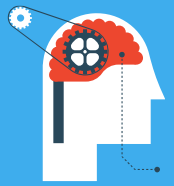


Spectrum Scale Expert Talks

Episode 3:
Spectrum Scale Strategy

Ted Hoover
Program Director Spectrum Scale Development

Wayne Sawdon
CTO for Spectrum Scale and ESS

Join our conversation:
www.spectrumscaleug.org/join

Show notes:
www.spectrumscaleug.org/experttalks

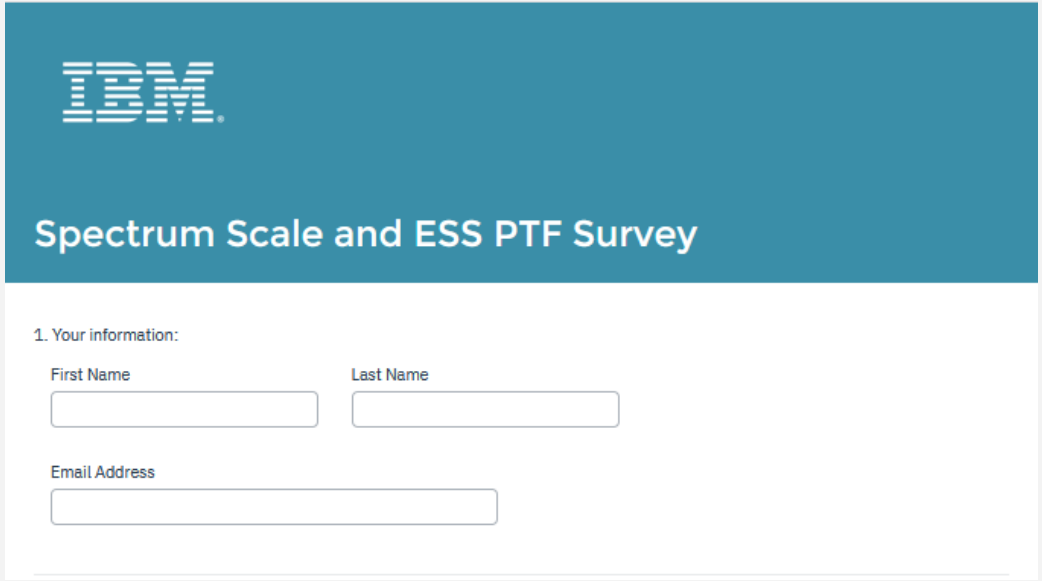
IBM

Survey on Spectrum Scale and ESS PTF Frequency

IBM Spectrum Scale

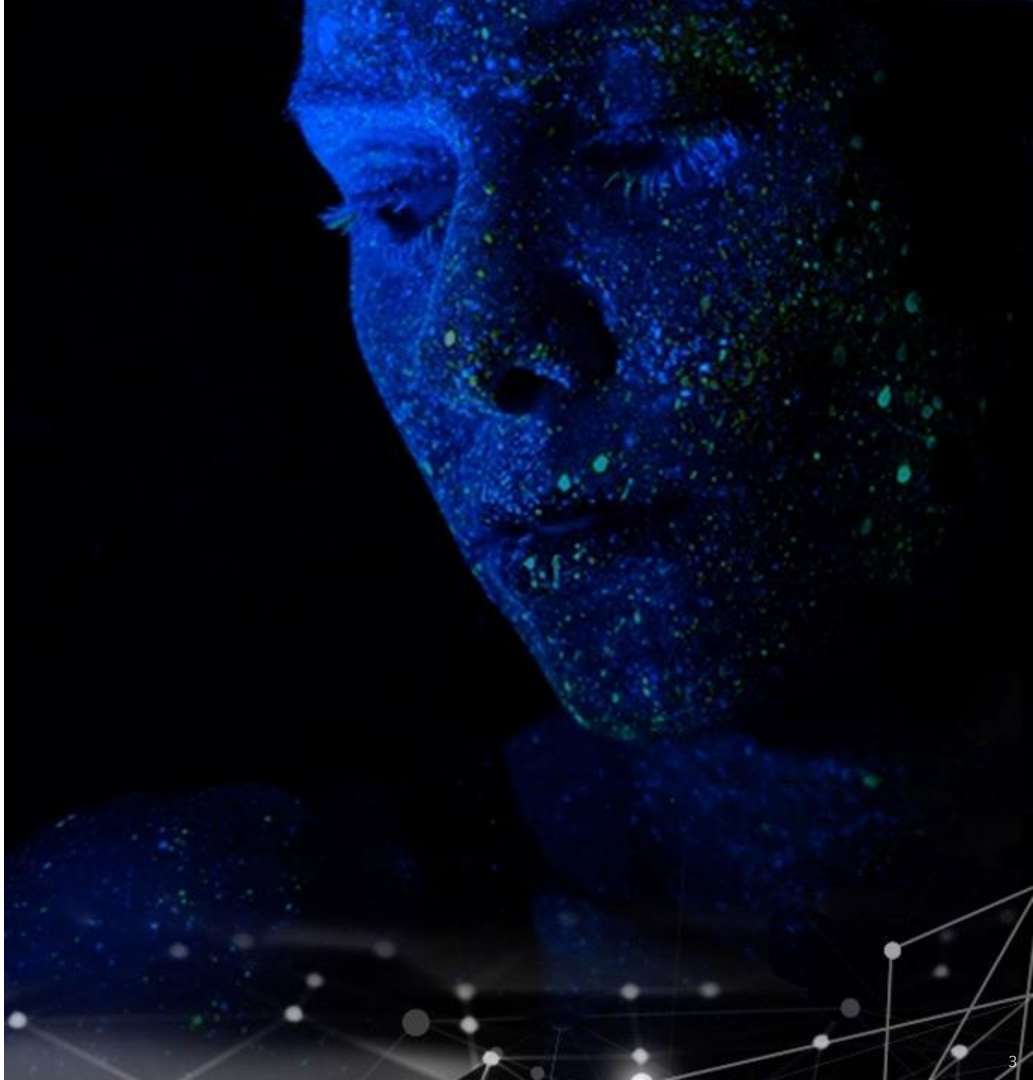
Please share your feedback:

https://www.surveygizmo.com/s3/5727746/47520248d614



The first 20+ years

A history of High Performance Storage from supercomputers in the national labs to today's data-intensive workflows and analytics across commercial markets



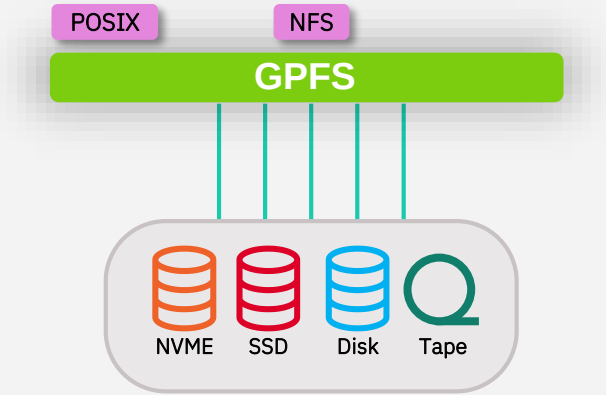
GPFS has evolved …

IBM Spectrum Scale

- In 1993: Started as “**Tiger Shark**” research project at IBM Research Almaden as high performance file system for accessing and processing multimedia data
- In 1998: Grew up as General Parallel File System (GPFS) to power the world’s largest supercomputers
- In 2014: Rebranded as IBM Spectrum Scale

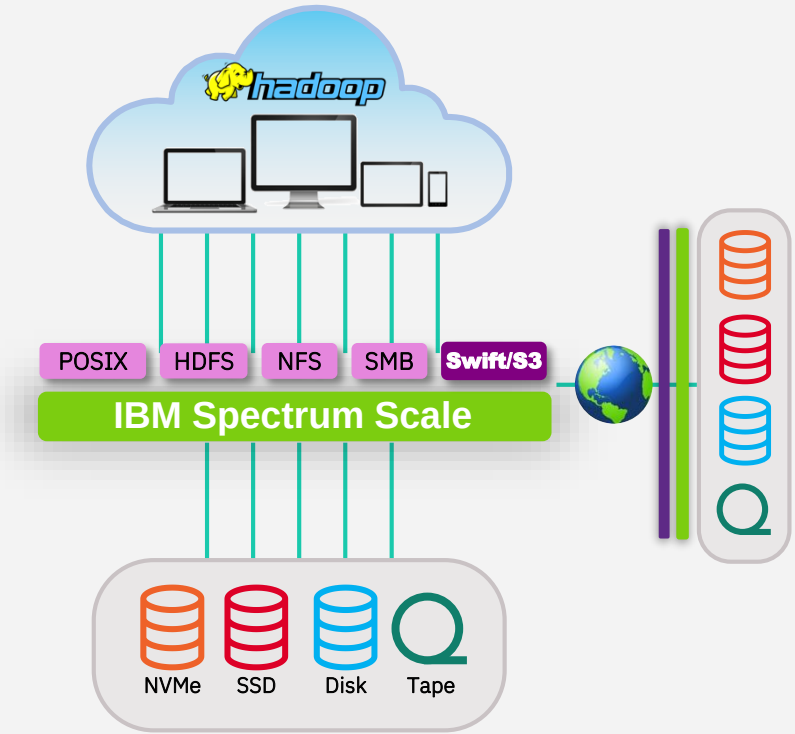
Although the workload has remained the same – high performance analytics on vast quantities of unstructured data – the product features are more focused on commercial markets.

This focus continues to evolve to today’s cloud and container markets.



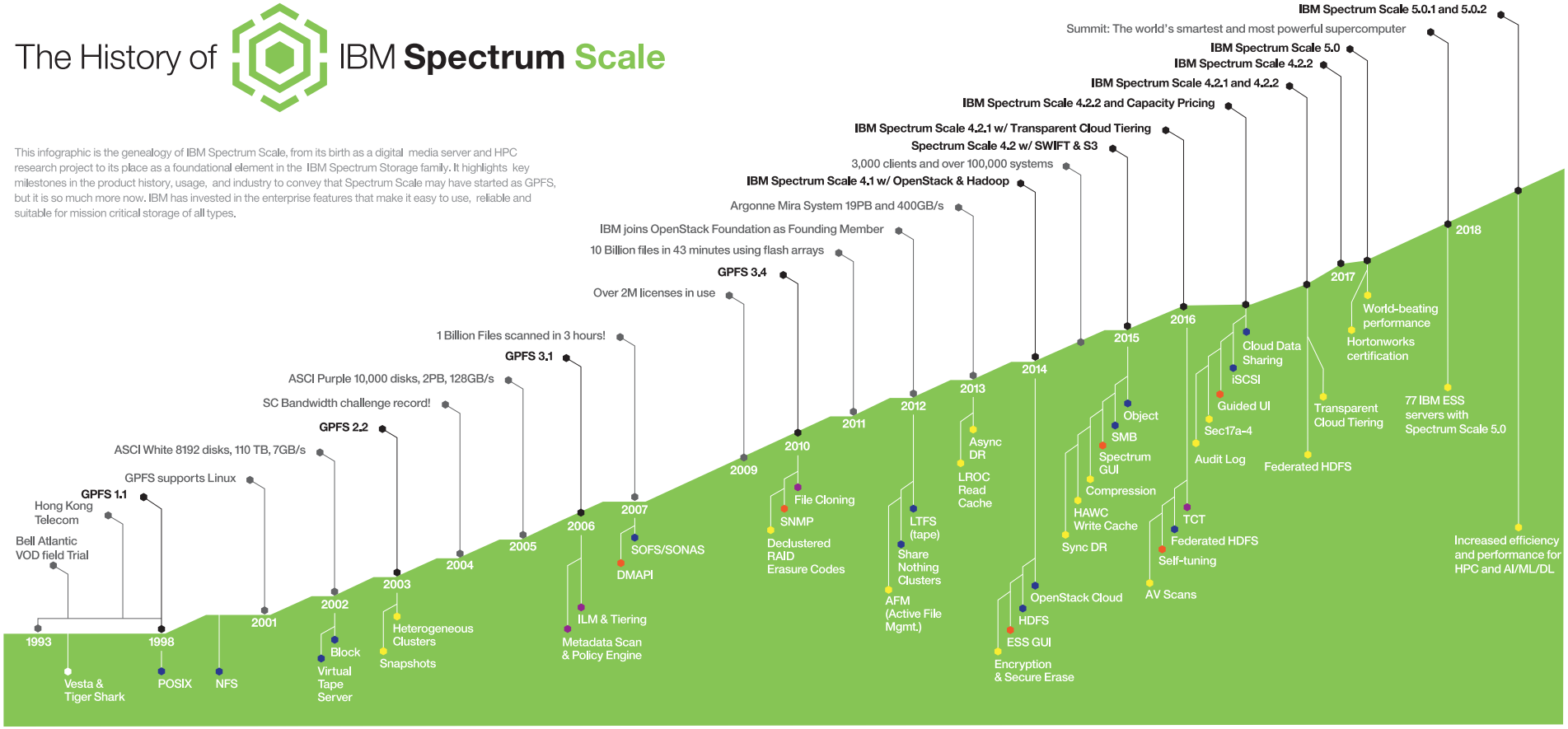
IBM Spectrum Scale

- GPFS is known for scale-out high performance on the world's largest supercomputers…
- BUT: If you still just think GPFS, you miss:
  - Support for workflows which for example inject data via object, analyze results via Hadoop/Spark and view results via POSIX
  - Storing and accessing large and small objects (S3 and Swift) with low latency
  - Storing and starting OpenStack VMs without copying them from object storage to local file system
  - Common namespace between Spectrum Scale clusters on-prem and in the cloud
  - Namespace includes Data Management to automatically destage cold data to on premise or off premise tape or object storage
  - GUI , REST API, Grafana Bridge
  - HA, DR, Real time Audit & Security
  - And much, much more



The History of IBM Spectrum Scale

This infographic is the genealogy of IBM Spectrum Scale, from its birth as a digital media server and HPC research project to its place as a foundational element in the IBM Spectrum Storage family. It highlights key milestones in the product history, usage, and industry to convey that Spectrum Scale may have started as GPFS, but it is so much more now. IBM has invested in the enterprise features that make it easy to use, reliable and suitable for mission critical storage of all types.



The History of IBM Spectrum Scale

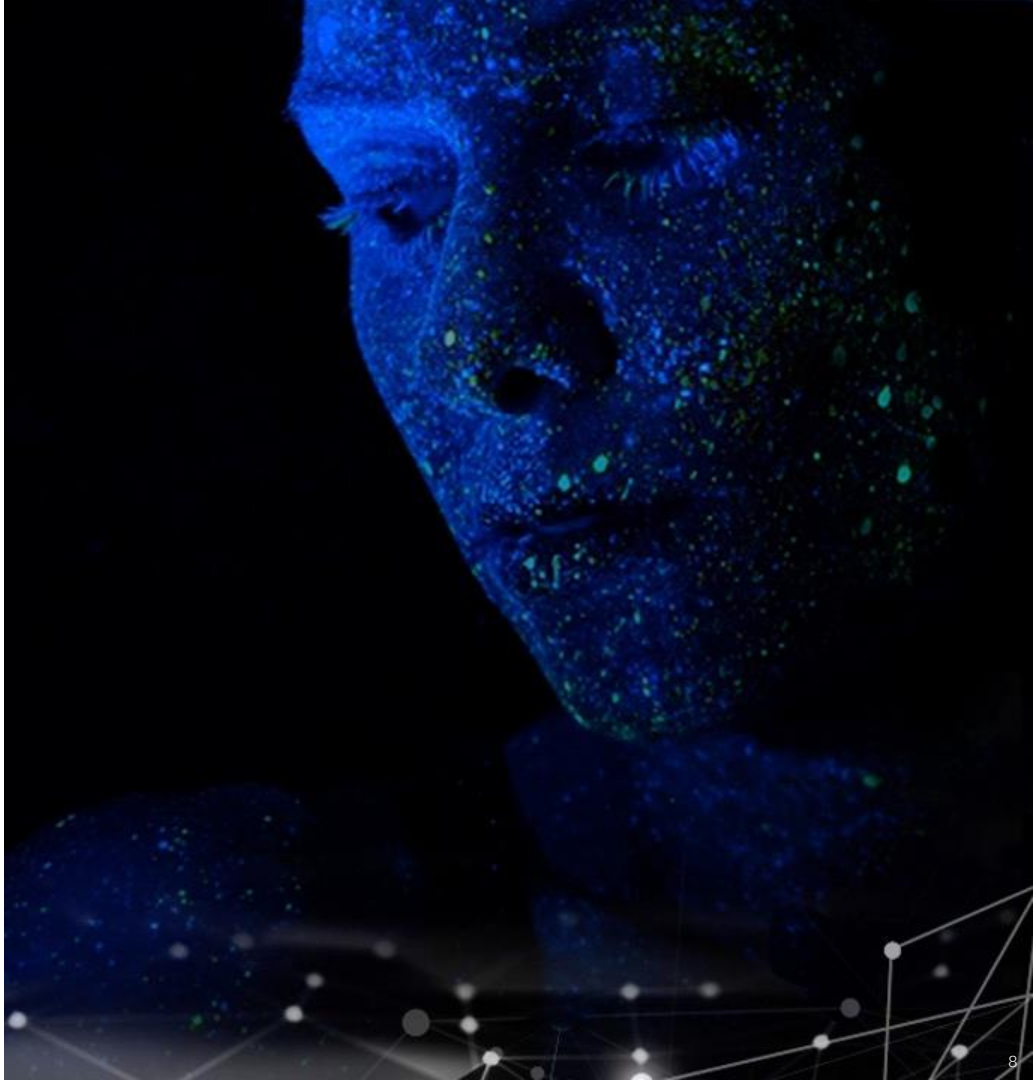
57 of the Global 100 run IBM Spectrum Scale

9 of the top 10 automobile manufacturers
9 of the top 10 investment banks
18 of the top 25 banks
8 of the top 10 global retailers
4 of the top 5 insurance companies

High performance analytics on vast quantities of unstructured data

Strategic Trends

- Connected Clouds
- Dev Ops
- Inescapable AI
- AI Data Management Challenges
- Security
- Performance



Hybrid multicloud is the platform

Companies average almost

5

private and public clouds

80%

of companies moved their applications or data from public clouds in 2018

Reasons to migrate from public cloud

- Security
- Performance
- Cost
- Control

IDC Survey

85%

of companies operate in a hybrid multicloud environment today

98%

of companies will be hybrid multicloud in three years

IDC; IBM IBV C-Suite Study; Rightscale

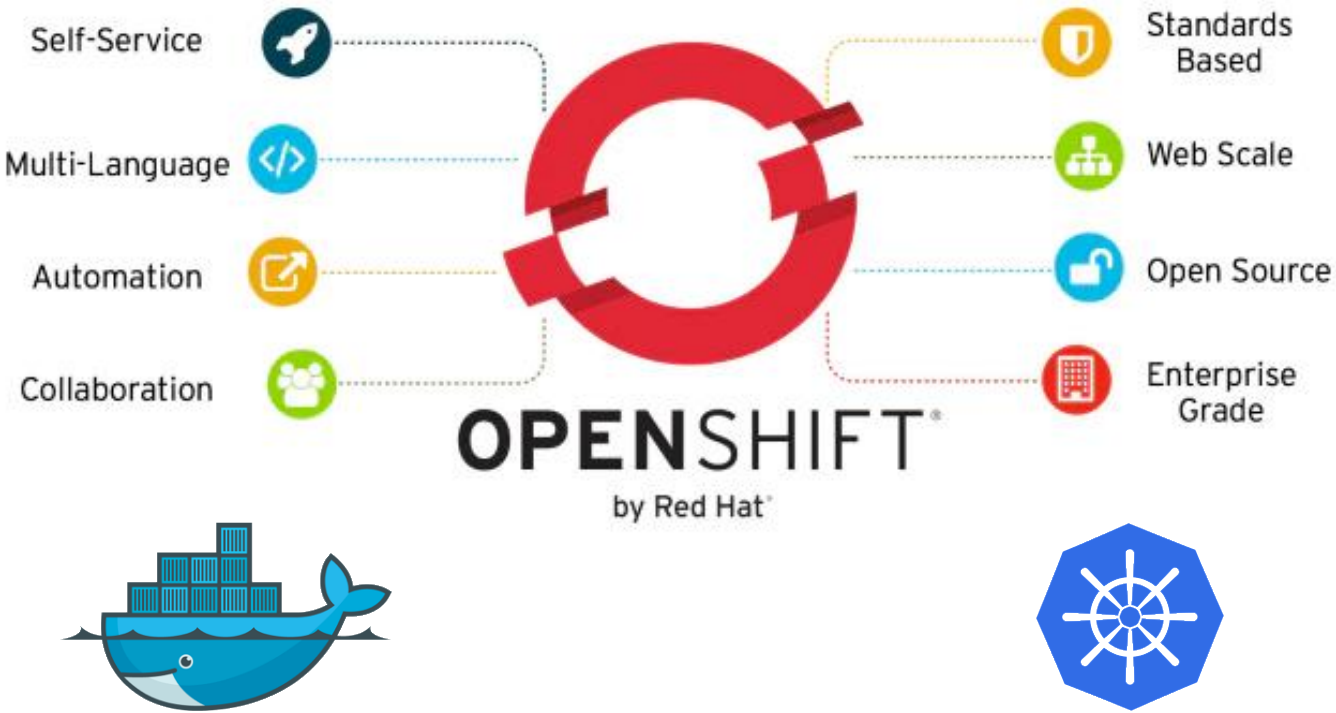
Source: IDC's Cloud and AI Adoption Survey, January 2018

Two simultaneous evolutions are taking shape in the data center today

1. Hybrid multicloud usage
2. Taking advantage of more data for competitive advantage

IBM's ACQUISITION OF RED HAT IN JULY 2019 COMPLETELY CHANGED THE CLOUD LANDSCAPE TO BECOME THE WORLD'S #1 HYBRID MULTI-CLOUD PROVIDER.

The shift to containers



Evolving Storage Market

Traditional Storage

- Deliver underlying infrastructure needs to support enterprise requirements.
- Centralized administration for organization.

Examples: DS8900, FlashSystem, IBM Spectrum Scale

Container-Ready Storage

- Leverage existing investments in traditional storage to support container deployments.
- Allows use of snapshots, clones, and replication but doesn't take advantage of container framework and related benefits.
- Not optimized for Kubernetes so can be a bottleneck to achieving increased agility and elasticity.

Examples: DS8900, FlashSystem, IBM Spectrum Scale

Container-Native Storage

- Storage deployed inside containers with enterprise level data management services to support mission critical applications deployed in containers.
- Direct attach and external storage support varying performance and capacity needs.
- Kubernetes control plane allows self service capabilities driving higher levels of efficiencies.
- Future: Cloud Native Spectrum Scale

Cloud Native Storage

Goal: Deliver High Performance File Services to Containerized Application Workloads

Support Workloads that Require High Performance File Services

- Analytics & Cognitive
- High Performance Computing
- AI Data Pipeline

Support the Workload Ecosystem in the Cloud

- Containerized Applications, Storage
- Ephemeral and Persistent Storage Volumes

Flexible Deployment

- Dynamic Provisioning, Configuration, Upgrade

Support for Multiple Clouds

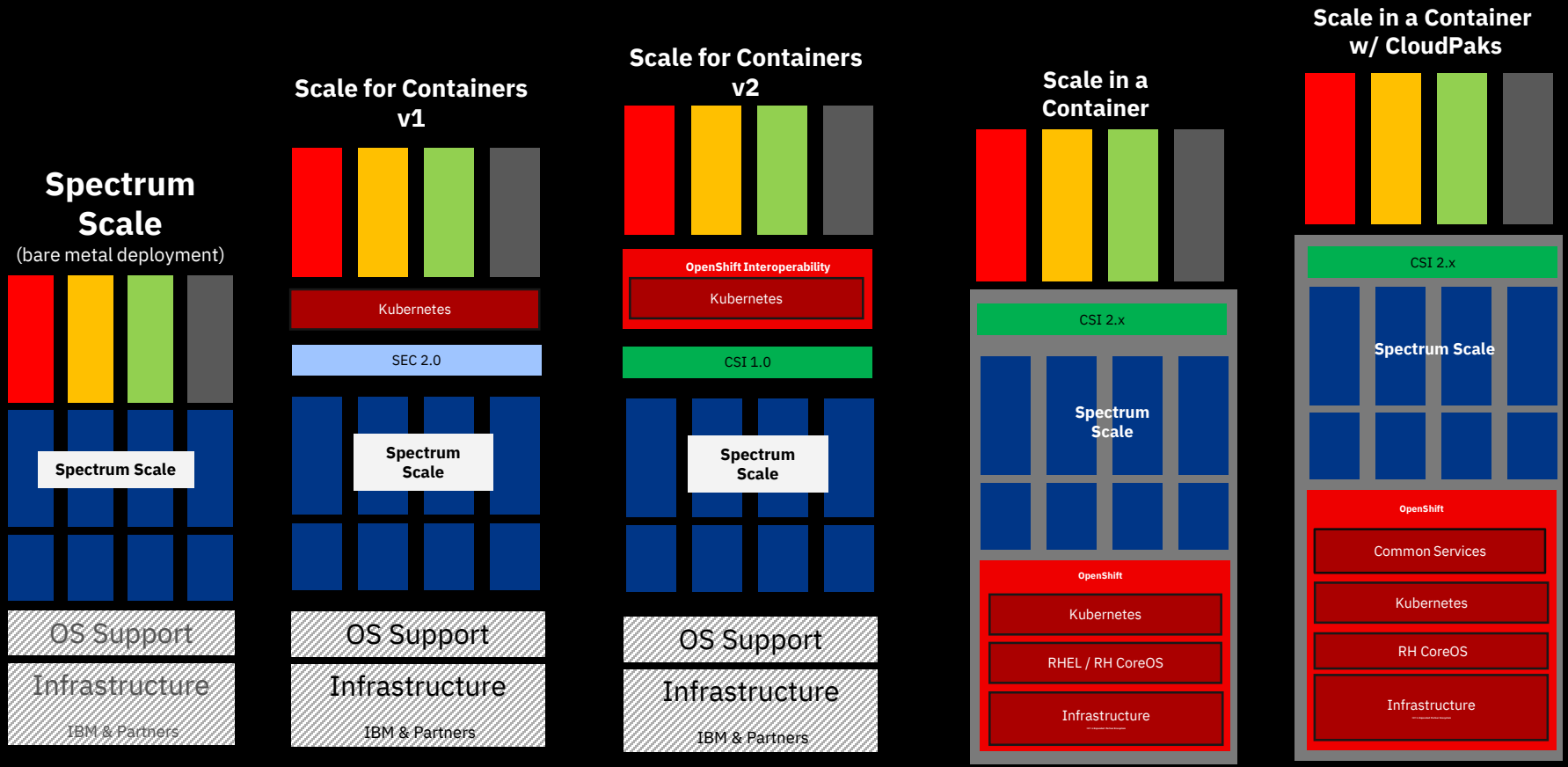
- Public, Private, Hybrid

Support Hybrid Use Cases

- Cloud Burst – Single Name Space
- Multi Cloud Data Sharing
- Archive
- Data Accelerator (High Performance Tiering)

Solution Integration (Partners)

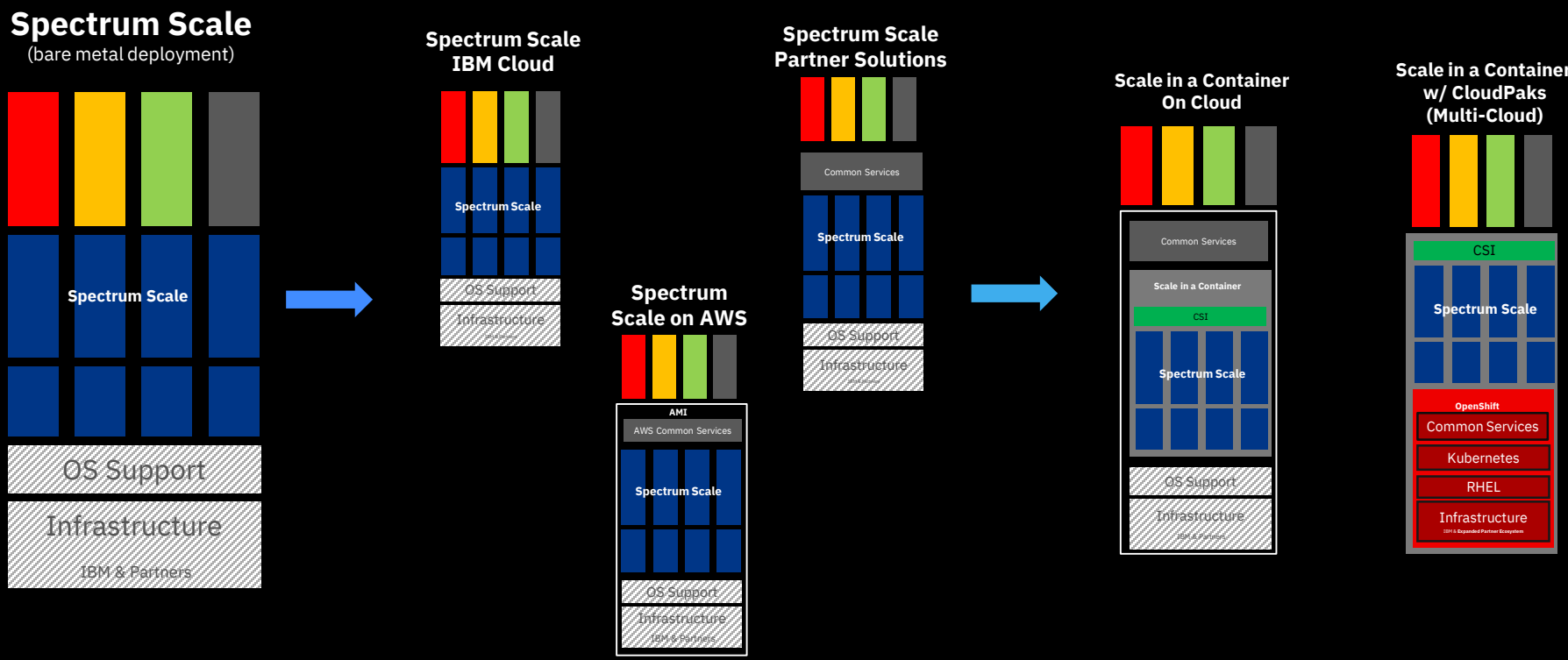
Evolution of IBM Spectrum Scale Containers



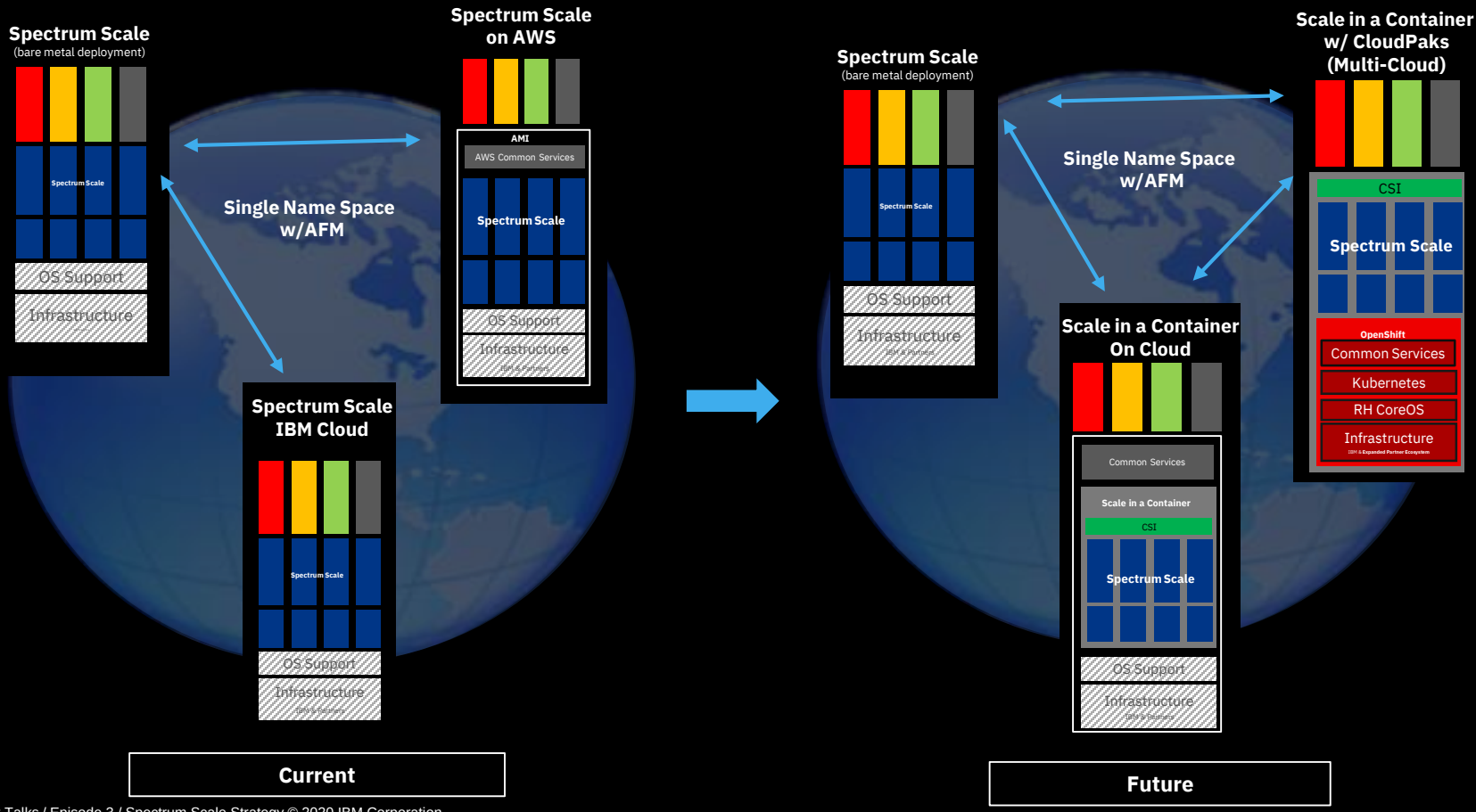
Evolution of IBM Spectrum Scale on Cloud

Current
Partner and Scale Offerings

Future
Partner and Scale Offerings



Evolution of Hybrid Cloud with IBM Spectrum Scale



Why DevOps?

Flexible Provisioning and Deployment

Consistency across On-Prem, Multi Cloud , Hardware Solutions

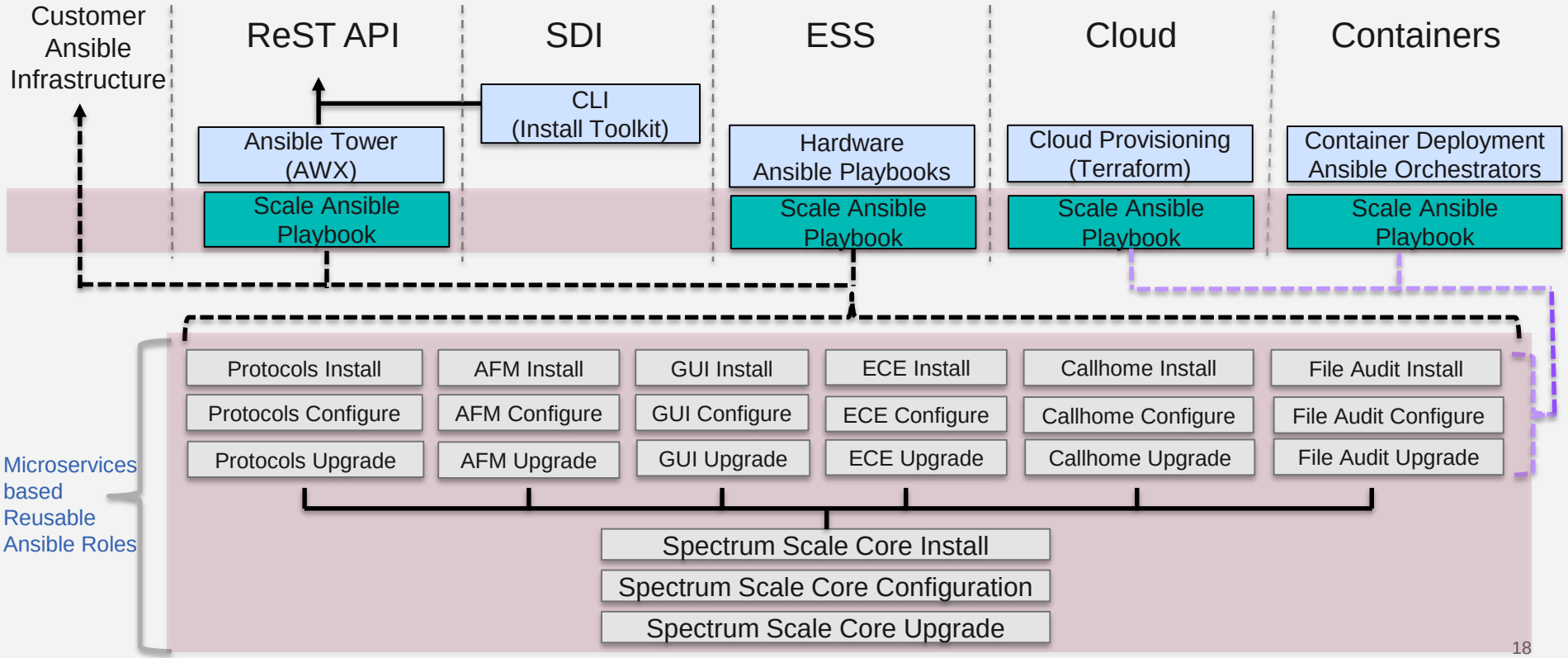
Needs to be Highly Customizable

- Microservices
- Integrated with Workload
- Open Source

Spectrum Scale Deployment: Strategy

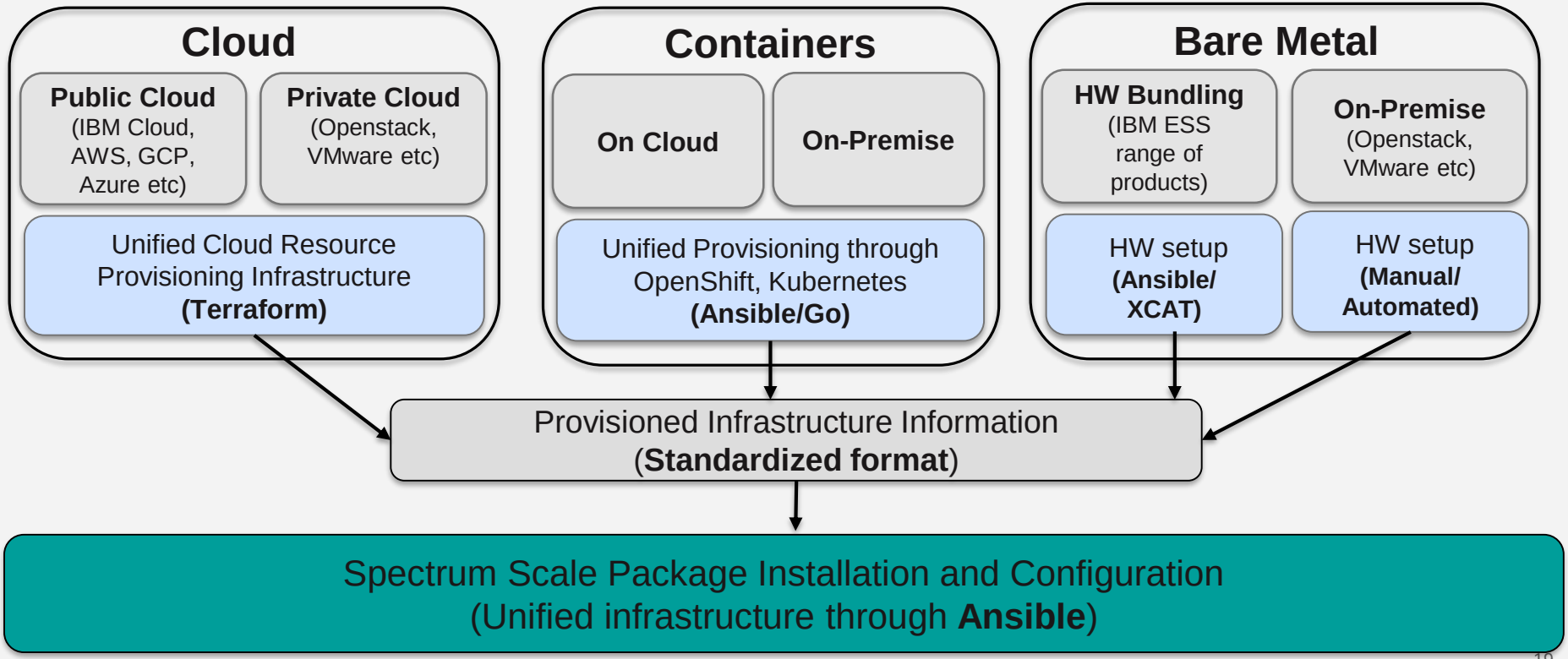
Microservices based reusable Ansible infrastructure

(Provides installation, configuration and upgrade capabilities for all Spectrum Scale form factors)



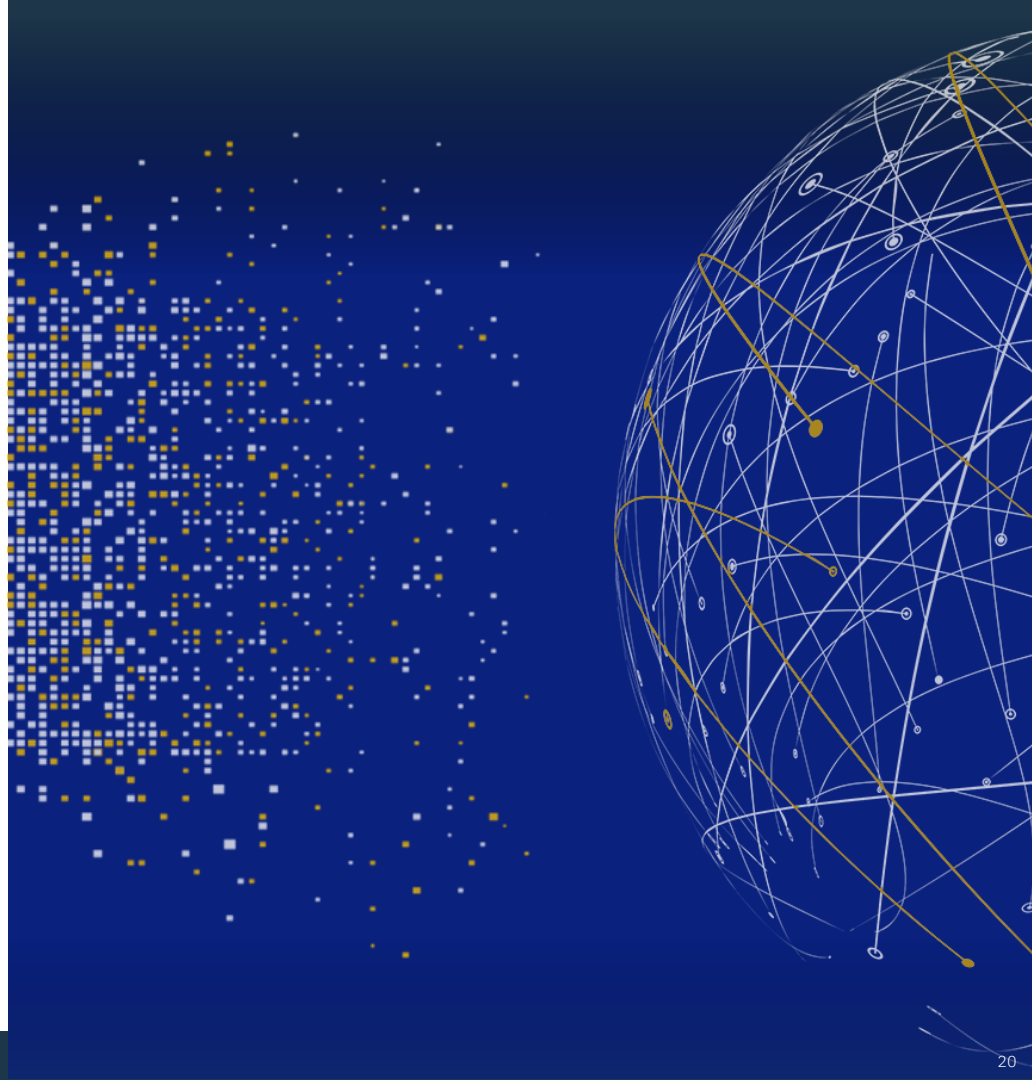
Spectrum Scale Deployment: Strategy

Infrastructure specific resource provisioning
Unified Installation and Configuration through reusable Ansible playbooks

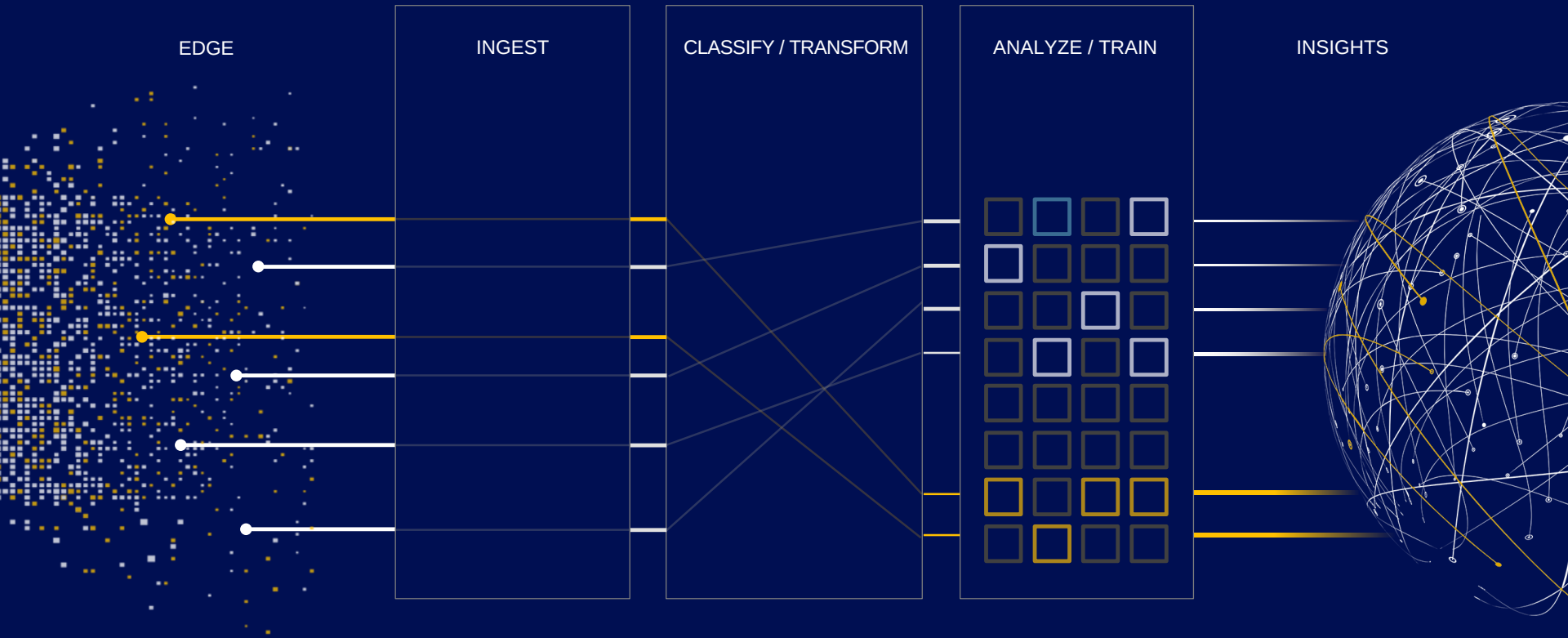


Data Management Challenges in AI and Analytics

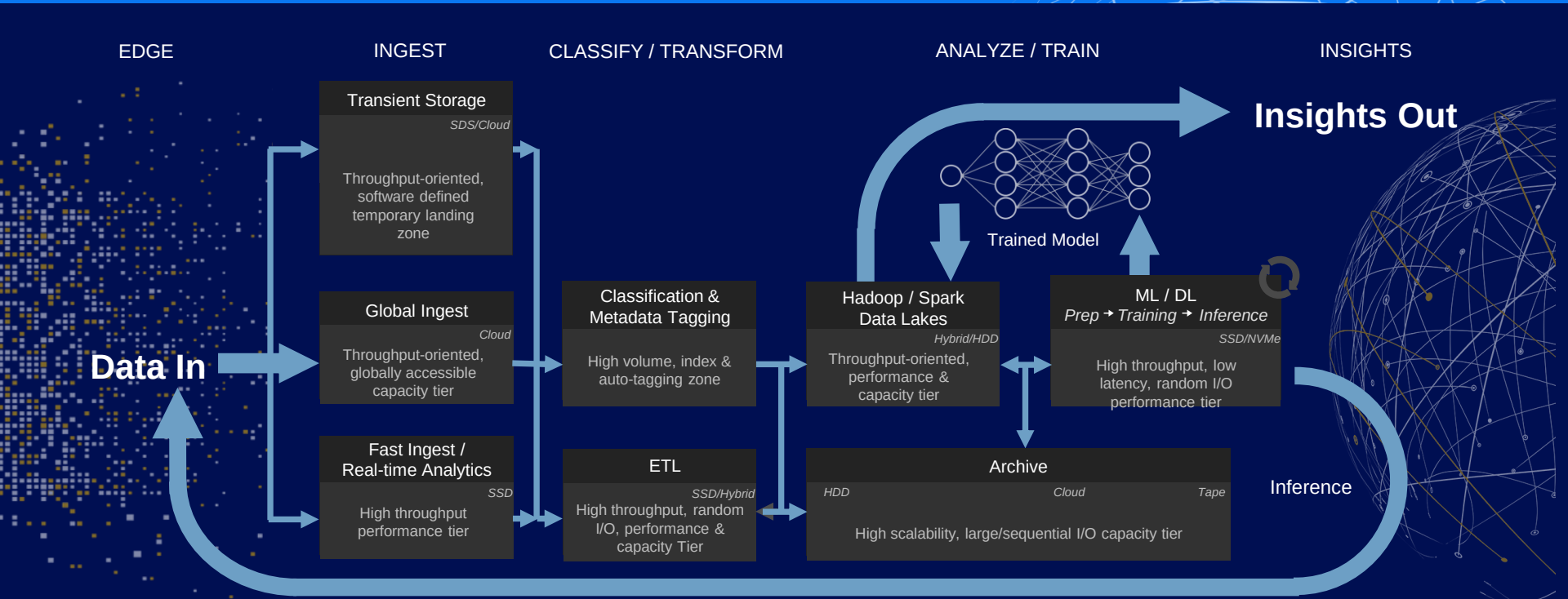
- Data ingest and preparation cycle are too time consuming
- Multi-source data aggregation
- Silos of infrastructure for various analytics use cases
- Multiple copies of same data without a single source of truth
- Analytics on stale data
- Need to securely manage and protect data provenance for repeatability
- Need for global accessibility and collaboration



The Goal: *Move Data from Ingest to Insights*



AI Data Pipeline



1. Single name space across storage platforms
2. Global collaboration / Hybrid Multi-Cloud
3. Indexing, Auto tagging / metadata management
4. Integrated analytics platform

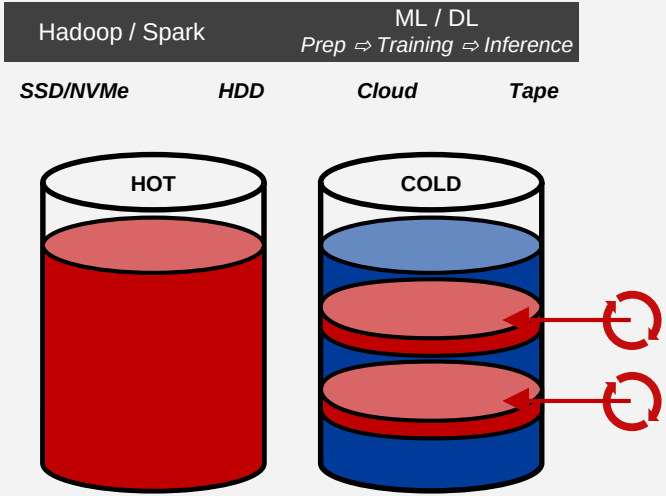
IBM Spectrum Scale
IBM Cloud Object Storage
IBM Spectrum Discover
IBM Cloud Paks

Spectrum Scale & ESS
Cloud Object Storage
Spectrum Discover
IBM Cloud Paks

The Problem

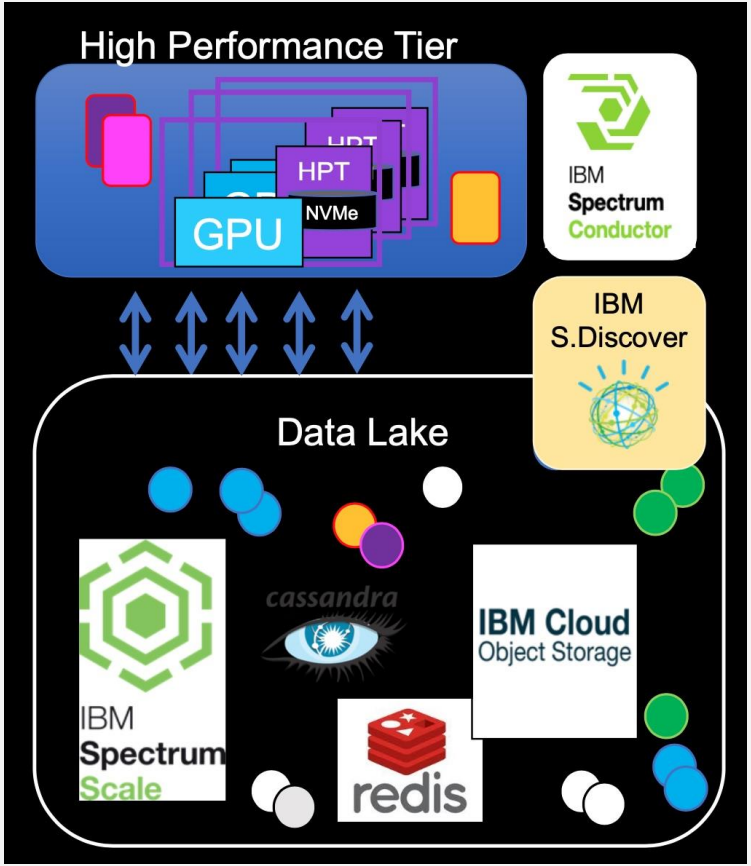
We see:

- Customers across all verticals are creating large PB to EB data stores.
- Vast majority of data is relatively cold, but still required for periodic trend analysis.
- But AI / Analytics require high performance, low latency storage to keep expensive CPU / GPU / TPU / FPGA busy.



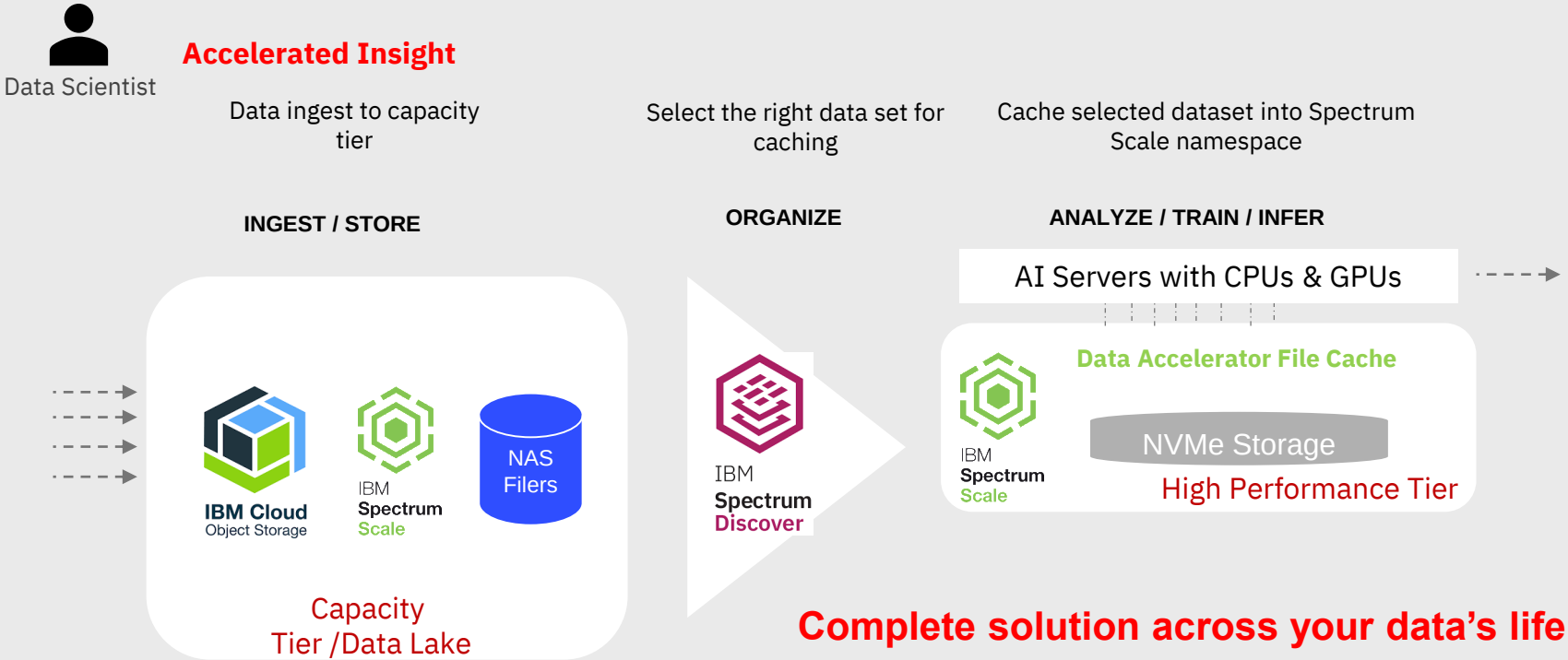
Moving Data as Close as Possible to Compute

- 2 storage tiers with different storage characteristics
  - Data Lake – minimize storage cost
  - High Performance Tier (HPT) – maximize storage performance
- HPT – high-performance data analytics on shared data, scale-out cluster FS, common namespace, no data transformations, data on-demand or prefetch, periodically revalidates cache
- S.Discover – curates data lake, metadata search engine, loads HTP, starts analytics, overall governance
- S.Conductor – Intelligent workload manager
- Data Lake – COS, Cloud or any high capacity data store
- Performance – single node & scale out
- End to end security and monitoring
- Can be deployed on-prem to on-prem, one-prem to cloud or cloud to cloud.

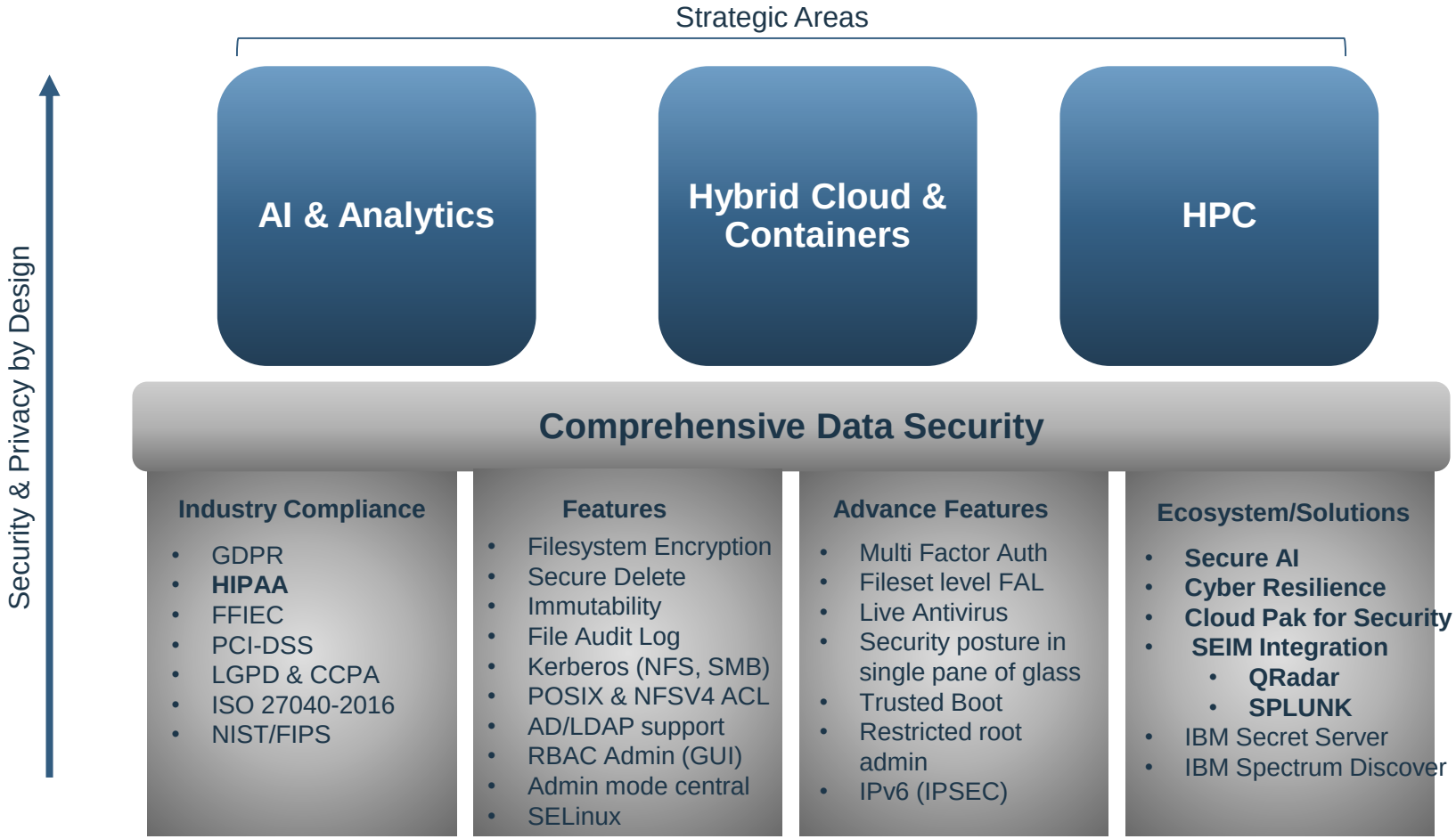


Spectrum Scale as Data Acceleration for AI and Analytics (DAAA)

- Accelerate model training output by prefetching selected dataset real-time in your ML/DL environment from the Hadoop/Spark data lake.
- Accelerate real time analytics / inference output by prefetching selected dataset real-time in a near-edge environment from the remote centralized data lake.



Spectrum Scale Strategic Areas: Security Feature Outlook



IBM Spectrum Scale and IBM QRadar: Threat Detection and Data Protection

Motivation

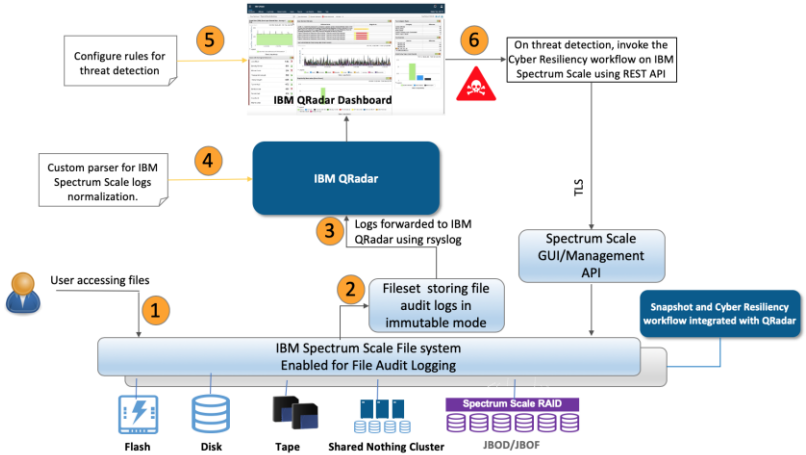
- Attacks against businesses have almost doubled in five years, and incidents that would once have been considered extraordinary are becoming more and more commonplace.
- If Data is the ‘Crown Jewel’ then Storage (Spectrum Scale) is the ‘Jewel Safe’ – lets make it more safe.
- IBM QRadar is a leading SIEM+ which analyzes event data in real time for early detection of targeted attacks and data breaches.

Benefits to Customers

Integrating IBM Spectrum Scale with IBM QRadar allows:

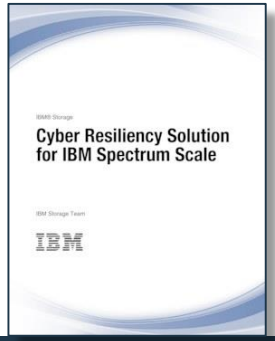
- Customers to proactively safeguard their data residing on Spectrum Scale or be alerted on potential threats (internal / external) in real time.
- Auto trigger data protection and backup on threat detection integrating with Cyber Resiliency solution.
  Solution Brief Released (Q1 2020)

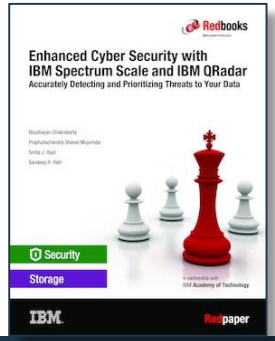
Solution Architecture

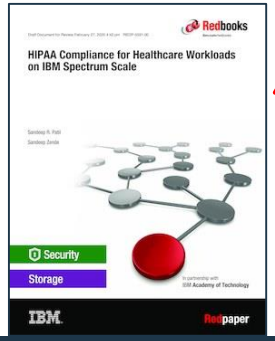


Blueprint & Redpapers:

http://www.redbooks.ibm.com/redpieces/abstracts/redp5591.html







New!

Next Generation Performance

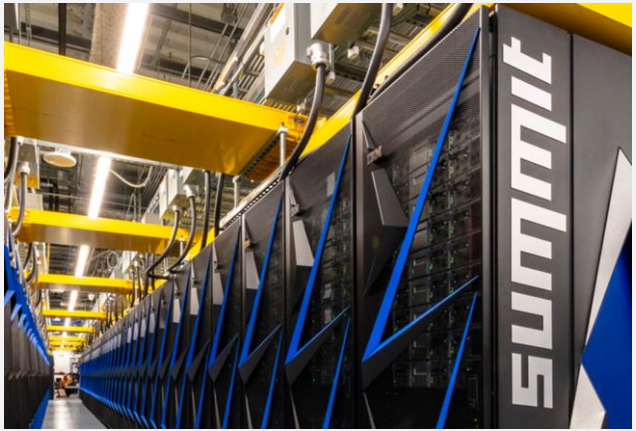
Scale's **Erasure Code Edition (ECE)** was announced in May 2019. Deploy the same ESS erasure encoding across storage rich servers for low cost reliable storage from commodity hardware.

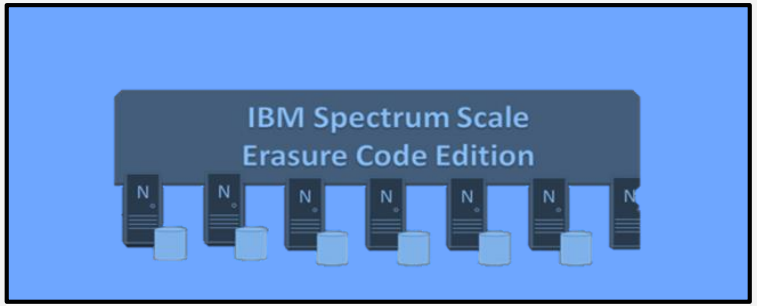
Leverage ECE with **Persistent Memory** to create Extreme Performance to local store.

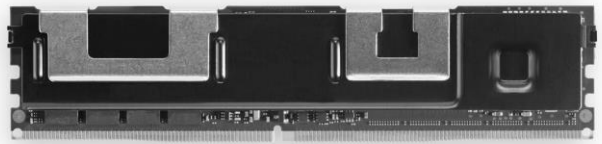
Research project to exploit persistent memory:

- ✓ Cooperative consistent client cache
- ✓ Topology Aware Replication with data affinity
- ✓ Read Ahead / Write Behind job scheduling
- ✓ Eventual Durability

Invest to accelerate time to value







Intel Optane DC Persistent Memory

Next Generation Performance (part 2)

Besides Persistent Memory, Spectrum Scale is continuing to invest in high throughput, low latency storage for AI and Analytics, HPC, Cognitive and Mission Critical workloads.

- PCIe Gen3 (1x) -> PCIe Gen4 (2x)-> PCIe Gen5 (4x)
- Network IB/RoCE/TCP: 100 gb -> 200 gb -> 400 gb
- NVMeoF creates “Composable Storage Infrastructure”
- Smart NICs: TCP offload, Encryption, Compression, Erasure Encoding, QoS, vLan, dynamic flow control, etc

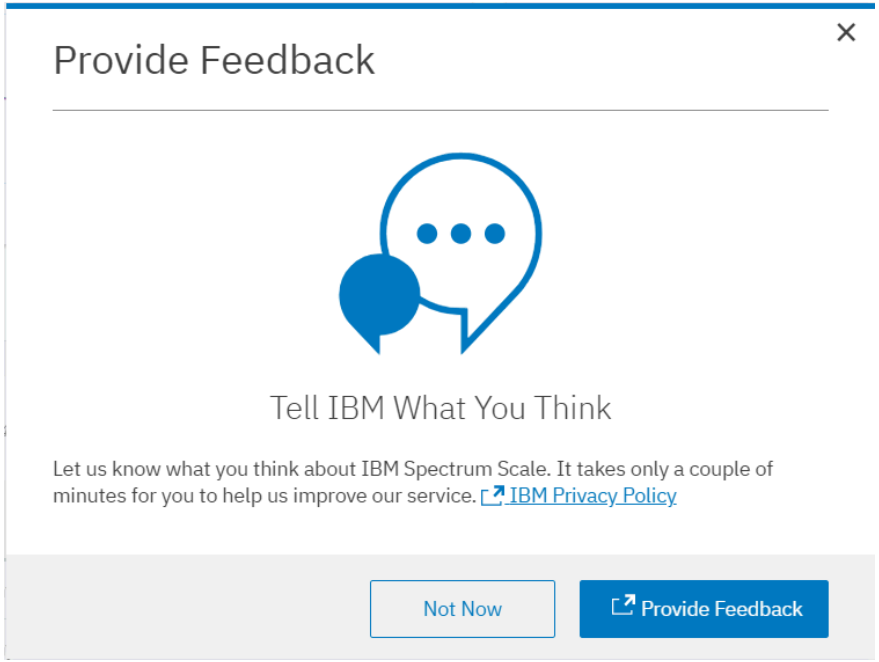
Hardware performance will increase by a factor of 10 in next few years. Spectrum Scale and ESS are making the investment required to continue its performance leadership.

Thank you!

IBM Spectrum Scale

Please help us to improve Spectrum Scale with your feedback

- If you get a survey in email or a popup from the GUI, please respond
- We read every single reply



Spectrum Scale User Group

The Spectrum Scale (GPFS) User Group is free to join and open to all using, interested in using or integrating IBM Spectrum Scale.

The format of the group is as a web community with events held during the year, hosted by our members or by IBM.

See our web page for upcoming events and presentations of past events. Join our conversation via mail and Slack.

www.spectrumscaleug.org