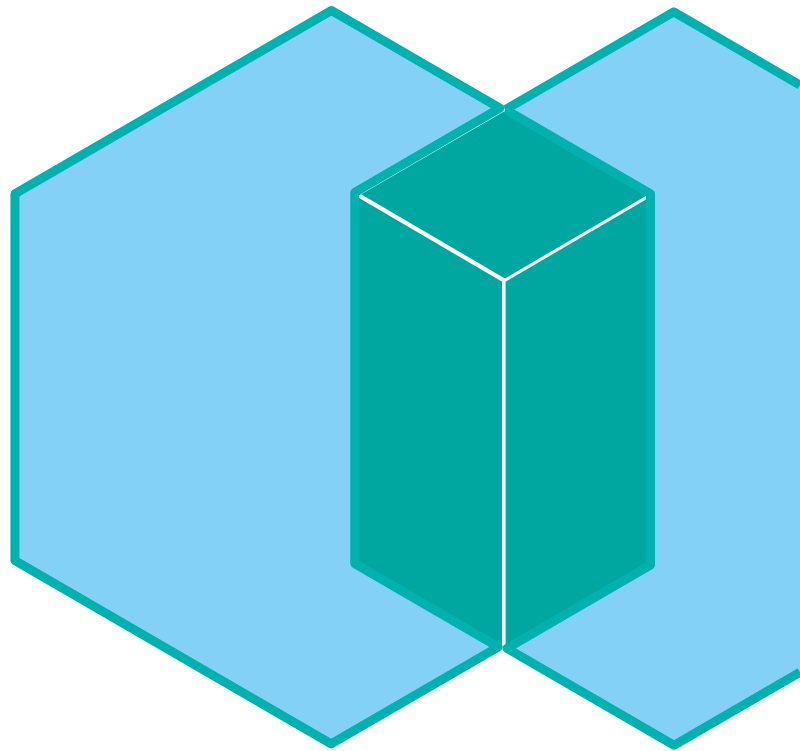# Make Smarter Business with IBM Big Data Analytic

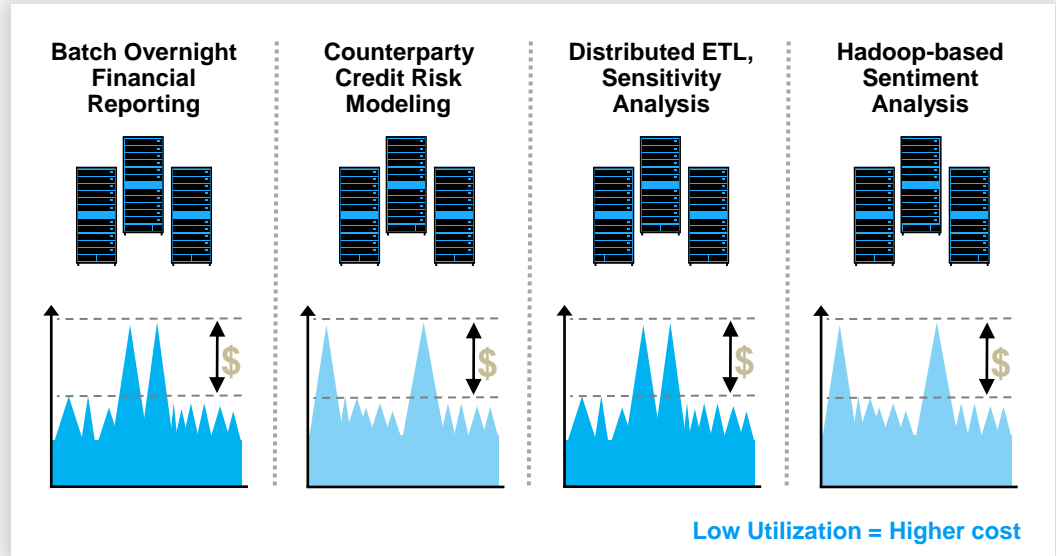The digital world is generating oceans of data

Organizations need the ability to confidently navigate and
rapidly explore these oceans to find valuable insights

Oceans of data must be cost and performance optimized,
efficiently stored, managed, and protected as well as
accessible to the right applications at the right time

# Creating silos to analyze all data is costly & inefficient

**New modern scale-out applications** and **frameworks** …

…lead to **costly, complex, siloed, & underutilized** infrastructure and replicated data



Batch Overnight Financial Reporting

Counterparty Credit Risk Modeling

Distributed ETL, Sensitivity Analysis

Hadoop-based Sentiment Analysis

**Low Utilization = Higher cost**

**Low Utilization + Poor Performance = Higher cost**

# IBM Spectrum Computing

*Powerful. Comprehensive. Intuitive*.

**High Performance Computing**
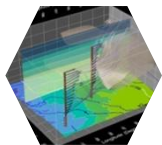
**IBM Spectrum LSF**

Scalable, comprehensive workload management **accelerates throughput up to 150X** for simulation, design & research

**Analytics and Big Data**

**IBM Spectrum Symphony**

Deliver infrastructure as a shared service with low-latency workload and resource management for up to **100x faster analytics**

**New Generation**

**IBM Spectrum Conductor**

**Accelerate results up to 60%** by efficiently analyzing, accessing and protecting data on an integrated application & data-optimized platform.

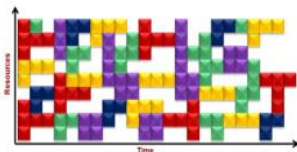# IBM Spectrum Computing for Next Generation Workloads

**▪*Traditional***

**▪ *Next Generation Workloads***

**▪Application Examples**
- Simulation
- Analysis
- Design
- Big data

**▪Application**

**▪IT constrained**
- Long wait times
- Low utilization
- Data access bottlenecks
- IT Sprawl

**SILO Compute / Storage**

**▪Simulation and Modeling** **▪Big Data** **▪Analytics** **▪Long Running Services**

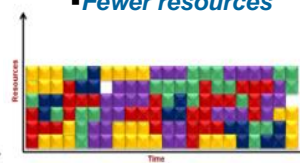**▪IBM Software Defined Infrastructure**

**Cloud**

**Compute / Storage**

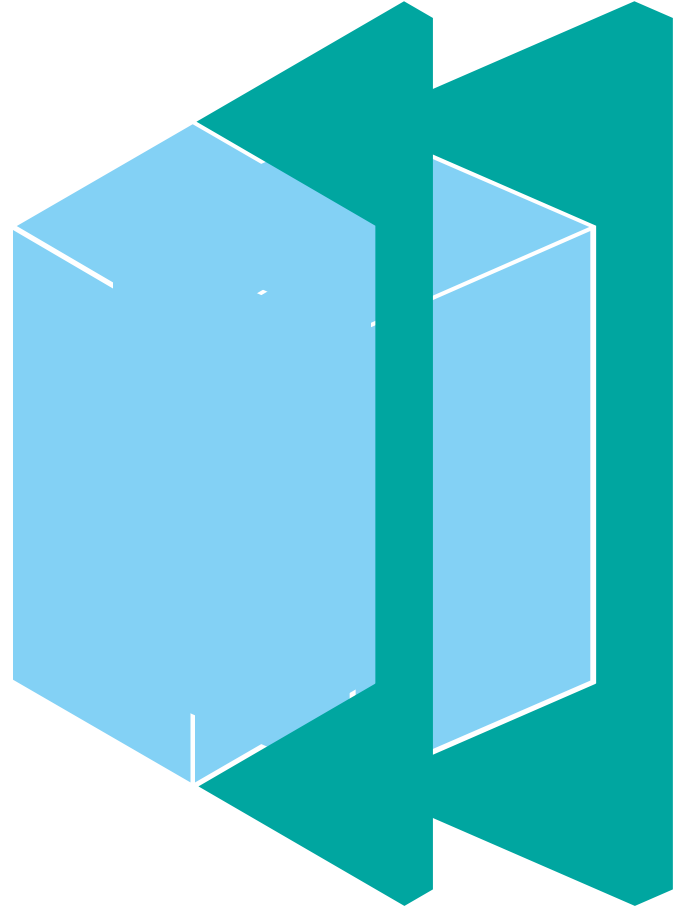**▪*Faster results*** **▪*Fewer resources***

**▪Make lots of computers look like one**

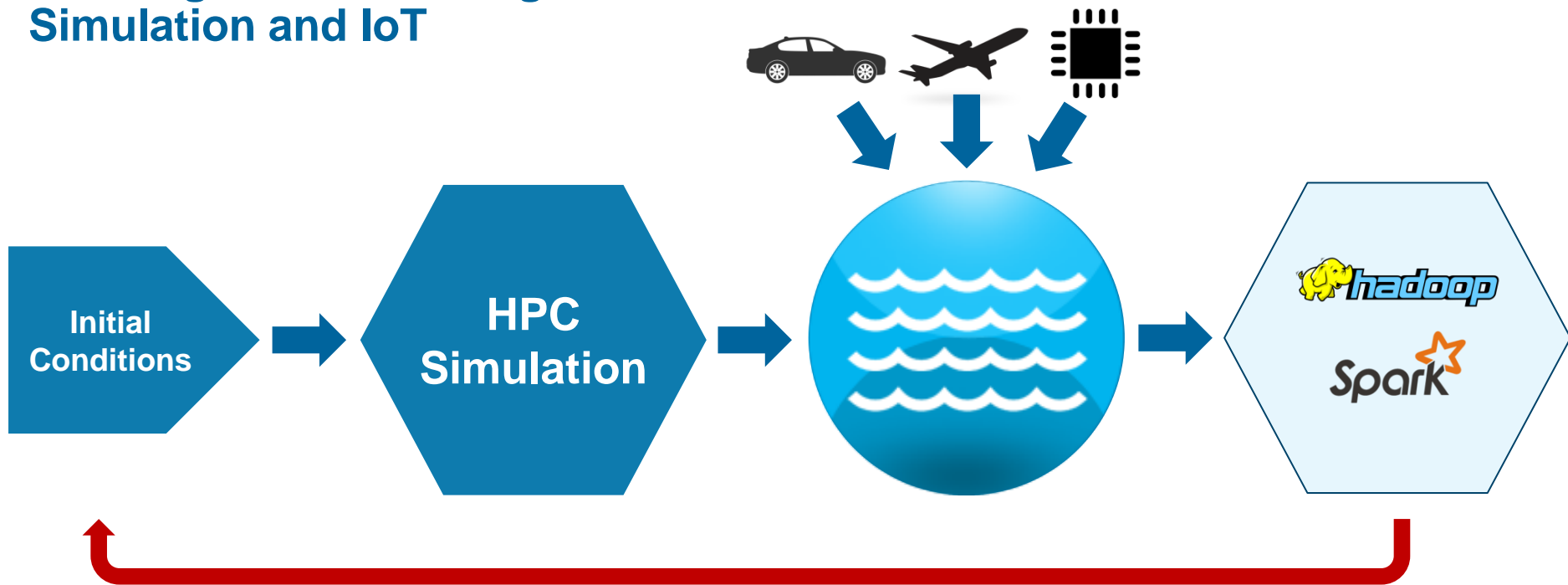**▪Prioritized matching of supply with demand**

**▪Benefits**
- High utilization
- Throughput
- Performance
- Prioritization
- Reduced cost

# Introducing IBM Spectrum Computing with Spark

# Converged HPC and Big Data: Simulation and IoT

**Initial Conditions** → **HPC Simulation** → [waves icon] → hadoop / Spark

Combine Sensor/IoT Data with Simulation Data to generate more accurate simulations to improve products

# Organizations' Spark Deployment Challenges

Integration of Spark into existing environments

Investment in new expertise, tools and workflows

Proliferation of numerous ad hoc Spark clusters ("Spark silos")

Fast-moving Spark lifecycles

# IBM Spectrum Computing with Spark
## The most complete enterprise-grade solution for Apache Spark

**Competitive advantage from faster time to results for big data analytics**

- Spark multitenancy allows Spark jobs to run on resources that would otherwise be idle
- Proven, high-efficiency resource scheduling technology

**Simplified deployment and management**

- End-to-end integrated solution incorporating resource scheduling, data management, monitoring, alerting, reporting and diagnostics
- Easier Spark lifecycle management from multitenancy support for simultaneous running of multiple simultaneous Spark versions

**Lower infrastructure costs from improved efficiency**

- Shared infrastructure maximizes efficient use of resources

*Spark*

Workload & Resource Management

Data Management

Infrastructure Management

Support & Services
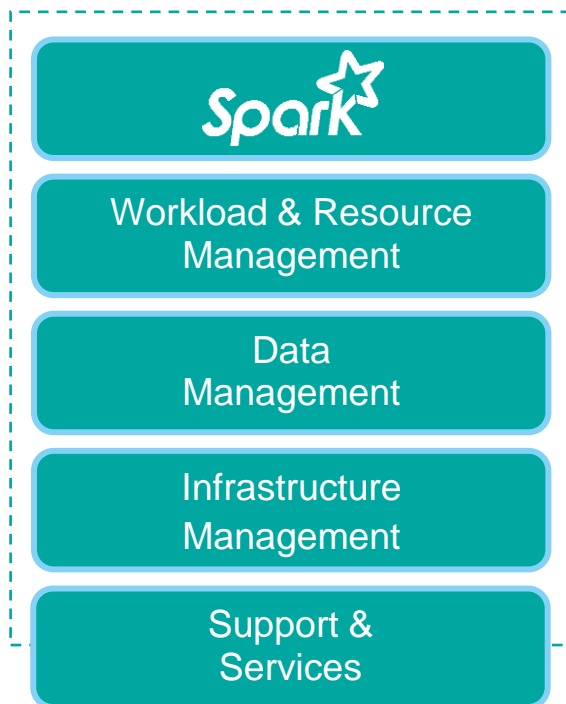
# Competitive advantage through faster, more predictable analytics

## Throughput: 41% over YARN; **57%** over Mesos

| | Spectrum Computing with Spark | Spark / YARN | Spark / Mesos |
|---|---|---|---|
| *When minutes count* | 10 minutes | 14.1 minutes | 15.7 minutes |
| *At quarter-end* | 80 hours | 112.8 hours | 125.6 hours |
| *Product development* | 26 weeks | 36.7 weeks | 40.8 weeks |

## Predictability: longest job duration compared with median *(lower is better)*

| Spectrum Computing with Spark | Spark / YARN | Spark / Mesos |
|---|---|---|
| 1.51X | 1.62X | 66.32X |

Source: STAC Report: Spark Resource Managers, Phase 1 (March 28, 2016)

*Note: IBM is an active contributor in the Mesos community, helping to advance its capabilities and integration with IBM solutions*

# Scale Out with Confidence across the Entire Organization

**Based on established technology**

**Scalability proven in some of the world's most demanding customer environments**

- Production customers with over **100,000 cores**
- **Thousands of nodes** deployed
- **Hundreds of applications** on a single cluster
- Over **1 billion tasks** per day
- Multiple sessions running over **1 million tasks** per session
- Fully resilient with HA recovery
- Continued investment in roadmap and support



## Citi

- **100X performance improvement** combined with on-demand access to compute power drastically speeds time to results

- Resources used more effectively with **hardware utilization increasing from 20% to 80%**

# Faster Time to Results and Reduced Infrastructure Costs

## Multi-Tenancy – Shared Infrastructure
*Dynamic Response to Changing Workload Demands*



**Workload** (y-axis) / **Time** (x-axis)
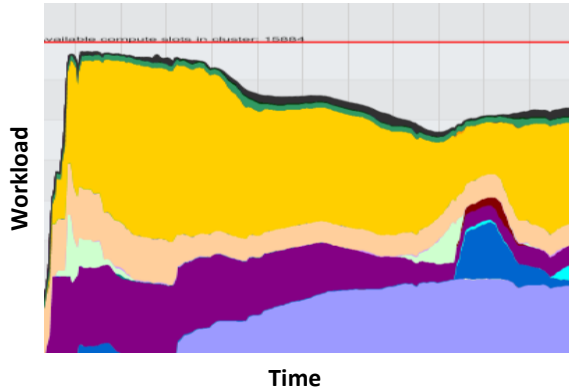
### 43% reduction in required infrastructure*

- **Siloed infrastructure**
  **28,000** cores would be required to handle peak capacity
- **Shared infrastructure**
  **16,000** total cores required

> * Based on actual customer data

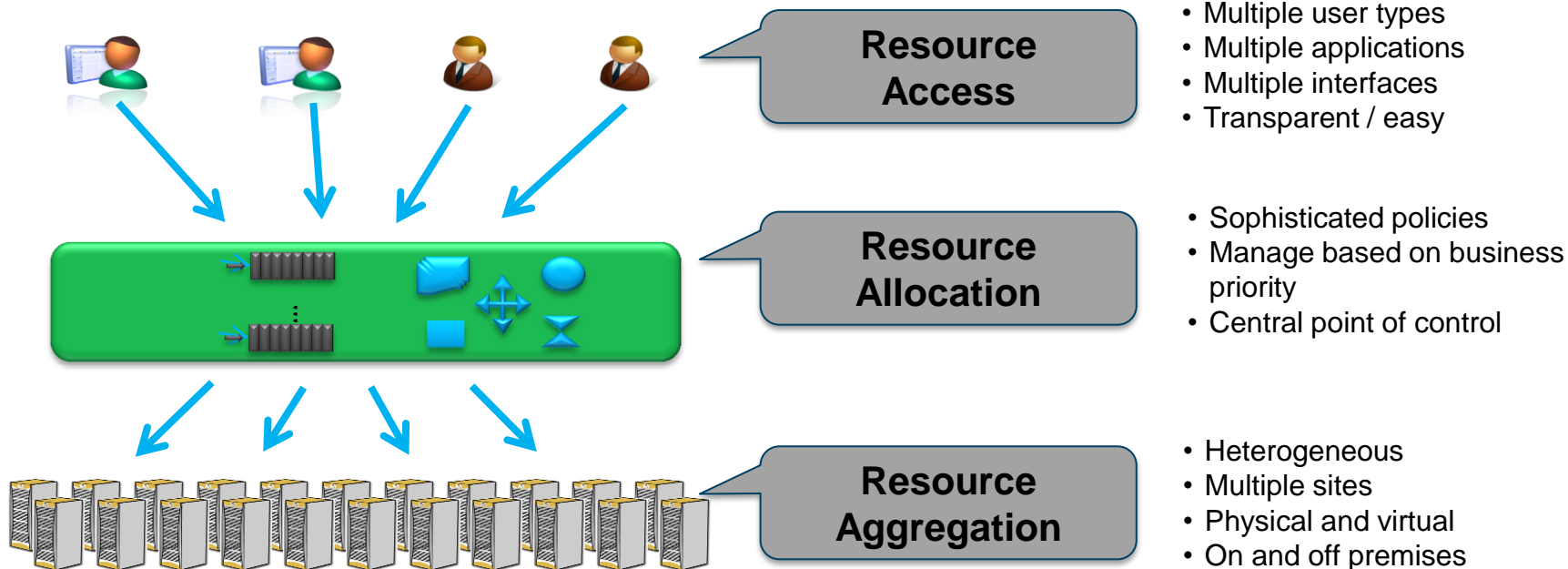**If application spikes above its silo cluster?**
- Performance slows way down waiting for resource allocation
- More hardware is purchased to manage peak in the future for each cluster

**With Spectrum Conductor with Spark**
- Application spikes are managed **dynamically** across application boundaries

# IBM Spectrum Computing Core Technology Concept

**Use** and **Manage** 1000s of computers as one



**Resource Access**

- Multiple user types
- Multiple applications
- Multiple interfaces
- Transparent / easy

**Resource Allocation**

- Sophisticated policies
- Manage based on business priority
- Central point of control

**Resource Aggregation**

- Heterogeneous
- Multiple sites
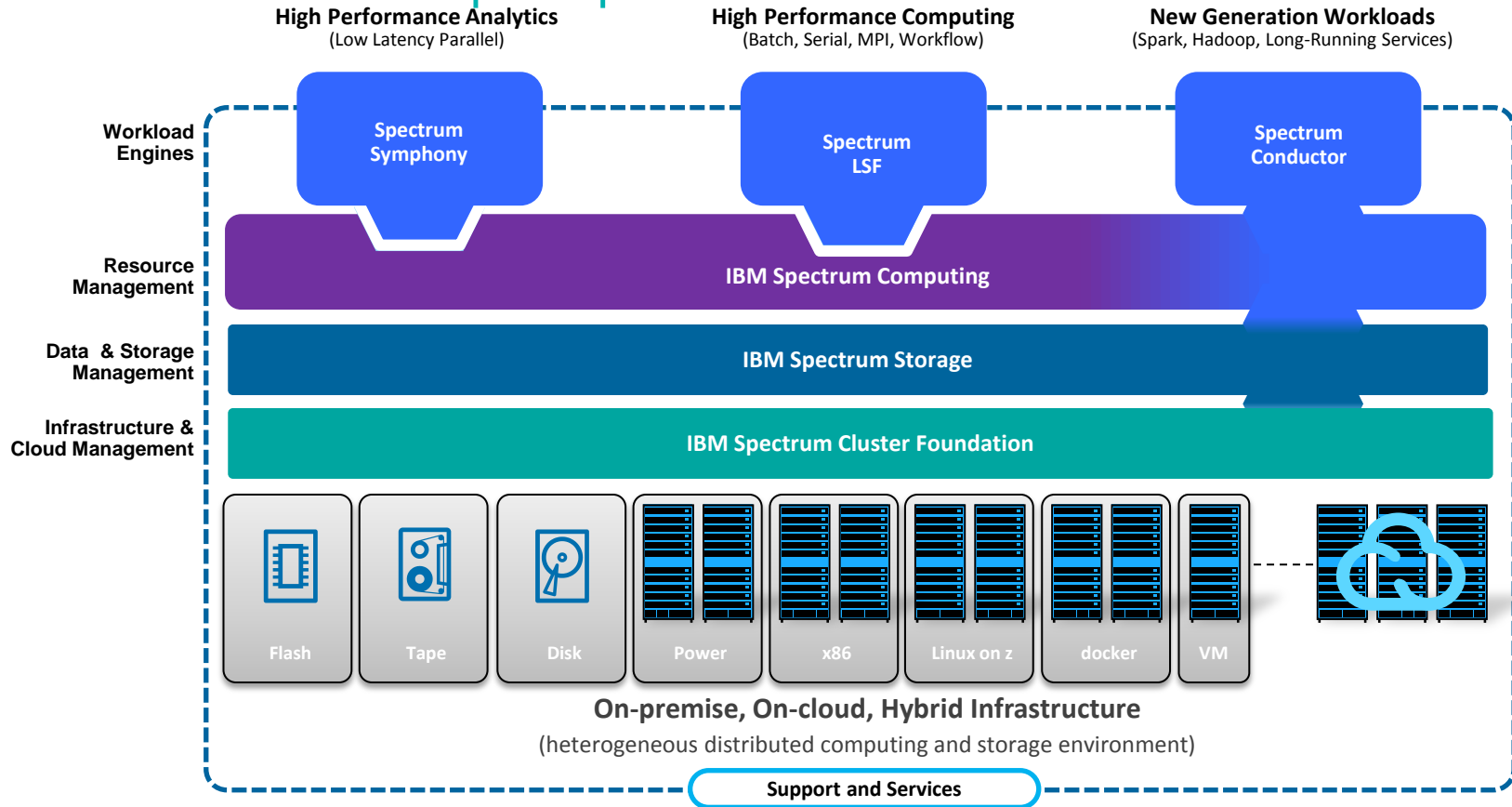- Physical and virtual
- On and off premises
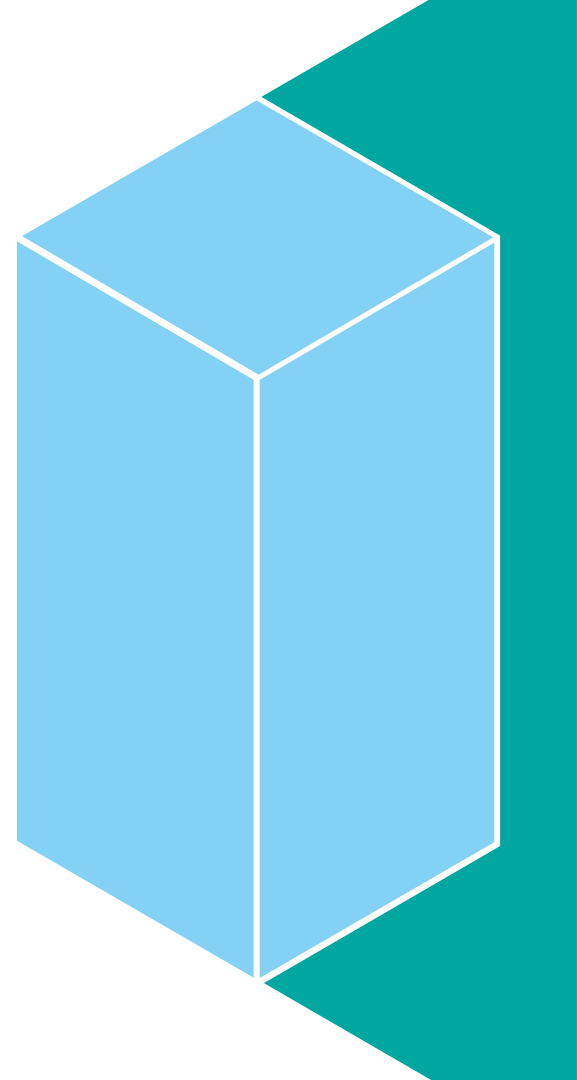
# Supporting Complex Workflows

## Different workload engines can be combined in a single computational workflow
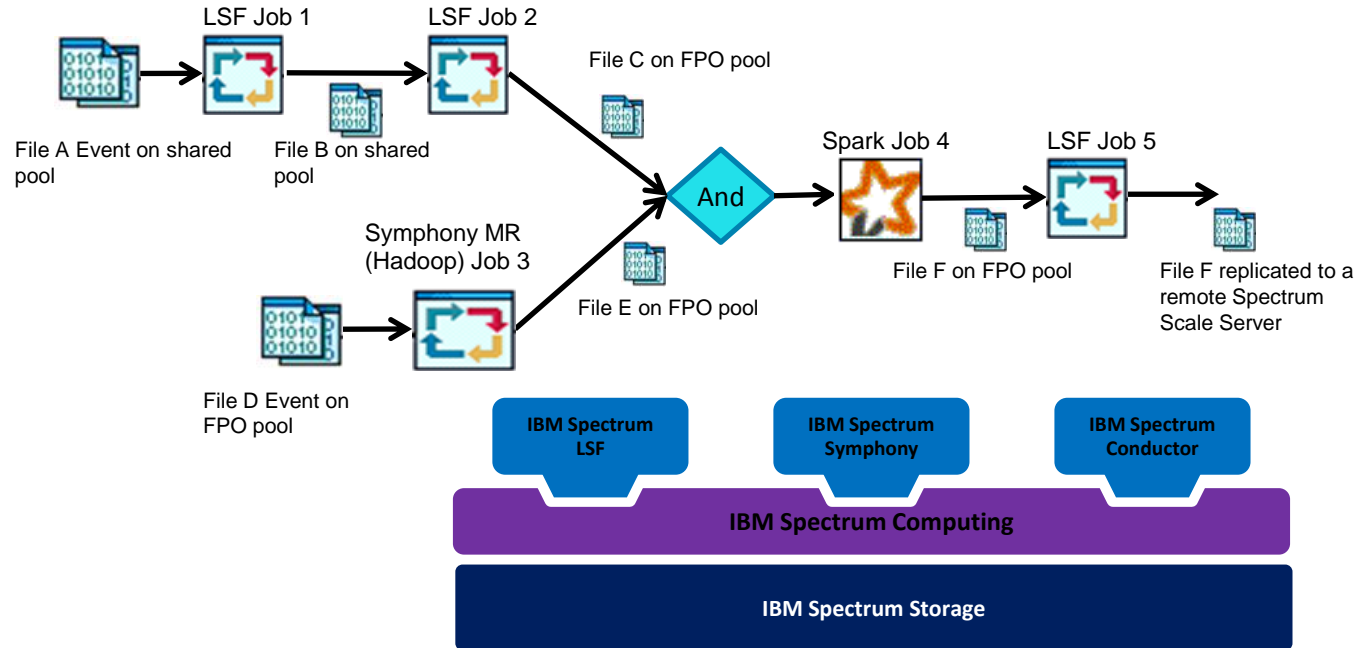
# Part of a complete portfolio of software-defined solutions
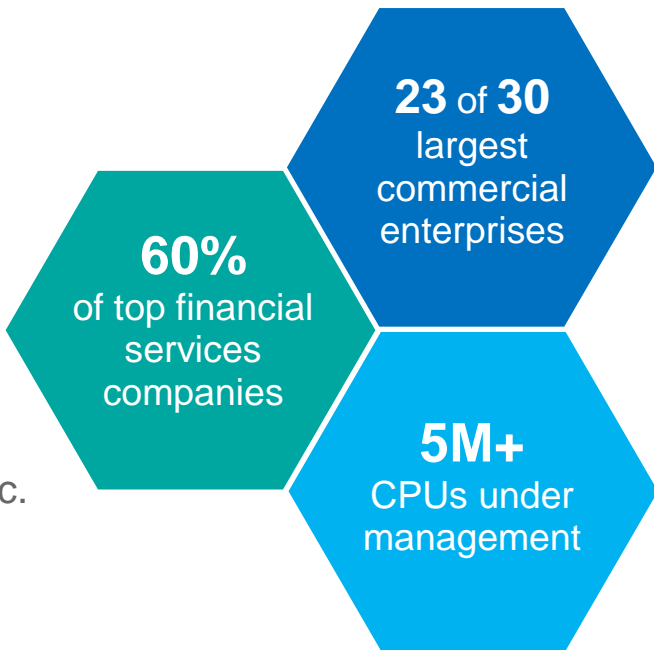
# Summary

# Software Defined Computing: IBM Spectrum Computing

*The leader in workload, resource and infrastructure management solutions*

- 20+ years managing distributed scale-out systems with 2000+ customers in many industries

- Market leading scheduling engine with intelligent infrastructure automation

- Unmatched scalability (small clusters to global grids) and production-proven reliability

- Heterogeneous –optimizes utilization of Power and System z plus 3rd party systems, virtual and bare metal, accelerators / GPUs, etc.

- Shared services for both compute and data intensive workloads

- On-premises, in the cloud and end-to-end hybrid cloud solutions

**23** of **30** largest commercial enterprises

**60%** of top financial services companies

**5M+** CPUs under management

IBM Systems

# IBM Software-Defined Infrastructure

*Delivering value across a multitude of industries*



## Red Bull Racing

- Designs and quickly executes multiple complex, interdependent simulations and analysis

- Leverages **20% increase in performance and throughput** to run more simulations is less time on less infrastructure



## Cypress Semiconductor

- Eliminated data access bottlenecks and has **increased performance 10x** using the same hardware

- Continuous data availability across hardware outages



## Caris Life Sciences

- Correlates molecular data for 65,000 patients and supporting 7,000 oncologists worldwide

- **Manages nearly a terabyte of data** per patient enabling precision cancer treatment



## Citi

- **100X performance improvement** combined with on-demand access to compute power drastically speeds time to results

- Computing resources used more effectively with **hardware utilization increasing from 20% to 80%**

# IBM Spectrum Computing Help …

### 1 Accelerate results

Run Spark natively on a shared infrastructure *without* the dependency of Hadoop.  Increase throughput up to 60%, accelerating time to results.

### 2 Reduce administration costs

Proven architecture at extreme scale, with enterprise class workload management, monitoring, reporting, and security capabilities.

### 3 Increase resource utilization

Fine grain, dynamic allocation of resources maximizes efficiency of multiple Spark instances and versions sharing a common resource pool to eliminate cluster sprawl.

### 4 End-to-end enterprise class

A tightly integrated offering that combines the IBM supported Spark distribution with workload, resource and data management as well as IBM support and services.

# How can semiconductor firms cut design costs and shorten time-to-market?

## IBM EDA

designs state-of-the-art IBM POWER processors

Aimed to deliver higher-quality chip designs to fabrication partners faster, cutting costs on both sides

Deployed IBM Platform LSF grid scheduling software to run complex simulations more efficiently

Achieves **multi-million dollar savings** annually by finding and fixing bugs before sending chip designs to fabrication

## 10%

**10% higher utilization** of compute resources enables EDA to **do more with less** hardware investment

## 10x

**10x faster scheduling** saves staff time and **accelerates** the delivery of high-quality processors to the market

Original Schedule

Smarter Schedule

Time

Resources

Time

Resources

Processing Schedule

# Thank You

For more information:
http://www.ibm.com/systems/spectrum-computing/products/conductor/

DCP03273USEN-01

**IBM** Systems