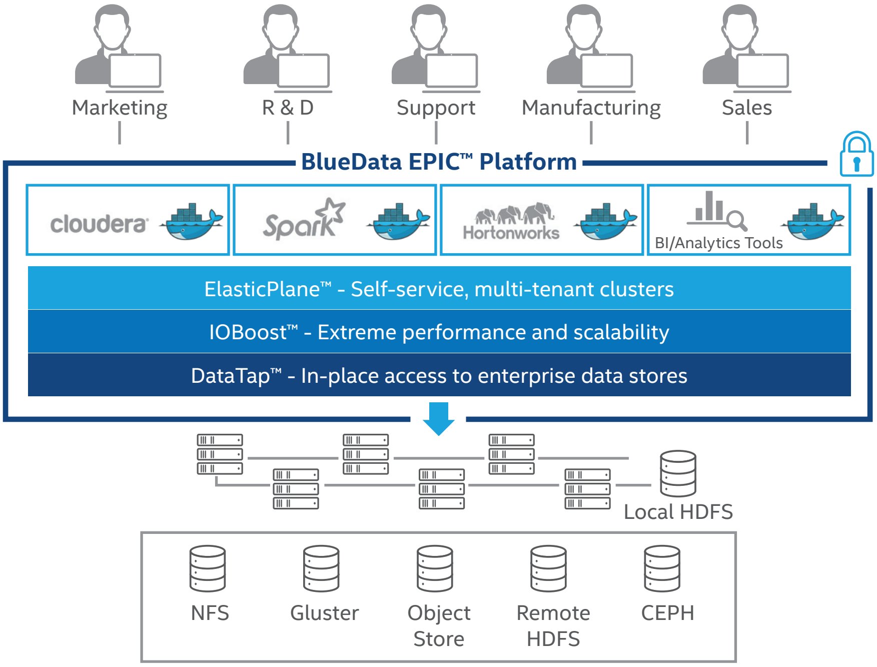

Figure 1. BlueData EPIC* software architecture.

ElasticPlane: Virtual Clusters on Demand

ElasticPlane enables spinning up virtual clusters on demand via self-service in a secure multi-tenant environment, with a policy engine for automated QoS and SLA management. End users can easily create virtual Hadoop or Spark clusters with BlueData EPIC 's ElasticPlane functionality and self-service interface. BlueData also provides multi-tenancy and data isolation to help ensure logical separation between each group within the organization.

The solution enables different project teams or departments across the enterprise to share the same physical infrastructure—and access the same data sources—for their big data analytics. The platform integrates with enterprise security and authentication mechanisms such as LDAP, Active Directory, and Kerberos*.

IOBoost: Enhanced Performance

IOBoost enhances the I/O performance and scalability of virtual clusters with hierarchical data caching and tiering, plus single-copy data transfer from physical storage to the virtual cluster. The IOBoost functionality of the BlueData EPIC platform provides application-aware caching and elastic resource management that adapts dynamically to changing application requirements, helping drive up performance.

Write-dominant workloads in particular benefit from IOBoost, which takes advantage of knowing how the application will access data. BlueData's IOBoost technology provides a non-persistent memory cache, the behavior of which changes to improve the efficiency of access to physical storage devices. IOBoost accesses the external file system by means of BlueData's DataTap file system connector.

Specifically, as the application writes data to the Hadoop Distributed File System (HDFS*), IOBoost functions as a write-behind cache, optimizing the performance of sequential writes.

DataTap: In-Place Processing of Data

DataTap allows in-place processing of data, eliminating the need to duplicate data across Hadoop systems. DataTap provides HDFS protocol abstraction that allows big data applications to run unmodified with fast access to data sources other than HDFS. With BlueData EPIC's DataTap capability, organizations can access data from any shared storage system (including HDFS as well as NFS*, GlusterFS*, CEPH*, and Swift*) for big data analytics.

That means organizations don't need to make multiple copies of data or move data into HDFS before running their analysis. Sensitive data can stay in their secure storage system with enterprise-grade data governance, without the cost and risks of creating and maintaining multiple copies.

DataTap effectively decouples compute from storage, providing the ability to independently scale compute and storage on an as-needed basis. This approach helps enable more effective utilization of infrastructure resources and lower data center operating costs.

Deployment Considerations and Guidance

BlueData's software applies patent-pending innovations to enable virtualization that is specifically tailored to the needs of big data. The use of Docker containers is completely transparent, but BlueData customers benefit from greater performance and deployment flexibility due to their lightweight nature. This enables enterprises to quickly and easily deploy Hadoop or Spark in a lightweight container environment, running on either bare-metal physical servers or on virtual machines.

BlueData EPIC supports Hadoop and Spark applications without requiring those applications to be modified in any way. Likewise, the platform utilizes the underlying features of the physical storage devices for data backup, replication, and high availability, so it is not necessary for organizations to modify their existing processes to facilitate security and durability of their data.

Synergies with Intel® Architecture

The ability to run on large clusters of mainstream two-socket servers extends the cost-performance advantages of virtualizing big data workloads. BlueData software running on the Intel® Xeon® processor E5-2600 v3 product family is a powerful combination to overcome key virtualization challenges such as network latency, infrastructure security, and power inefficiencies. Deploying BlueData EPIC environments on two-socket servers powered by these processors takes advantage of the following benefits:

- **Improved performance and virtualization density**. With increased core counts, larger cache, and higher memory bandwidth, the processor delivers dramatic improvements over its predecessors.
- **Hardware-based security**. Intel® Platform Protection Technology, including Intel® Trusted Execution Technology, Intel® OS Guard, and BIOS Guard, enhances protection against malicious attacks.
- **Increased power efficiency**. Per-core P states dynamically respond to changing workloads and adapt power levels on each individual core, to deliver better performance per watt than predecessor platforms.

Beyond the processor platform used with BlueData deployments, using Intel® Solid-State Drives (Intel® SSDs) helps optimize the execution environment at the system level. For example, the Intel® SSD Data Center (Intel SSD DC) P3608 Series[1] delivers high performance and low latency that help accelerate virtualized Hadoop workloads, using connectivity based on the Non-Volatile Memory Express (NVMe) standard and eight lanes of PCI Express* (PCIe*) 3.0.

Created by an industry coalition including Intel, NVMe replaces the older SATA standard with a new technology developed specifically to deliver latency and throughput advantages for high-speed SSDs and other non-volatile memory-based storage. The Intel SSD DC P3608 Series builds on those capabilities with a unique dual-controller architecture that improves scaling across the execution cores of Intel® Xeon® processors. It is available in a low-profile PCIe form factor, in capacities up to 4 TB.

Intel® Ethernet Controllers help accelerate workloads including those based on virtualized Hadoop with purpose-built capabilities for virtualization, such as intelligent offload of traffic management to network hardware. By handling traffic functions in network silicon, Intel Ethernet removes the associated burden from the processor, freeing execution resources for other work.

Configuration Best Practices

Ongoing experimentation by Intel and BlueData indicates that the following suggested guidelines may help data center operators achieve optimal throughput on virtualized Hadoop and Spark workloads using the BlueData EPIC software platform. While detailed examination is beyond the scope of this paper, the following guidance is particularly relevant to I/O-bound workloads.

- **Configure systems to enhance disk performance**. The performance of the storage where the files are stored must be sufficient to avoid a bottleneck.

- **Provide sufficient network throughput**. The performance of the network connectivity between the EPIC hosts must be sufficient to avoid bottlenecks.
- **Deploy using current-generation Intel® processors**. The architecture of each generation of Intel Xeon processors provides advances in terms of performance and power efficiency, which can have significant benefits to Hadoop or Spark workloads running on BlueData EPIC.

Conclusion

The BlueData EPIC software platform enables virtualization of Hadoop and Spark workloads as a viable alternative to bare-metal implementations. This breakthrough capability makes BlueData and Intel Xeon processors key ingredients in big data deployments for customers that want to take advantage of containers and virtualization to enhance the performance, efficiency, and agility of their implementations.

As a result, Hadoop or Spark on BlueData EPIC software and Intel architecture has become the solution stack of choice for many big data initiatives, providing an optimized set of building blocks to meet emerging needs faced by businesses of all types and sizes. Going forward, big data deployments can be a part of broader virtualization initiatives, including taking advantage of ongoing improvements to Intel Xeon processor-based platforms.

In particular, companies can provide Hadoop-as-a-Service or Spark-as-a-Service for their business users, empowering them to design and implement analytics on demand while leveraging on-premises infrastructure, governance, and security. These capabilities allow them to put data to use more easily as changing business needs warrant, supporting the evolving demands of business units for self-determination while also making IT more responsive to the needs of the business as a whole.

Supporting Resources

- Intel blog: Simplify Big Data Deployment http://blogs.intel.com/evangelists/2015/08/25/simplify-big-data-deployment/
- BlueData blog: New Funding and Strategic Collaboration with Intel http://bluedata.com/blog/2015/08/new-funding-and-strategic-collaboration-with-intel

For more information, visit intel.com/bigdata and bluedata.com

[1] www.intel.com/content/www/us/en/solid-state-drives/solid-state-drives-dc-p3608-series.html.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Copyright © 2015 Intel Corporation. All rights reserved. Intel, the Intel logo, and Xeon are trademarks of Intel Corporation in the U.S. and other countries.

*Other names may be trademarks of their respective owners.

Printed in USA 1115/MW/MESH/PDF ♻ Please Recycle 332621-001US