

IBM Spectrum Conductor and IBM Spectrum Conductor with Spark

Dino Quintero

Nicolas Joly



 **Cloud**

Infrastructure Solutions

Find and read thousands of IBM Redbooks publications

- ▶ Search, bookmark, save and organize favorites
- ▶ Get personalized notifications of new content
- ▶ Link to the latest Redbooks blogs and videos

Get the latest version of the Redbooks Mobile App



iOS

Download
Now

Android



Promote your business in an IBM Redbooks publication

Place a Sponsorship Promotion in an IBM® Redbooks® publication, featuring your business or solution with a link to your web site.

Qualified IBM Business Partners may place a full page promotion in the most popular Redbooks publications. Imagine the power of being seen by users who download millions of Redbooks publications each year!



ibm.com/Redbooks

About Redbooks → Business Partner Programs

THIS PAGE INTENTIONALLY LEFT BLANK



IBM Spectrum Conductor and IBM Spectrum Conductor with Spark

In an era of rapid change and increasing competition, organizations need a smarter infrastructure that is agile, flexible, and resilient. IBM Spectrum Conductor™ delivers those capabilities with robust features that enable organizations to store, analyze, and protect their data with optimum efficiency.

IBM Spectrum Conductor is the next evolution of IBM® software-defined infrastructure technology. It is an integrated application and data-optimized platform that enables organizations to achieve up to 60%¹ faster results with new generation, cloud-native applications and open source frameworks. The product achieves these results by efficiently analyzing, accessing, and protecting data.

Note: IBM Spectrum Conductor is compatible with IBM Spectrum™ Computing, complements IBM Spectrum Storage™ offerings, and is part of the software-defined infrastructure portfolio.

The IBM Spectrum Conductor offerings include the following products:

- ▶ IBM Spectrum Conductor V2.1
- ▶ IBM Spectrum Conductor with Spark V2.1

This IBM Redpaper™ publication describes IBM Spectrum Conductor and IBM Spectrum Conductor with Spark. It includes the following topics:

- ▶ The data growth challenge
- ▶ IBM Spectrum Computing family offering names
- ▶ IBM Spectrum Computing
- ▶ What is IBM Spectrum Conductor?
- ▶ Why IBM Spectrum Conductor?
- ▶ What is IBM Spectrum Conductor with Spark?
- ▶ Why IBM Spectrum Conductor with Spark?
- ▶ IBM Spectrum Computing resource manager
- ▶ Specified operating environment
- ▶ Specified operating environment
- ▶ IBM Spectrum Computing product family
- ▶ Licensing and ordering the solution
- ▶ Additional references

¹ STAC REPORT: Evaluates leading resource managers running Spark workloads in a multi-user environment:
https://www.ibm.com/marketing/iwm/dre/signup?source=STG-Web-SDI-NA&S_PKG=ov47883

The data growth challenge

The digital world is generating tremendous amounts of data. This “ocean” of data is creating big challenges for organizations that require the ability to navigate and rapidly explore these vast oceans of data to find valuable insights. These oceans of data must be cost and performance optimized, and be efficiently stored, managed, and protected. In addition, the data must be accessible to the correct applications at the correct time.

The amount of data is growing at exponential rates. Although some of the data does not need to be kept, the rest of the data is a critical asset to any corporation. To gain competitive advantages from this data, your organization needs to access it, analyze it, and protect it, as shown in Figure 1.

Increasingly complex application portfolios and traditional infrastructure solutions add to the complexity of managing the data. All of these components have created many data silos (oceans of data), under-used infrastructures, and duplicated data. Therefore, it can become cost-prohibitive and inefficient to analyze all of the data.

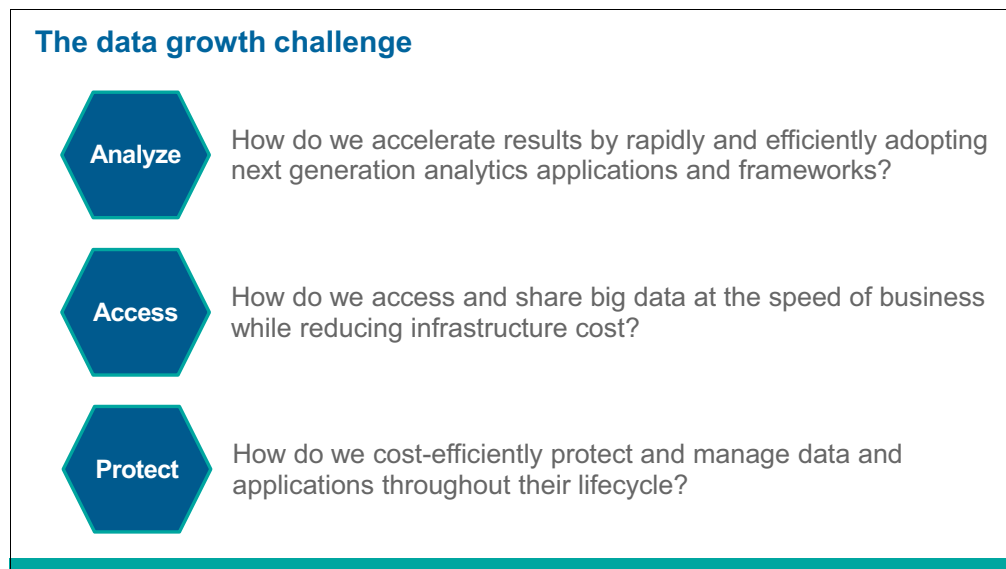


Figure 1 Data growth challenge aspects

The new generation of modern, scale-out applications and frameworks is helping simplify increasingly complex application portfolios and traditional infrastructure solutions. With the introduction of the IBM Spectrum Conductor and the IBM Spectrum Conductor with Spark, IBM provides a multiscale platform for these next generation applications.

These solutions provide multidimensional, independent application and data scalability. They can be deployed on premises, in the cloud, or as an integrated solution, as shown in Figure 2.

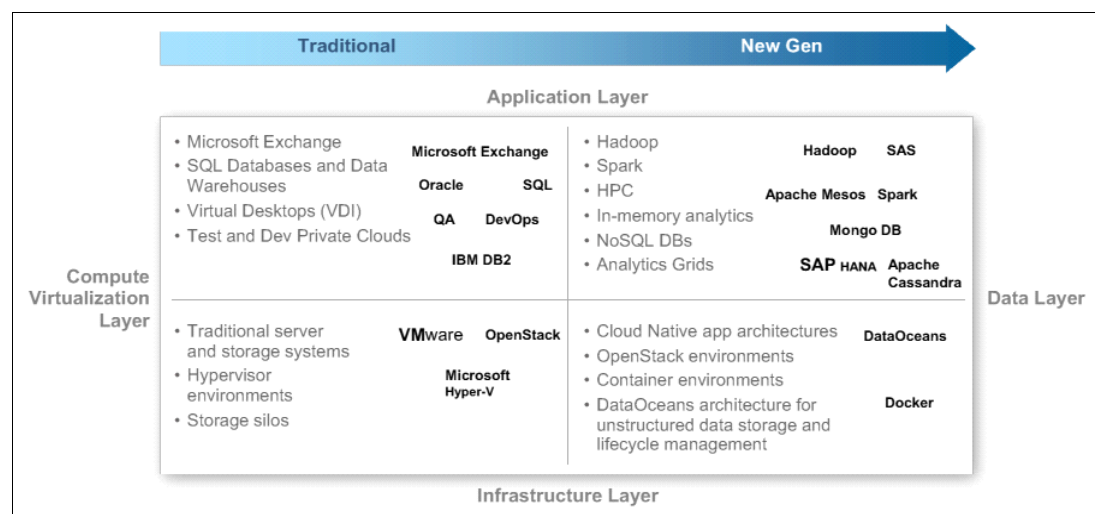


Figure 2 IBM Spectrum Conductor simplifies complex application portfolios and infrastructure

These solutions are based on the IBM Software Defined Infrastructure technology, which has been deployed and proven in some of the world's most demanding environments.

IBM Spectrum Computing family offering names

The IBM Spectrum Computing family (formerly known as *IBM Platform Computing*) includes three core members plus two complementary offerings. Table 1 lists the family offering names that were rebranded.

Note: The IBM Platform Cluster Manager *is* part of the IBM Spectrum Computing family. However, its name remains unchanged.

Table 1 IBM Spectrum Computing family

New name	Previous name
IBM Spectrum Load Sharing Facility (IBM Spectrum LSF®)	IBM Platform LSF
IBM Spectrum Symphony™	IBM Platform for IBM Symphony®
IBM Spectrum Conductor	IBM Platform Converge
IBM Spectrum Message Passing Interface (IBM Spectrum MPI)	IBM Platform MPI

For a more complete overview of the IBM Spectrum Computing family of offerings, see "IBM Spectrum Computing product family" on page 17.

For more information about the IBM Spectrum Computing family offerings, current announcements, and rebranding information, see the following website:

<http://www.ibm.com/systems/spectrum-computing/>

IBM Spectrum Computing

IBM Spectrum Computing uses intelligent workload-driven and policy-driven resource management to optimize resources across the data center, whether on premises or in the cloud. IBM Spectrum Computing is now faster, more flexible, and scalable to over 160,000 cores. IBM uses advances in software-defined infrastructure to help you take advantage of the power of your distributed computing environment, as shown in Figure 3.

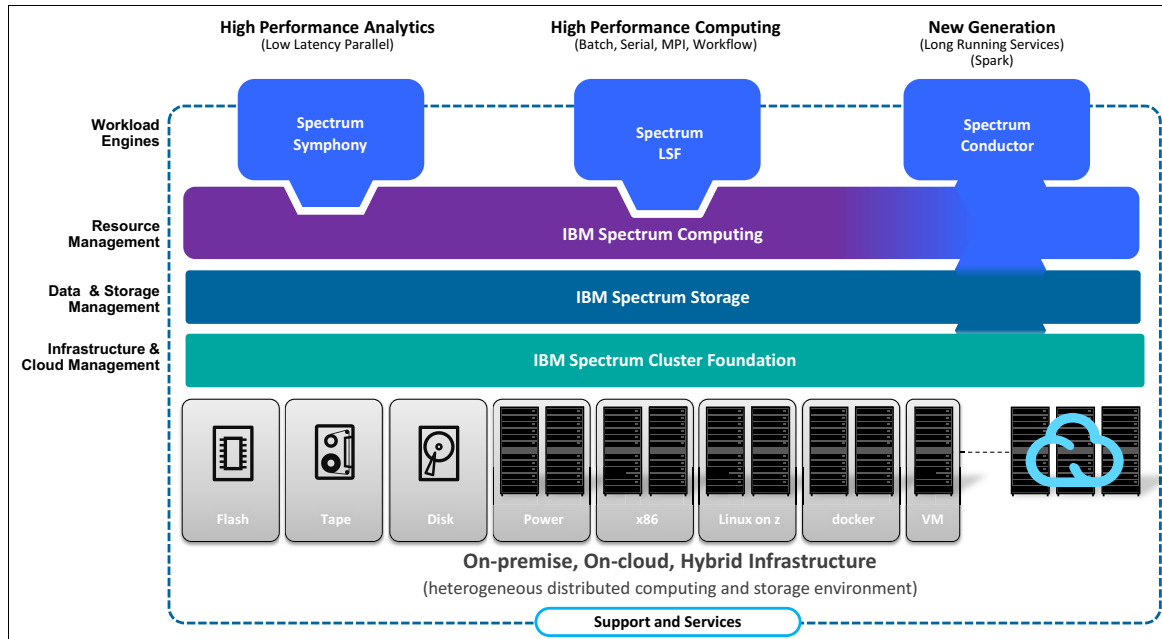


Figure 3 IBM software defined infrastructure solution

For a more complete overview of the IBM Spectrum Computing family of offerings, see “IBM Spectrum Computing product family” on page 17.

What is IBM Spectrum Conductor?

IBM Spectrum Conductor is a multiscale platform for new generation, cloud-native applications and open source frameworks, such as Hadoop, Spark, and NoSQL databases (MongoDB, Hbase, and so on). The platform stores, analyzes, and protects unstructured big data on an application-optimized converged infrastructure.

The ultimate aim of a software defined infrastructure is to yield an application-aware and data-aware environment that captures workload requirements, provides policy-based automation across data center environments, and includes analytics to optimize in real time.

IBM Spectrum Conductor achieves these goals in the most cost-efficient way possible through three core capabilities (Figure 4 on page 5):

- **Analyze.** Workload and data-aware platform increases usage of existing hardware resources and speeds analysis.
- **Access.** Policy-driven data placement with global access enables organizations to tackle all facets of storing data and sharing resources for distributed teams or data centers.
- **Protect.** Enterprise-class features, such as data encryption, automatic failover, and seamless system recovery, provide enhanced resilience and protection.

Figure 4 shows the core capabilities of IBM Spectrum conductor.

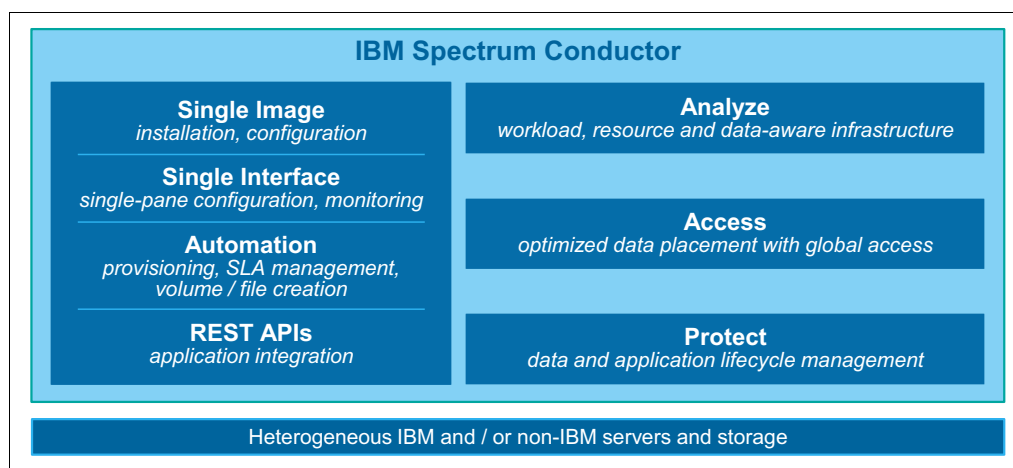


Figure 4 IBM Spectrum Conductor solution representation

Understanding the components of IBM Spectrum Conductor

IBM Spectrum Conductor has the following components (Figure 5 on page 6):

- ▶ The Elastic Stack (Elasticsearch, Logstash, Kibana, and Beats) is integrated within IBM Spectrum Conductor for data collection and visualization. Registered as system services, the Elastic Stack integration enables you to search, analyze, and visualize application data for efficient monitoring. Kibana is also embedded within the management console for data visualization.
- ▶ The IBM Spectrum Computing resource manager, also known as the *enterprise grid orchestrator* (EGO), manages the supply and distribution of resources, making them available to applications. It provides resource provisioning, remote execution, high availability, and business continuity.

The resource orchestrator provides cluster management tools and is akin to a distributed operating system that pools a heterogeneous set of nodes into a large distributed virtual computer. It allows multiple types of workloads to share resources efficiently, increasing usage while ensuring service-level agreements (SLAs) for individual applications.

- ▶ The application service controller daemon (ASCD) manages the lifecycle of application instances. Registered as an EGO service, the ASCD acts as a Representational State Transfer (REST) application programming interface (API) layer. It enables application instances to be described as a set of inter-related long-running services in a single specification. The services associated with the application instances can then be deployed and scaled.
- ▶ IBM Spectrum Scale™ File Placement Optimizer (FPO) is a Portable Operating System Interface (POSIX)-compliant storage management technology, and a more space-efficient alternative to Hadoop Distributed File System (HDFS), which is also supported.
- ▶ IBM Spectrum Cluster Foundation provides bare metal deployment capabilities to provide an infrastructure management layer.

The main components of IBM Spectrum Cluster are shown in Figure 5.

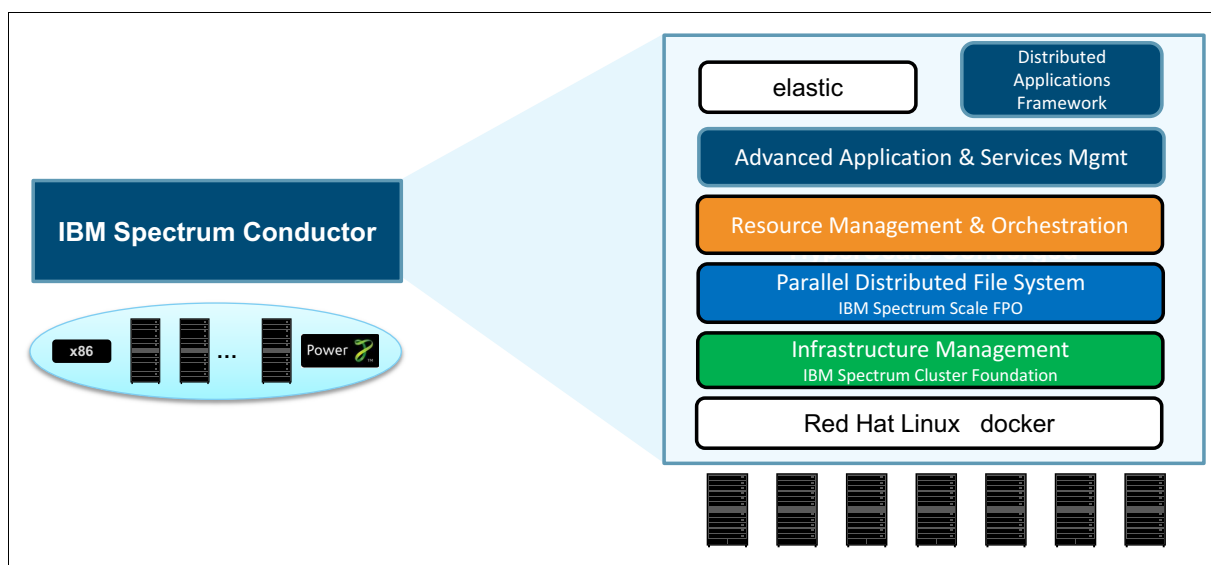


Figure 5 IBM Spectrum Conductor with Spark: End to end enterprise class solution

Why IBM Spectrum Conductor?

Organizations of all types are realizing data is their most valuable business asset, and that extracting full value from that data provides a critical competitive advantage. However, with today's massive data volumes continuing to grow, organizations are looking for a complete solution that answers the following questions:

- ▶ How do I accelerate business insights from all data by leveraging next generation applications and frameworks such as Spark and MongoDB?
- ▶ How do I store and access big data cost-effectively and not in disconnected silos?
- ▶ How do I best manage that data throughout its lifecycle while ensuring maximum security, reliability, and availability?

IBM Spectrum Conductor helps organizations answer these questions in the most cost-efficient way possible, by transforming their infrastructure into a tightly integrated data and application-optimized platform for next generation, cloud native applications.

IBM Spectrum Conductor is designed to enable organizations to adopt new generation cloud native applications and open source frameworks efficiently, effectively, and confidently. Resource and data sharing can be significantly increased without compromising availability or security. Unlike open source HDFS, IBM Spectrum Conductor is POSIX-compliant and provides significant storage efficiencies compared with HDFS.

Remember: IBM Spectrum Conductor also supports HDFS for users who prefer that option.

Users can share and manage the data over many applications:

- ▶ Data volumes are created on demand and owned by LOBs/individuals.
- ▶ Data is shared between users applications and seen as a shared file system.
- ▶ The wanted data volumes are connected to Docker containers and explain what users can do with it.

When a cluster is running, it is important to be able to monitor and maintain the system. IBM Spectrum Conductor helps simplify management with integrated cluster and application management control tools. Administrators can see where the application services are running, monitor the status of services, and see how resources are being allocated to meet service-level objectives. Figure 6 is an example of the IBM Spectrum Conductor dashboard, which presents a unified view of all the compute and storage resources of a managed cluster.

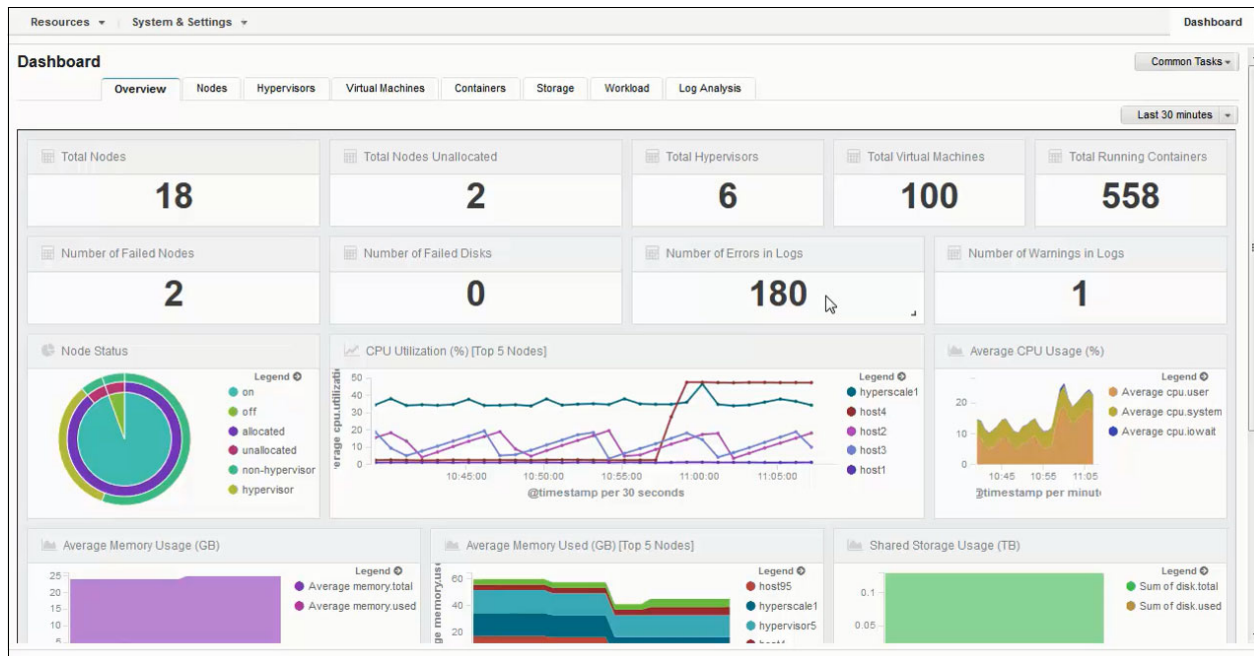


Figure 6 IBM Spectrum Conductor dashboard

What is IBM Spectrum Conductor with Spark?

IBM Spectrum Conductor with Spark is a complete enterprise-grade, multi-tenant solution for Apache Spark. It implements the concept of IBM Spectrum Conductor to address the requirements of users who need to adopt the Apache Spark technology, and to integrate it into their environment.

Apache Spark in the enterprise

Apache Spark is an open source cluster computing framework for large-scale data processing. Like Hadoop MapReduce, Spark provides parallel distributed processing, fault tolerance on commodity hardware, and scalability. With its in-memory computing capabilities, analytic applications can run up to 100 times faster on Apache Spark compared to other technologies on the market today.

Apache Spark is highly versatile, and known for its ease of use in creating algorithms that harness insight from complex data. In addition to its ease of use, this framework covers a wide range of workloads through its different modules:

- ▶ Interactive queries through Spark SQL.
- ▶ Streaming data, with Spark Streaming.
- ▶ Machine Learning, with the MLlib module.
- ▶ Graph processing with GraphX.

Applications can be built using simple APIs for Scala, Python, and Java:

- ▶ Batch applications leveraging Hadoop MapReduce compute model.
- ▶ Iterative algorithms that build upon each other.
- ▶ Interactive queries and data manipulation through *notebooks* (web-based interface).

Spark runs on Hadoop clusters, such as Hadoop YARN or Apache Mesos, or even as a stand-alone component with its own scheduler.

Figure 7 presents the Spark architecture.

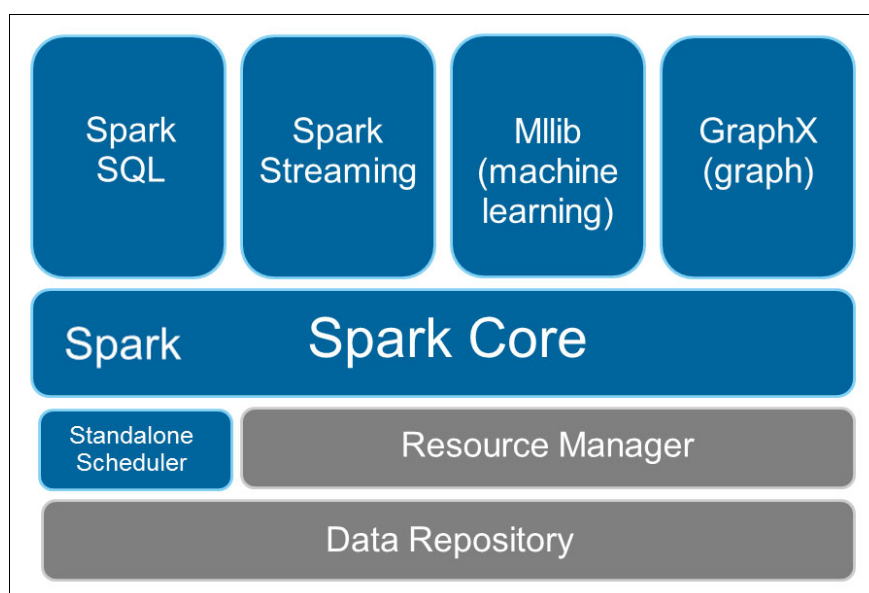


Figure 7 Spark software stack

Developed in the AMPLab at the University of California, Berkeley, Spark was elevated to a top-level Apache Project in 2014 and continues to expand today.

Despite its relative ease of use, Apache Spark deployments present multiple challenges in the enterprise:

- ▶ Lifecycle management. The Spark framework, like many other open source projects, experiences a quick release cycle.
- ▶ Expertise. Need to configure and install various tools for monitoring, management, and workflow.
- ▶ Cluster sprawl. Multiplicity of ad hoc Spark and Hadoop MapReduce clusters.
- ▶ Resource usage. Need to set up and configure a resource sharing manager to drive an efficient use of the compute resources.

IBM Spectrum Conductor with Spark value proposition

To address the challenges mentioned previously, IBM Spectrum Conductor with Spark delivers the following benefits:

- ▶ Accelerate results. Run Spark natively on a shared infrastructure without the dependency of Hadoop, helping reduce application wait time and accelerating time to results.
- ▶ Reduce administration costs. Proven architecture at extreme scale, with enterprise class workload management, monitoring, reporting, and security capabilities.

- Increase resource usage. Fine grain and dynamic allocation of resources maximizes the efficiency of Spark instances that share a common resource pool. Extends beyond Spark and eliminates cluster sprawl.
- End-to-end enterprise class solution. A tightly integrated offering that combines the IBM supported Spark distribution with workload, resource, and data management, as well as IBM support and services.

IBM Spectrum Conductor with Spark integrates notebook functionalities, and takes advantage of a notebook's graphical user interface (GUI) to manipulate and visualize data. In addition to the built-in Apache Zeppelin notebook, third-party notebooks, such as iPython-Jupyter, can be used.

For more information about notebook management and integrating third-party notebooks in IBM Spectrum Conductor with Spark, see the following website:

<https://ibm.biz/BdHuEW>

Figure 8 shows an example of a notebook running on Apache Zeppelin in an IBM Spectrum Conductor with Spark instance with embedded emulated log-on facility (ELK) visualization.

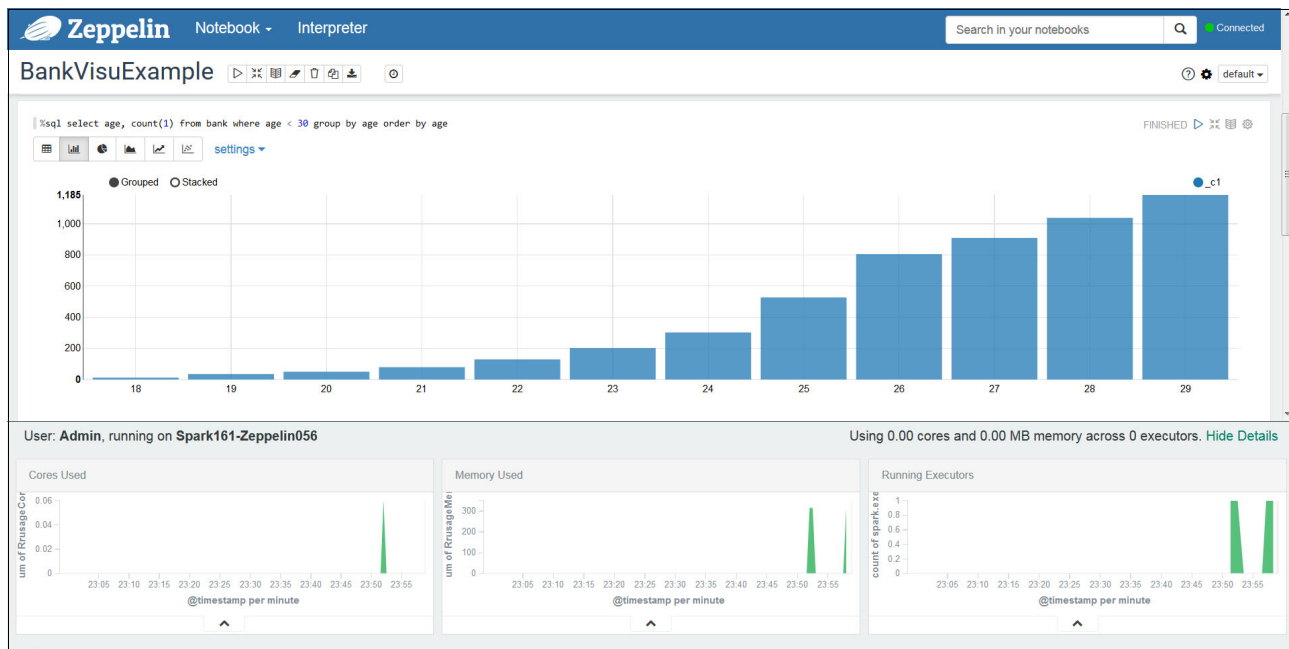


Figure 8 Zeppelin note running with Spark 1.6.1 core on IBM Spectrum Conductor with Spark

Note: As other Spark and Zeppelin versions become available, you can download them as part of a Spark package from IBM Fix Central at the following website:

<https://www.ibm.com/support/fixcentral/>

Understanding the components of IBM Spectrum Conductor with Spark

The IBM Spectrum Conductor with Spark framework shown in Figure 9 consists of the following components:

- ▶ A Spark distribution containing the core components of Apache Spark: Spark Core, Spark SQL, Spark Streaming, MLib Machine Learning Library, and GraphX.
- ▶ A notebook framework for data manipulation and visualization: Apache Zeppelin is built in, and other third-party notebook frameworks can be used.
- ▶ The Elastic Stack (Elasticsearch, Logstash, Kibana, and Beats) is integrated within IBM Spectrum Conductor with Spark for data collection and visualization: Registered as system services, the Elastic Stack integration enables you to search, analyze, and visualize Spark application data for efficient monitoring. Kibana is also embedded within the management console for data visualization.
- ▶ The IBM Spectrum Computing resource manager, also known as the enterprise grid orchestrator (EGO): Manages the supply and distribution of resources, making them available to applications. It provides resource provisioning, remote execution, high availability (HA), and business continuity. This component replaces the embedded stand-alone scheduler of the Apache Spark stack, and interfaces with Spark Core through a plug-in logic (Spark on EGO).
- ▶ The ASCD manages the lifecycle of a Spark instance group: Registered as an EGO service, the ASCD acts as a REST API layer and enables Spark instance groups to be described as a set of inter-related, long-running services in a single specification. The services that are associated with the Spark instance groups can then be deployed and scaled.
- ▶ IBM Spectrum Scale FPO is a POSIX-compliant storage management technology, and a more space-efficient alternative to HDFS (which is also supported). When using IBM Spectrum Scale, IBM Spectrum Conductor with Spark becomes an end-to-end IBM-supported Spark solution. This component provides the data layer.
- ▶ IBM Spectrum Cluster Foundation provides bare metal deployment capabilities to provide an infrastructure management layer.

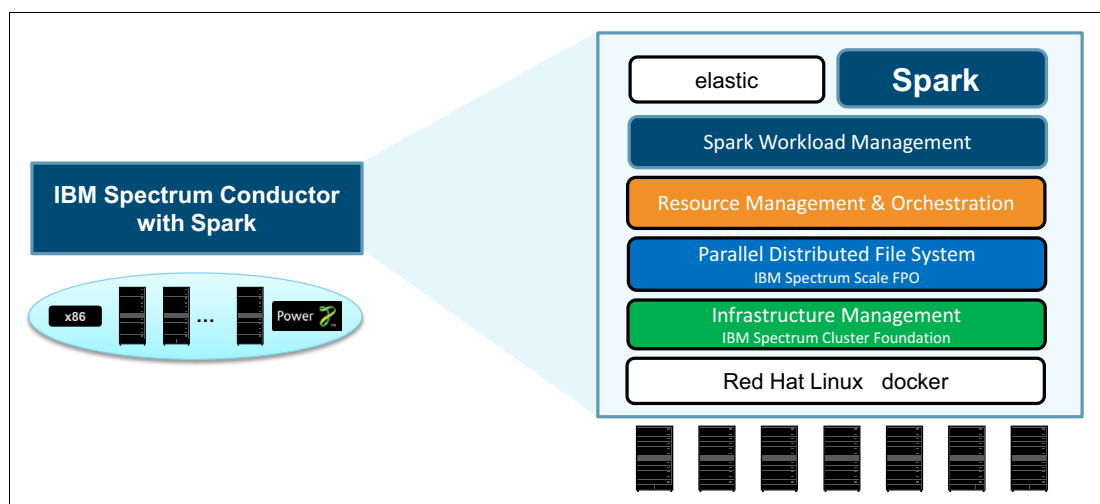


Figure 9 IBM Spectrum Conductor with Spark: End to end enterprise class solution

Both IBM Spectrum Scale and IBM Spectrum Cluster Foundation are optional in the following scenarios:

- ▶ You plan to deploy the solution in an existing HDFS cluster.
- ▶ You plan to deploy the solution in an existing infrastructure. The operating system is already deployed on the hosts.

Why IBM Spectrum Conductor with Spark?

Apache Spark, thanks to its speed, ease of use, and rich environment, is increasingly being adopted in the enterprise as the main platform for analytics workloads. More applications are being developed based on Apache Spark, but with different Spark versions, using different tools, and for different users. These heterogeneous environments become a challenge for the IT departments of today's enterprise.

The easiest way to handle these challenges is to set up isolated clusters for different lines of business, but this is expensive and inefficient in a modern IT infrastructure. IT demands a software product to provide Spark multitenancy on a shared physical cluster.

Organizations that are looking to deploy an Apache Spark environment should consider a purpose-built version, such as IBM Spectrum Conductor with Spark. It enables you to deploy Apache Spark efficiently and effectively, supporting multiple deployments of Spark, increasing performance and scalability, maximizing usage of resources, and eliminating silos of resources that would otherwise each be tied to separate Spark implementations.

Spark Instance Group

IBM Spectrum Conductor with Spark introduces a new concept, called *Spark Instance Group*, for the notion of a Spark *tenant*. Each Spark instance group is an installation of Apache Spark that can run Spark core services (Master, Shuffle, and History) and notebooks as configured.

A Spark instance group contains the following elements:

- ▶ Spark core services
- ▶ Spark tools and notebook
- ▶ User and applications
- ▶ Basic isolation

A Spark instance group can be created to serve a line of business or a group member of a business organization. Figure 10 shows an example of a Spark instance group as created in IBM Spectrum Conductor with Spark.

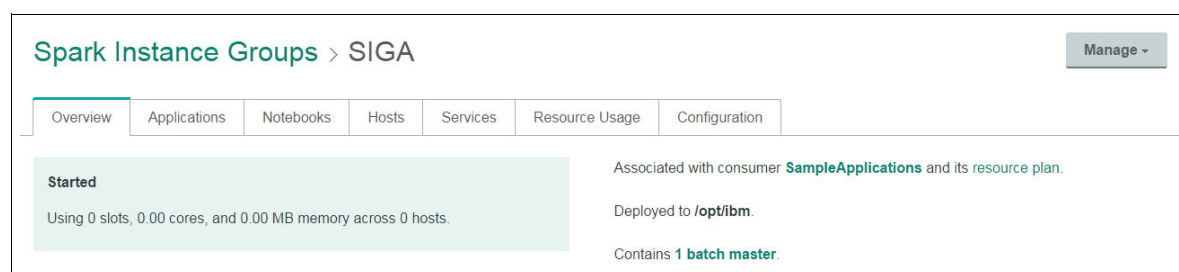


Figure 10 Spark instance group

Figure 11 shows an example of multiple applications that run in different instances, which are managed in the IBM Spectrum Computing Cluster management console.

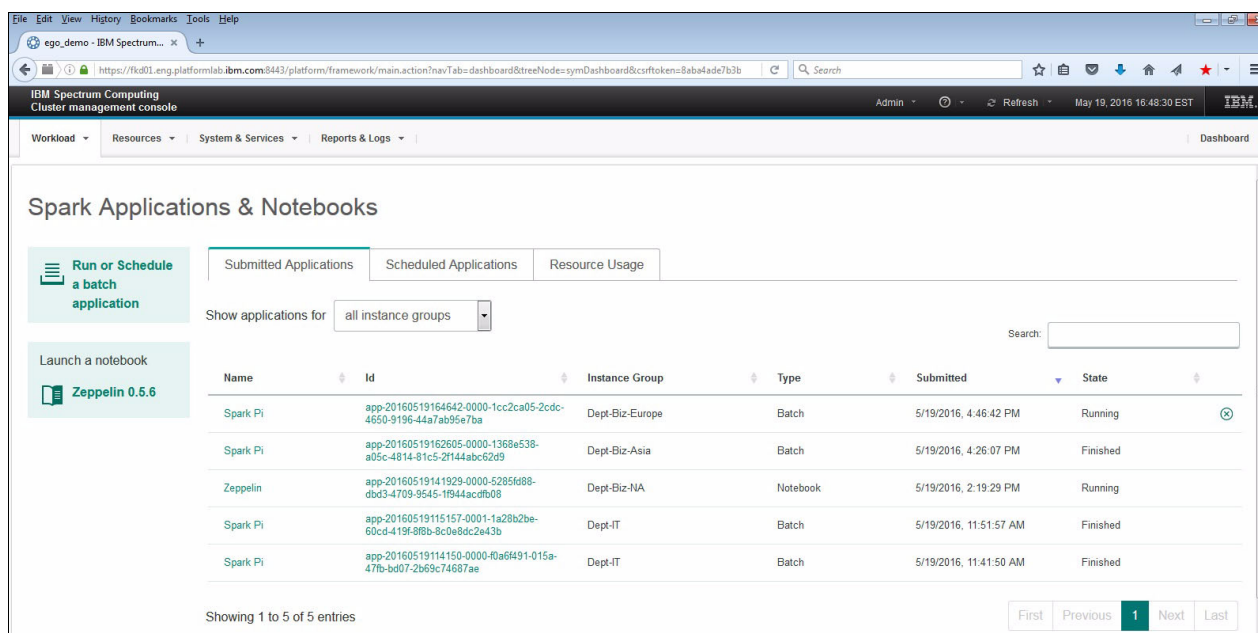


Figure 11 Multiple Spark applications running in different instances

IBM Spectrum Computing resource manager

IBM Spectrum Conductor product family multitenancy capabilities are possible due to the resource manager available as a core component of these solutions: IBM Spectrum Computing resource manager.

This resource orchestrator, also known as the enterprise grid orchestrator (EGO), manages the supply and distribution of resources, making them available to applications. It provides a full suite of services to support and manage resource orchestration in a cluster, including cluster management, configuration, and auditing of service-level plans, failover capabilities, monitoring, and data distribution.

Only the resource requirements are considered when allocating resources. Therefore, the business services can use the resources with no interference.

EGO uses resource groups to organize and manage the supply of resources, which are then allocated to different workloads according to policies. Resource groups can be static or dynamic, using different host attributes to define membership, or simply using tags. Resources can also be logical entities that are independent of nodes (bandwidth capacity, software licenses).

Those resources are used through consumers. A *consumer* is a logical structure that creates the association between the workload demand and the resource supply. Consumers are organized hierarchically into a tree structure to reflect the structure of a business unit, department, projects, and so on.

Demand - Workload

Client machines running App A

Application 'A'

Workload demand for App A

Consumer1

Consumer Structure

Consumer2

Consumers provide the right resources for a given workload

Client machines running both apps

Application 'A'

Application 'B'

Workload demand for App B

Client machines running App B

Application 'B'

Supply - resources

Resource Group 1: Made up of "x" type of machines

Resource Group 2: Made up of "y" type of machines

Individual use, project, department, LOB, or entire company

ORG

Dept-1

Team-1

Team-2

Dept-2

Proj-1

Proj-2

Dept-3

Consumers can be divided into lower level consumers, which can be subdivided. The lowest level consumer is the level at which the application is associated.

Next, policies that assign different amounts of resources to different consumers are defined in resource plans, as shown in Figure 13.

-
- Resource Plan**
- Resource Group: ComputeHosts Time Intervals and Settings
- Slot allocation policy
- | Consumer | Owned Slots | Consumer Rank | Lend Limit | Borrow Limit | Model type | Share Ratio | Limit |
|-----------------------|-------------|---------------|--------------|----------------|------------|-------------|-------|
| symphdemo | 100 | | | | | | |
| symTesting | 0 | 0 | | | | | 1 |
| symTesting1 | 0 | 0 | | | | | 1 |
| Total | 0 | | | | | | |
| Balance | 0 | | | | | | |
| SampleApplications | 0 | 0 | | | | | 1 |
| SOASamples | 0 | 0 | | | | | 1 |
| EclipseSamples | 0 | 50 | | | | | 1 |
| Total | 0 | | | | | | |
| Balance | 0 | | | | | | |
| SymExec | 0 | 0 | | | | | 1 |
| SymExec61 | 0 | 0 | | | | | 1 |
| Total | 0 | | | | | | |
| Balance | 0 | | | | | | |
| MapReduceConsumer | 40 | 0 | | | | | 1 |
| MapReduce61 | 10 | 0 | | | | | 1 |
| MapReduceHighPriority | 30 | 0 | | | | | 1 |
| MapReduceDefault | 0 | 0 | | | | | 1 |
| Total | 40 | | | | | | |
| Balance | 0 | | | | | | |
| SampleAppCPP | 0 | 0 | | | | | 1 |
| GpuTestApp | 0 | 0 | | | | | 1 |
| Total | 40 | | | | | | |
| Balance | 68 | | | | | | |
- Lend Details**
- MapReduce61 (ComputeHosts, 00:00-24:00)
- Total lend limit: 0
- Lend to these consumers
- Consumers to lend to
- symphdemo
- symTesting
- symTesting1
- SampleApplications
- SOASamples
- EclipseSamples
- SymExec
- SymExec61
- MapReduceConsumer
- MapReduce61
- MapReduceHighPriority
- MapReduceDefault
- SampleAppCPP
- GpuTestApp
- [Expand All](#)
- [Collapse All](#)
- [Apply](#) [Revert](#) [Close](#)
- Borrow Details**
- MapReduceDefault (ComputeHosts, 00:00-24:00)
- Total borrow limit: 0
- Borrow from consumers
- Consumers to borrow from
- symphdemo
- symTesting
- symTesting1
- SampleApplications
- SOASamples
- EclipseSamples
- SymExec
- SymExec61
- MapReduceConsumer
- MapReduce61
- MapReduceHighPriority
- MapReduceDefault
- SampleAppCPP
- GpuTestApp
- [Expand All](#)
- [Collapse All](#)
- [Apply](#) [Revert](#) [Close](#)

13

The resource plan allows SLAs to be created so that each line of business can meet its objectives, while sharing a common set of resources. The resource requirements can accommodate multi-dimensional resource allocations where each allocation can request different amounts of physical resource types, including but not limited to processors (CPUs), cores, memory, and number of disks.

Note: For more information about the technology behind IBM Spectrum Computing resource scheduler, see the following website:

<http://ibm.co/1TKU1Mg>

Tip: You need to create an IBM ID in order to access this website and retrieve the information.

Integration with Apache Spark

Apache Spark applications run as independent sets of processes on a cluster that is coordinated by the Spark Context object in the driver program. The Spark Context object can connect to the cluster manager (either Spark's own stand-alone cluster manager, Apache Mesos, Hadoop YARN, or EGO in IBM Spectrum Conductor with Spark) through the Session Scheduler, which allocates resources across applications, as shown in Figure 14.

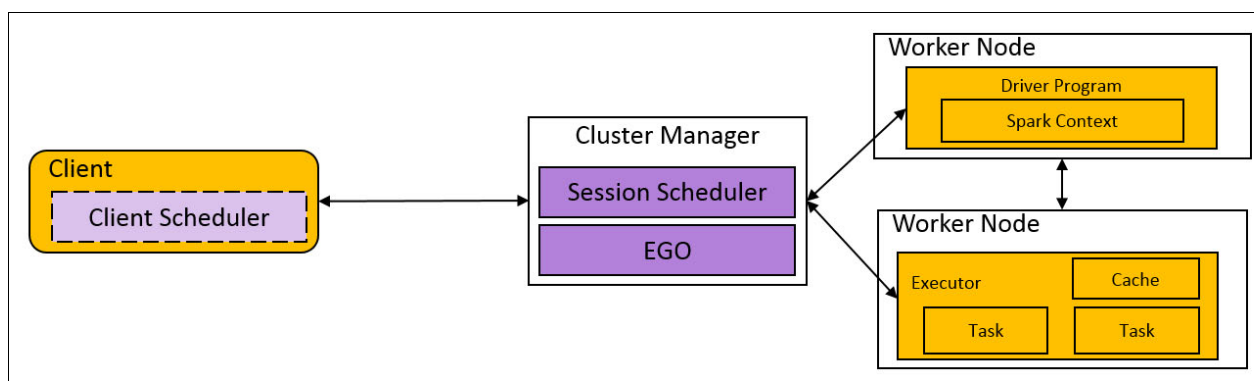


Figure 14 Spark integration with IBM Spectrum Computing resource manager

In IBM Spectrum Conductor with Spark, the resource orchestrator (EGO) acts as the cluster manager with Apache Spark, enabling Spark applications to benefit from resource sharing across the cluster. This provides the following benefits:

- ▶ **Fine-grained scheduling.** Apache Spark uses a coarse-grained or *fine-grained* resource scheduling policy. The Spark application requests a static number of resources and holds them in its lifecycle, which means each application gets more or fewer resources as it scales up and down. Based on the fine-grained scheduling policy, applications share resources at a very fine granularity, especially when many applications are running in a cluster concurrently.
- ▶ **Resource sharing and reclaim.** Tenants (known as a *Spark instance group*) share resources that are managed in consumers and resource groups by the resource orchestrator. Therefore, some tenants can acquire more resources than the resource distribution definition by borrowing them from other applications. When the lender needs more resources, it reclaims these borrowed resources from the borrower. This can keep the cluster in high usage status, and maintain consistency between all tenants.

- ▶ Multiple resource scheduling policy (first-in first-out (FIFO)/Fairshare) for each tenant. All tenants partition the resources based on the resource allocation plan.
- ▶ Multi-tenant scheduling with Session scheduler. A consumer can run many applications concurrently, and the session scheduler can schedule resources between applications that belong to one consumer. Therefore, the resource scheduling is hierarchical, as shown in Figure 15.

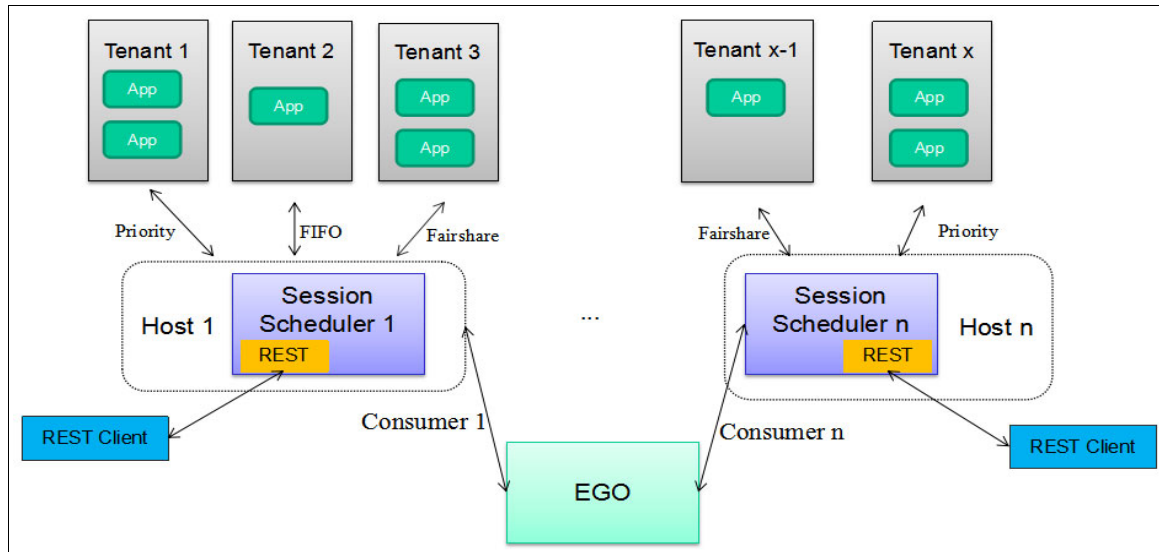


Figure 15 Session scheduler flexible scheduling plans

Comparison with other resource managers

There are several resource managers available for use with Apache Spark, including Spark stand-alone resource manager, Hadoop YARN, Apache Mesos, and IBM Spectrum Computing resource manager. Efficiency, or lack thereof, of the resource manager can have a significant impact on Spark application performance.

To help companies evaluate Spark resource managers, IBM introduced the Spark Multi-User Benchmark version 1 (SMB-1). More information about this new tool to drive Spark performance advancement and adoption can be found on the following website:

<https://www.ibm.com/blogs/systems/new-tool-drive-spark-performance-advancement-adoption/>

Note: To download and try SMB-1, see the following website:

<https://developer.ibm.com/open/spark-benchmarking/>

Graphics processor unit support

Apache Spark allows accelerating applications by using graphics processor units (GPUs) effectively and transparently. This is achieved by introducing a new API in resilient distributed dataset (RDD) implemented in both CUDA or OpenCL.

IBM Spectrum Conductor interfaces with Spark scheduler to ensure that GPU resources are assigned to the applications that can use them. A stage marked with GPU is scheduled to the GPU resource group, where others are scheduled to the default resource group.

Figure 16 presents a high-level overview of the GPU support functionality.

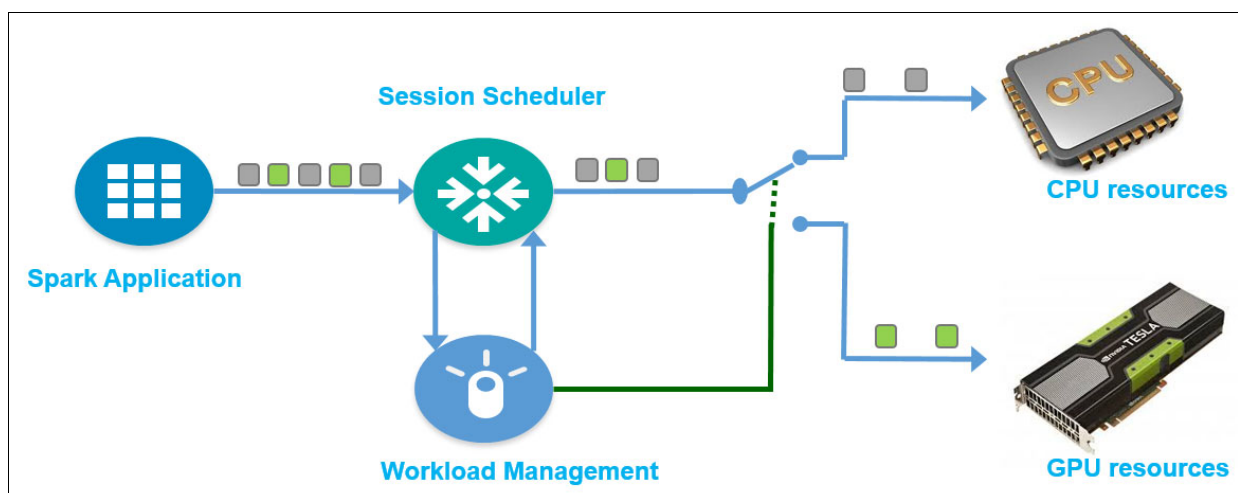


Figure 16 GPU resources support in IBM Spectrum Conductor with Spark

Specified operating environment

The following section provides details about the hardware and software requirements for implementing IBM Spectrum Conductor V2.1 and IBM Spectrum Conductor with Spark V2.1.

Key prerequisites: A distributed computing environment that consists of IBM Power Systems™ servers or x86 servers on supported Linux operating systems, running in containers or on bare metal.

Hardware requirements

IBM Spectrum Conductor V2.1 and IBM Spectrum Conductor with Spark V2.1 are supported on the following hardware:

- ▶ IBM Power Systems servers (IBM POWER8® LE)
- ▶ x86-based servers

Software requirements

IBM Spectrum Conductor V2.1 and IBM Spectrum Conductor with Spark V2.1 are supported on the following operating systems:

- ▶ On x86-based servers, Linux 64-bit:
 - Red Hat Enterprise Linux (RHEL) Application Server (AS) 6.4, or later
 - RHEL AS 7.0, or later up to 7.2
 - SUSE Linux Enterprise Server 11 SP2 and SP3
 - SUSE Linux Enterprise Server 12
 - Ubuntu Server 16.04 Long Term Support (LTS)
- ▶ On IBM Power Systems, IBM POWER® 64-bit LE:
 - RHEL AS 7.1 and 7.2
 - SUSE Linux Enterprise Server 12
 - Ubuntu 16.04 LTS

- ▶ Additionally, IBM Spectrum Conductor V2.1 is supported on the following Microsoft Windows platforms in 64-bit:
 - Windows Server 2012 and 2012 R2:
 - Standard
 - Datacenter
 - Windows Server 2008 and 2008 R2:
 - Standard Edition
 - Enterprise Edition
 - Windows HPC Server 2008 and 2008 R2 SUSE Linux Enterprise Server 11 SP2/SP3
 - Windows 8 and 8.1 SUSE Linux Enterprise Server 12
 - Windows 7 (compute host only)

Important: IBM Spectrum Conductor with Spark V2.1, when running fully integrated for bare metal deployment with the storage solution, is supported on RHEL AS 7.2 only. The program's specifications and specified operating environment information can be found in documentation accompanying the program, if available, such as a readme file, or other information published by IBM, such as an announcement letter.

IBM Spectrum Computing product family

IBM Spectrum Computing is a family of offerings. Individual offerings are members of the family. The renamed offerings are detailed in Table 2.

Table 2 IBM Spectrum Computing family of offerings

Prior IBM Platform Computing name	New IBM Spectrum Computing name
No prior offering	IBM Spectrum Conductor
IBM Platform Conductor for Spark	IBM Spectrum Conductor with Spark
IBM Platform Application Service Controller	Withdrawn
IBM Platform LSF	IBM Spectrum LSF
IBM Platform Symphony - Express, Standard, and Advanced Edition	IBM Spectrum Symphony Express, Standard, and Advanced Edition
IBM Platform Symphony - Desktop Harvesting	IBM Spectrum Symphony - Desktop Harvesting
IBM Platform Symphony - GPU Harvesting	IBM Spectrum Symphony GPU Harvesting
IBM Platform Symphony - Server and VM Harvesting	IBM Spectrum Symphony Server and VM Harvesting
IBM Platform Symphony - Co-Processor Harvesting	IBM Spectrum Symphony Co-Processor Harvesting
IBM Platform Symphony - LSF Standard for Symphony	IBM Spectrum LSF Standard Edition for Symphony

Note: IBM Platform Symphony has been renamed to *IBM Spectrum Symphony*, and is updated to version 7.1.2.

Licensing and ordering the solution

The licensing model is as follows:

- ▶ Per socket
- ▶ Per terabyte (TB)

Contact your IBM sales representative for further information.

To use an existing installation of IBM Spectrum Scale with IBM Spectrum Conductor with Spark, you must meet the following requirements:

- ▶ You must have IBM Spectrum Scale Standard Edition or Advanced Edition versions 4.1.1, 4.2.1, or later installed.
- ▶ You must have an existing IBM Spectrum Scale FPO license.

The solution's availability date was 23 June, 2016. Table 3 shows the offering program numbers.

Table 3 Solutions program numbers

Program number	VRM	Program name
5725-Y98	2.1.0	IBM Spectrum Conductor
5725-Y38	2.1.0	IBM Spectrum Conductor with Spark

IBM Spectrum Conductor with Spark binaries are available in two versions:

- ▶ IBM Spectrum Conductor with Spark, which contains the core components of the solution. Use this package if the hosts are already deployed with a supported operating system.
- ▶ IBM Spectrum Conductor with Spark with integrated infrastructure, which contains the bare metal installation package in addition to the standard components. Use this package to deploy the solution on a bare metal cluster.

Authorized customers can download the installation packages, and other IBM Spectrum Conductor with Spark packages, on IBM Passport Advantage®.

Tip: If you install IBM Spectrum Conductor with Spark as an integrated infrastructure, IBM Spectrum Scale is automatically deployed in your cluster. Therefore, you are not required to install it separately.

IBM Spectrum Conductor with Spark is available for trial from the following website:

<http://www.ibm.com/developerworks/servicemanagement/tc/pcs/downloads.html>

Additional references

This section provides websites for information about IBM Spectrum Computing, IBM Spectrum Storage solutions, and more:

- ▶ IBM Spectrum LSF
<http://www.ibm.com/spectrum-lsf>
- ▶ IBM Spectrum Symphony
<http://www.ibm.com/spectrum-symphony>
- ▶ IBM Spectrum Conductor
<http://www.ibm.com/spectrum-conductor>
- ▶ IBM Spectrum Conductor with Spark Wiki
<http://ibm.co/22os7I5>
- ▶ IBM Spectrum Computing
<http://www.ibm.com/systems/spectrum-computing/>
- ▶ IBM Spectrum Storage
<http://www.ibm.com/systems/storage/spectrum/>
- ▶ IBM software define infrastructure blog
<https://www.ibm.com/blogs/systems/topics/storage/software-defined-infrastructure/>

The following IBM Spectrum Conductor documentation publications are delivered with the program, and are available from the IBM Knowledge Center on the following website:

<https://www.ibm.com/support/knowledgecenter/SSZU2E>

- ▶ Quick Start Guide for IBM Spectrum Conductor V2.1
- ▶ Quick Start Guide for IBM Spectrum Conductor with Spark V2.1
- ▶ Release Notes for IBM Spectrum Conductor V2.1
- ▶ Release Notes for IBM Spectrum Conductor with Spark V2.1
- ▶ Install Guide for IBM Spectrum Conductor V2.1
- ▶ Install Guide for IBM Spectrum Conductor with Spark V2.1

Authors

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, Poughkeepsie Center.

Dino Quintero is a Complex Solutions Project Leader and an IBM Level 3 Certified Senior IT Specialist with the ITSO in Poughkeepsie, New York. His areas of expertise include enterprise continuous availability, enterprise systems management, system virtualization, technical computing, and clustering solutions. He is an Open Group Distinguished IT Specialist. Dino holds a Master of Computing Information Systems degree and a Bachelor of Science degree in Computer Science from Marist College.

Nicolas Joly is a pre-sales architect with IBM Systems in New York City, New York. His areas of expertise include software defined infrastructure, analytics solutions, storage, technical computing, and clustering solutions. He works with major customers in the finance and telecommunication industry. Before joining IBM US, Nicolas worked for IBM France, where he worked as a technical sales specialist for analytics and technical computing solutions. Nicolas holds a Master degree in Computer Science with a major in parallel and distributed computing from Institut Polytechnique de Bordeaux (ENSEIRB-MATMECA), France.

Thanks to the following people for their contributions to this project:

Louise Westoby, Scott Campbell, Daniel de Souza Casali
IBM US

Now you can become a published author, too

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time. Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run 2 - 6 weeks in length, and you can participate either in person or as a remote resident working from your home base.

Learn more about the residency program, browse the residency index, and apply online:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:
<http://twitter.com/ibmredbooks>
- ▶ Look for us on LinkedIn:
<http://www.linkedin.com/groups?home=&gid=2130806>
- ▶ Explore new IBM Redbooks® publications, residencies, and workshops with the IBM Redbooks weekly newsletter:
<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:
<http://www.redbooks.ibm.com/rss.html>

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.


COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

IBM®	IBM Spectrum Symphony™	POWER8®
IBM Spectrum™	LSF®	Redbooks®
IBM Spectrum Conductor™	Passport Advantage®	Redpaper™
IBM Spectrum Scale™	POWER®	Redbooks (logo)  ®
IBM Spectrum Storage™	Power Systems™	Symphony®

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.



REDP-5379-00

ISBN 0738455385

Printed in U.S.A.

Get connected

