

# IBM Fusion HCI – on prem cloud for OpenShift + watsonx for enterprise GenAI

**Ivo Gomilsek**

ivog@at.ibm.com

Senior Infrastructure Architect, IBM NCEE

**Novica Nivic**

novica.ninic@rs.ibm.com

Silver sponsor



# Agenda



IBM Fusion  
by IBM Storage

- ✓ IBM Fusion HCI - on prem cloud for OpenShift
- ✓ watsonx for enterprise GenAI



IBM Fusion  
by IBM Storage

# IBM Fusion HCI - on prem cloud for OpenShift

# Containers are the engine of digital transformation

Why are so many projects stuck in pilot?

## What's missing?

**Simple** and **consistent**  
ways to support the **data**  
needs of **mission-critical**  
stateful applications

## Organizations need ways to:

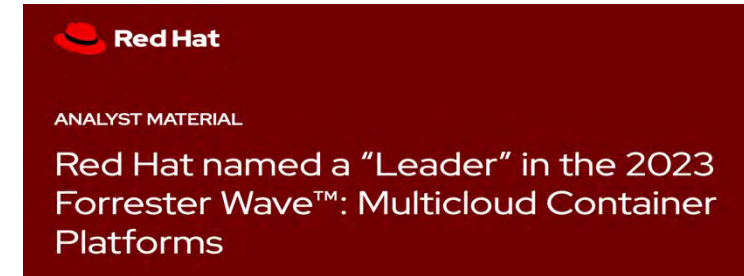
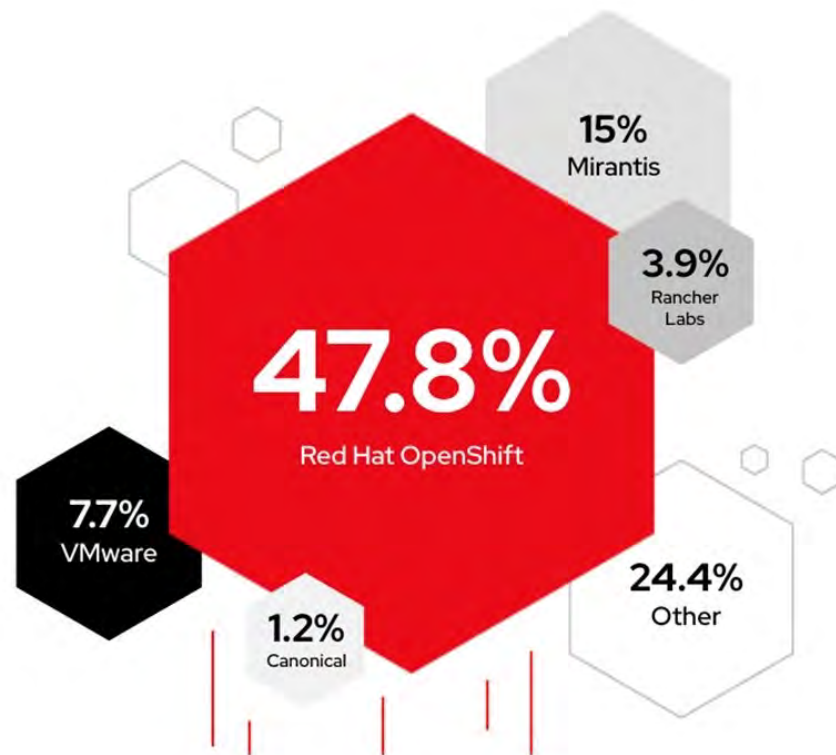
- **Protect application data**  
against disasters, theft, and cyber attacks
- **Ensure application availability**  
to maintain business operations continuity
- **Securely access data, anywhere**  
to support hybrid and multi-cloud  
use cases efficiently and securely
- **Achieve business objectives**  
for performance, scaling, and cost

# RHOS goes from pilot to production

Red Hat OpenShift is the market leader for hybrid cloud and multicloud Kubernetes deployments



FORTUNE  
500



**Challenges** remain getting applications out of pilot and into production

- Ensuring application availability (HA)
- Protecting and securing data (DR/backup)
- Integrating with legacy storage
- Achieving performance and scalability objectives
- Operationalizing container-native applications (the skills gap)



**Ad-hoc** data management can and do produce adverse business outcomes

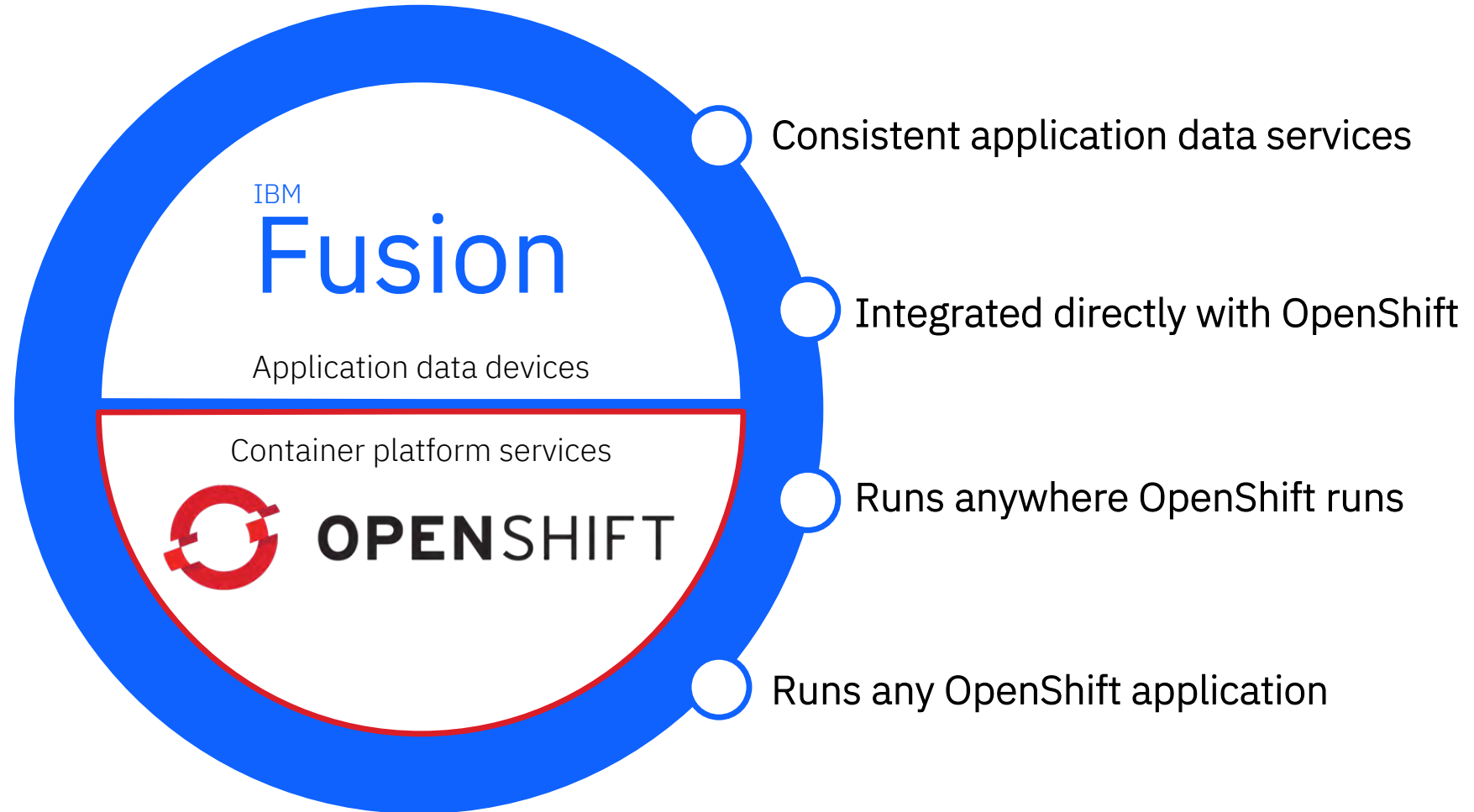
- Project delays and cost overruns
- Application outages
- Poor data governance and oversight sensitive data exposed to theft, loss, and corruption
- Reduced productivity
- Inefficient use of infrastructure

There is often a **disconnect** between **app owners** and **IT operations** who support the infrastructure



# IBM Fusion

Modern cloud-native application data services platform





# IBM Fusion

Container data services that are **simple to use, consistent everywhere,** and **strategic**

## Protect application data

Configure application backup policies;  
recover applications to any point in time

## Ensure application availability

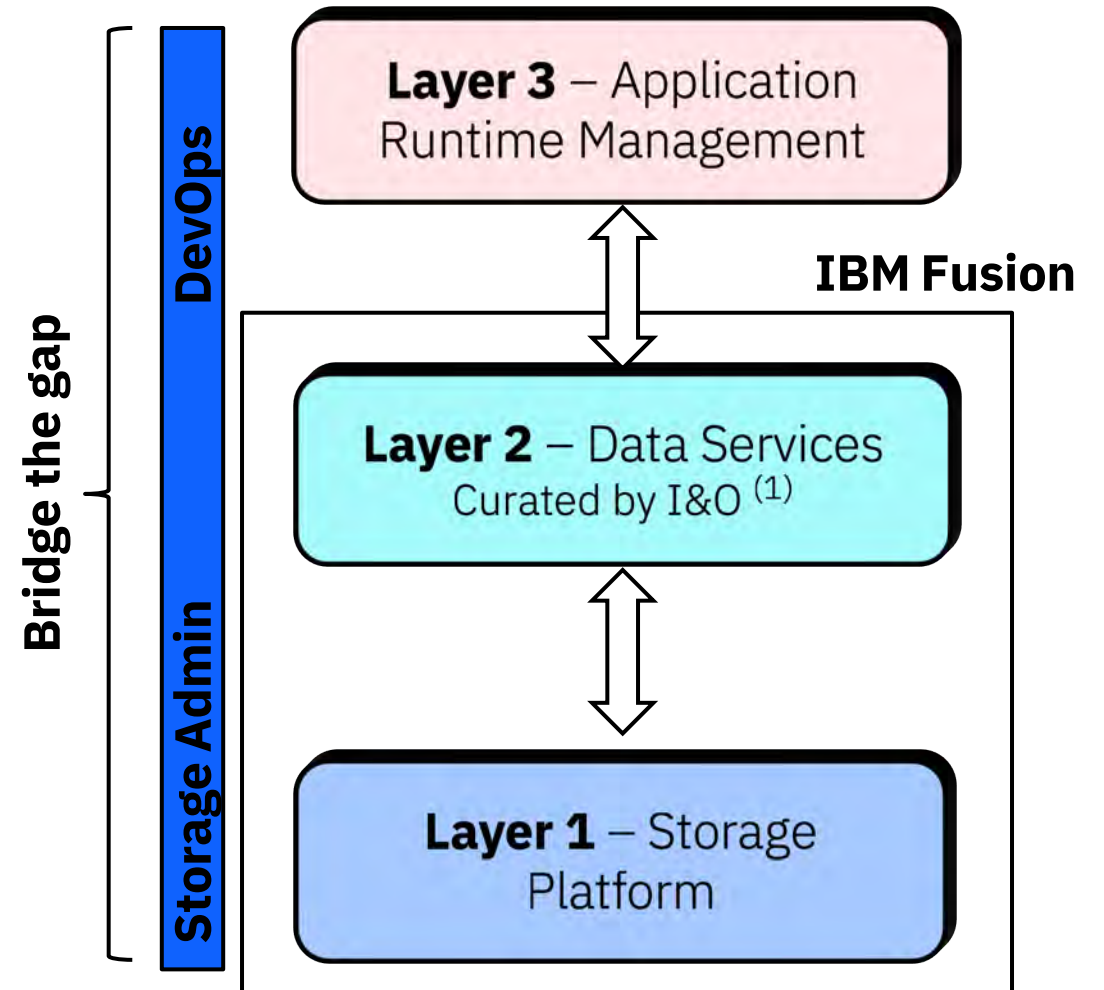
Configure cross-zone data  
replication; manage availability with  
policies to RPO / RTO objectives

## Access any data, anywhere

Manage access to data with  
policies; connect applications  
to any data source, anywhere

## Data cataloging built-in service

Orchestrate your data by integrated  
governance, inspection and data classification engine



1. Infrastructure and Operations

# IBM Fusion with two deployment options

OpenShift appliance or standalone software

Customer Apps



IBM Cloud Paks

- Data
- Integration
- Security
- Cloud Satellite
- Business automation
- Watson AIOps
- Network automation



Databases

- Cassandra
- MongoDB
- RabbitMQ
- Elastic search
- PostgreSQL
- Spark



Off the shelf

- Pega
- Mulesoft
- TIBCO



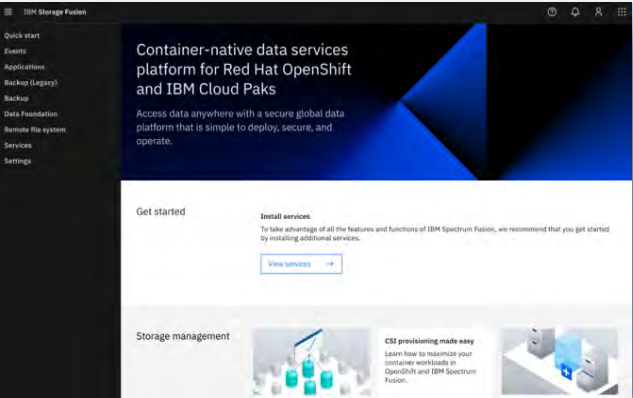
Custom apps

Home grown

Offering

## Fusion software

Data services for stateful OpenShift applications

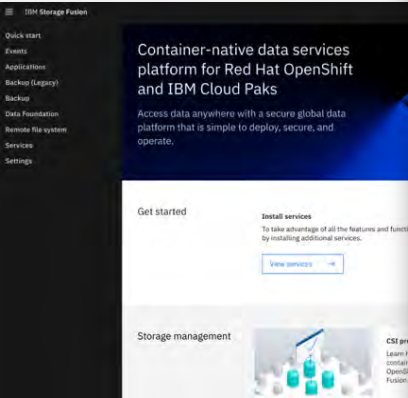


### Consistent experiences

- APIs
- Fusion console
- Data protection
- Disaster Recovery
- Fusion Data Foundation (FDF)

## Fusion HCI

Integrated Application Platform for OpenShift



Built on

Deployments



AWS EBS



Azure block



Persistent disk



IBM block



SAN, vSAN



DS8K



FlashSystems



Bare metal

Server-attached drives

Hyper-Converged Infrastructure for OpenShift

Switches, servers, storage

# IBM Fusion HCI

## Bare Metal is better



### Higher Performance

- Eliminate hypervisor to reduce overhead
- More resources available for workloads

### Lower Cost, Simplify Operations

- Reduce OpenShift license cost
- Eliminate VM operational complexity

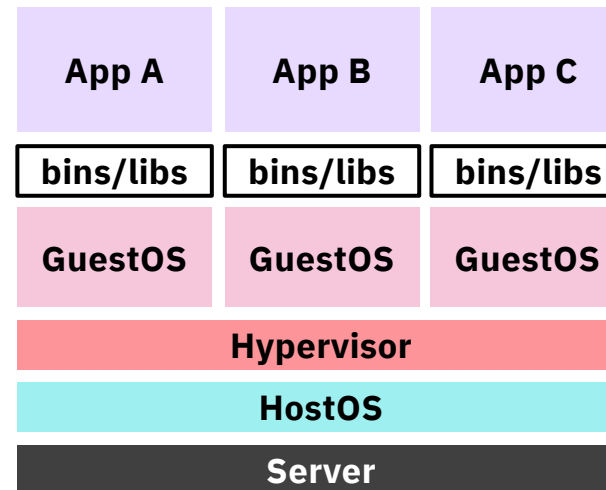
### Improve Security

- Immutable CoreOS reduces attack surface
- Sandboxed containers

### Support Windows and Linux VMs

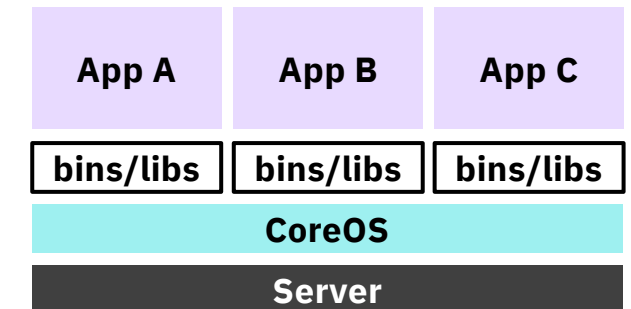
- Manage with OpenShift Virtualization

#### Good



Virtualized Infrastructure  
with many layers

#### Better



Bare metal on  
IBM Fusion HCI

# IBM Fusion HCI

A better way to run mission-critical applications on bare-metal OpenShift

## IBM Cloud Paks

- Data, Security, Integration
- Network Automation,
- Business Automation,
- Watson AI Ops

## Client/Partner

- Any OpenShift application
- NVIDIA CUDA
- Edge applications
- ...

## Open Source

- AI/ML: TensorFlow, PyTorch, NumPy, run.ai, etc.
- PostgreSQL, MongoDB
- ...

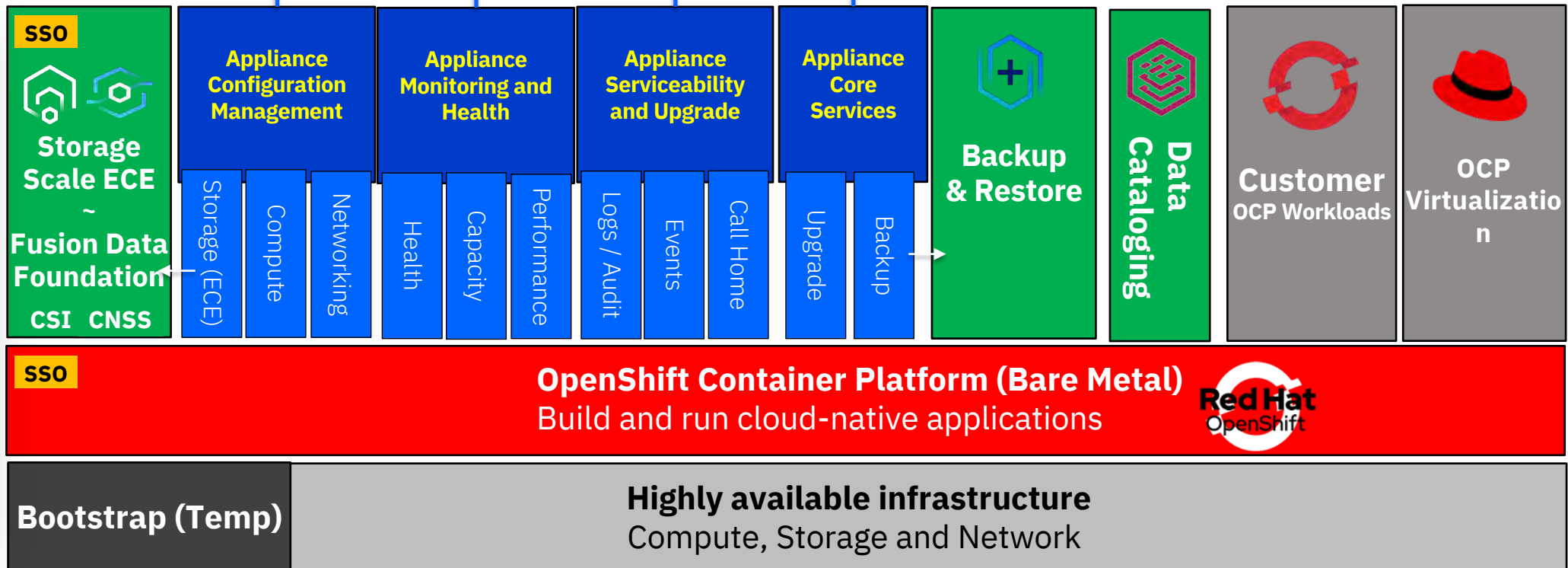


Red Hat  
Advanced Cluster  
Management

SSO  
Appliance UI



Satellite

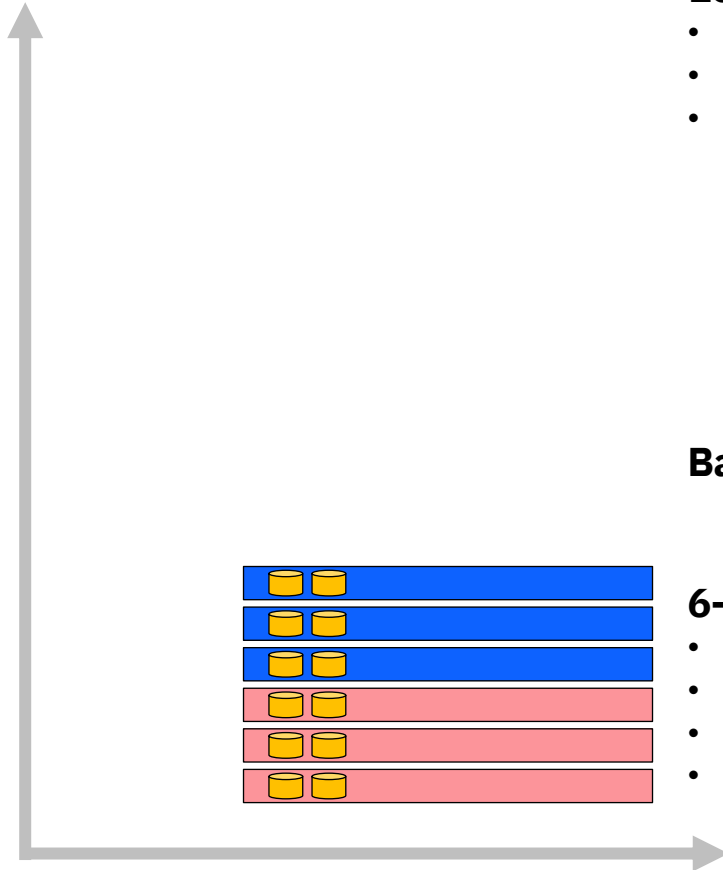


# IBM Fusion HCI

## Scale-up & Scale-out (single rack version)

### Compute & Storage scalability

6 nodes to 16 nodes



#### 16-node Fusion HCI rack (max size)

- Up to 1,117 TiB raw, 772 TiB usable storage
- Up to 823 usable cores (1646 vCPU)
- Up to 31.4 TB usable RAM

#### Base configuration

#### 6-node Fusion HCI rack (min size)

- v. 3.84TB: 41.9 TiB raw, 26.7 TiB usable storage
- v. 7.68TB: 83 TiB raw, 53.5 TiB usable storage
- 60 cores (120 vCPU) usable
- Starting with 528GB RAM

#### Storage scalability

2-10x 3.84TB or 7.68TB NVMe flash drives (in pairs) per node

### Configuration Options

X86 Compute Only

X86 Compute Only

Compute only nodes (nodes 7 – 16)

NVIDIA 8x L40s GPU

NVIDIA 8x L40s GPU

4x 3U GPU enhanced nodes with  
32x NVIDIA L40s GPUs per Rack

X86 communication nodes

X86 communication nodes

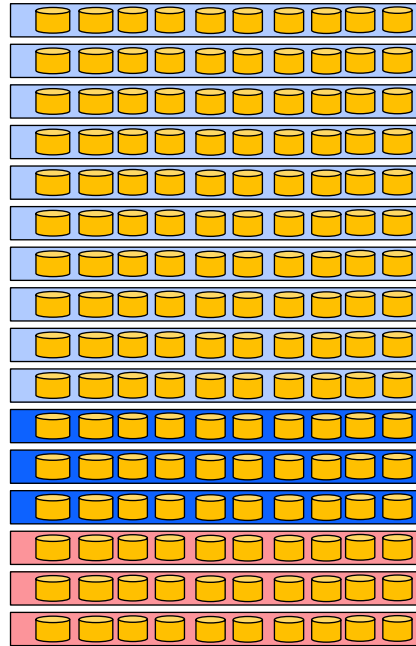
One pair 1U Global data access nodes

# IBM Fusion HCI

## Scale-up & Scale-out (single rack version)

### Compute & Storage scalability

6 nodes to 16 nodes



#### 16-node Fusion HCI rack (max size)

- Up to 1,117 TiB raw, 772 TiB usable storage
- Up to 823 usable cores (1646 vCPU)
- Up to 31.4 TB usable RAM

#### Maximum Compute/Storage Nodes Maximum Disk amount

#### 6-node Fusion HCI rack (min size)

- v. 3.84TB: 41.9 TiB raw, 26.7 TiB usable storage
- v. 7.68TB: 83 TiB raw, 53.5 TiB usable storage
- 60 cores (120 vCPU) usable
- Starting with 528GB RAM

#### Storage scalability

2-10x 3.84TB or 7.68TB NVMe flash drives (in pairs) per node

### Configuration Options

X86 Compute Only

X86 Compute Only

Compute only nodes (nodes 7 – 16)

NVIDIA 8x L40s GPU

NVIDIA 8x L40s GPU

4x 3U GPU enhanced nodes with  
32x NVIDIA L40s GPUs per Rack

X86 communication nodes

X86 communication nodes

One pair 1U Global data access nodes

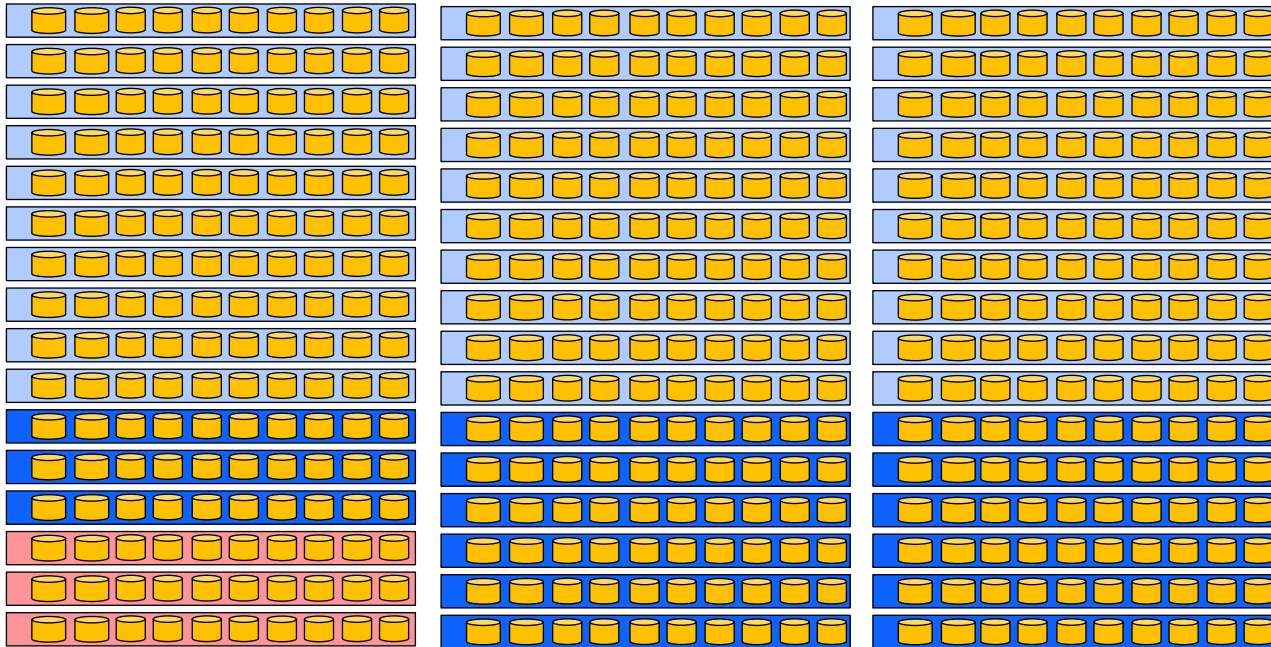


# IBM Fusion HCI

## Scale-up & Scale-out (3 rack cluster – rack expansion)

### Cluster scalability

3 x 16 node Racks



### Cluster 3 x 16 node racks

- 2.1 PiB usable capacity NVMe
- Up to 2487 usable cores (4974 vCPU)
- Up to 95.8 TB RAM

### Configuration Options

X86 Compute Only

Compute only nodes (nodes 7 – 16)

NVIDIA 8x L40s GPU

NVIDIA 8x L40s GPU

4x 3U GPU enhanced nodes with  
32x NVIDIA L40s GPUs per Rack

X86 communication nodes

X86 communication nodes

One pair 1U Global Data Access nodes –  
per rack

### Highlights

- 3 x Control Nodes in 1<sup>st</sup> rack
- Single Rack, 2 Rack or 3 Rack systems
- Up to 45 Worker Nodes expansion
- Base 6 Nodes configuration per Rack (new recovery group)
- Expansion by adding 1 Node

**Maximum Compute/Storage Nodes**  
**Maximum Disk amount**

### Storage scalability

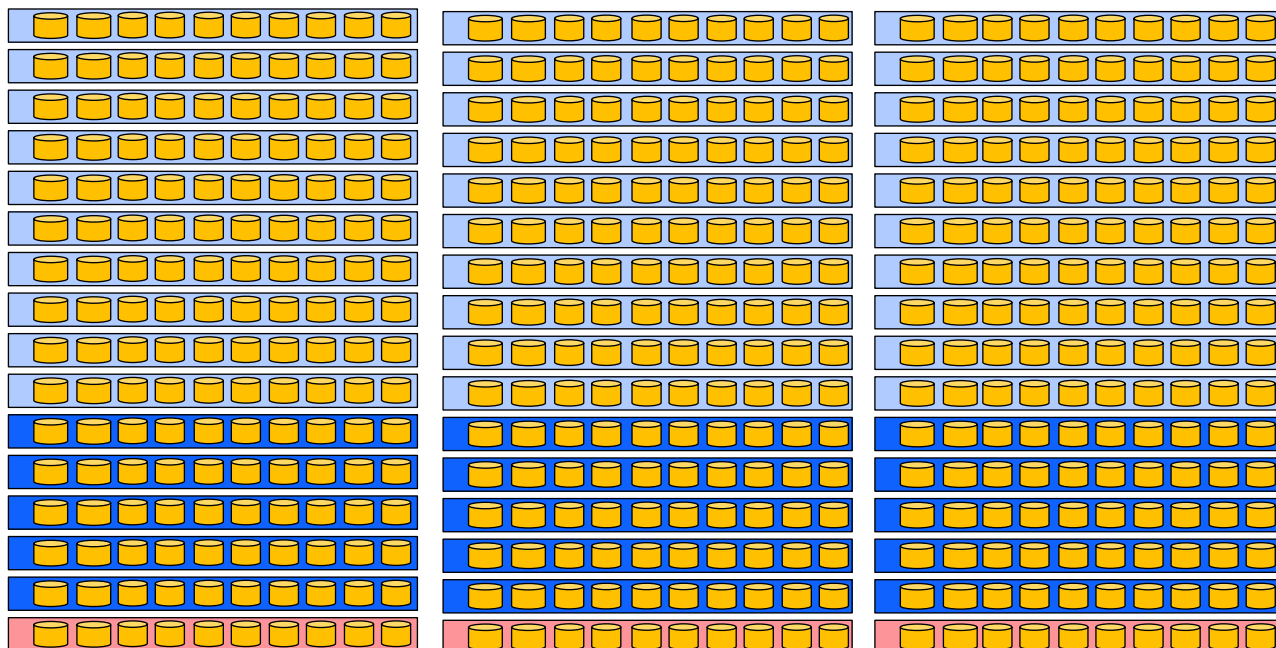
2-10x 3.84TB or 7.68TB NVMe flash drives (in pairs) per node

# IBM Fusion HCI

## Scale-up & Scale-out (3 rack cluster – rack-HA configuration)

### Cluster scalability

3 x 16 node Racks



### Cluster 3 x 16 node racks

- 2.1 PiB usable capacity NVMe
- Up to 2487 usable cores (4974 vCPU)
- Up to 94.4 TB RAM

### Configuration Options

X86 Compute Only

Compute only nodes (nodes 7 – 16)

NVIDIA 8x L40s GPU

NVIDIA 8x L40s GPU

4x 3U GPU enhanced nodes with  
32x NVIDIA L40s GPUs per Rack

X86 communication nodes

X86 communication nodes

One pair 1U Global Data Access nodes –  
per rack

### Highlights

- 3 x Control Nodes distributed in 3 racks
- Single Rack or 3 Rack systems
- Up to 45 Worker Nodes expansion
- Base 6 Nodes configuration per Rack (new recovery group)
- Expansion by adding 3 Nodes

**Maximum Compute/Storage Nodes**  
**Maximum Disk amount**

### Storage scalability

2-10x 3.84TB or 7.68TB NVMe flash drives (in pairs) per node



# IBM Fusion HCI

## Gen2 – Compute/GPU expansion options



<b>NODES:</b>	16 Compute	15 Compute 1 GPU Node	14 Compute 2 GPU Node	13 Compute 3 GPU Node	12 Compute 4 GPU Node
---------------	------------	--------------------------	--------------------------	--------------------------	--------------------------

<https://www.ibm.com/docs/en/sfhs/2.8.x?topic=prerequisites-hardware-overview-single-rack>

# IBM Fusion HCI

## Gen2 - Hardware layout (Intel)

### Base configuration includes (gray):

- 42U rack cabinet
- 2x Ethernet 100 GbE high-speed switches
- 2x Ethernet 1 GbE management switches
- 6x Storage/Compute servers with 2x NVMe drives/server \*

### Options available (blue):

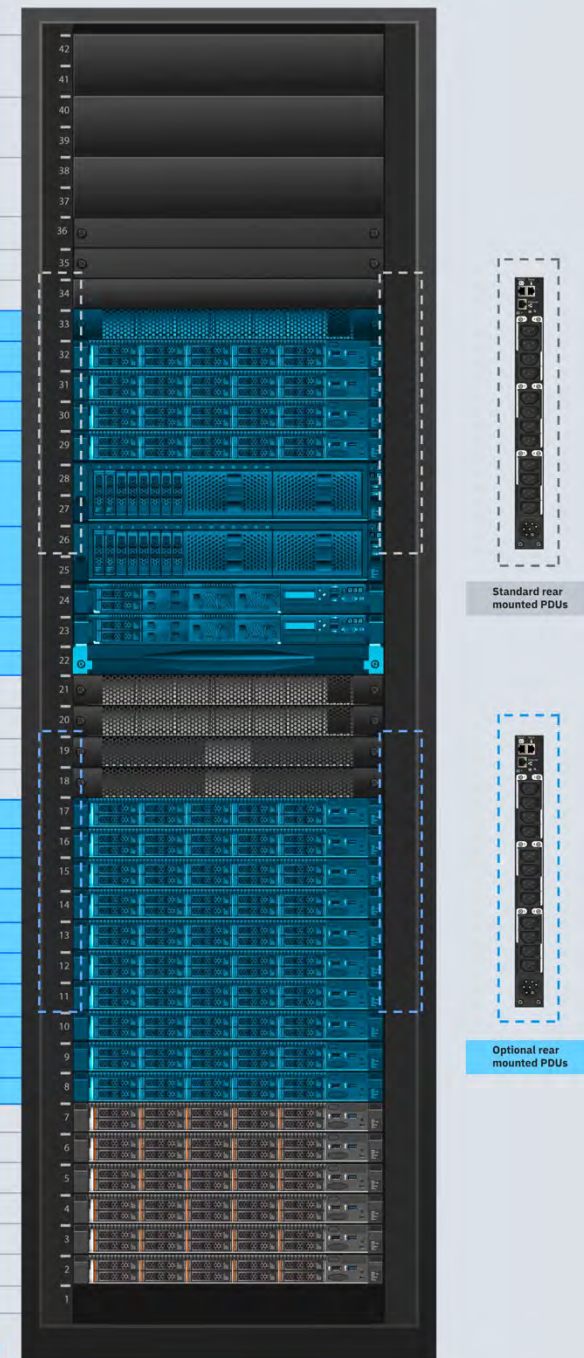
- 32c or 64c multithread nodes \*
- Storage rich and compute only nodes
- Memory options 8 GB to 32 GB RAM per physical core
- Up to 16 nodes per rack, in-field scalable  
(20 nodes possible, but in single rack configuration only)
- 3U GPU servers, each with 0 up to 8 NVIDIA L40s GPUs
- Increased storage by adding pairs of drives to storage/compute servers
  - 3.84TB or 7.68TB NVMe PCIe Gen5 drives/server to a max of 10 drives/server
- AFM (Active File Manager) NFS/S3 gateway nodes

#### \* CPU details:

2x Intel Gold 6426Y 16C 2.5 GHz 185 W CPU ("Sapphire Rapids")  
2x Intel Gold 6438N 32C 2.0 GHz 205 W CPU ("Sapphire Rapids")

42	
41	2U Filter
40	
39	2U Filter
38	
37	2U Filter
36	1U PDU (Horizontally mounted)
35	1U PDU (Horizontally mounted)
34	1U Filler
33	32-Port 100 GbE Ethernet spine switch
32	Storage / Compute server
31	Storage / Compute server
30	Storage / Compute server
29	Storage / Compute server
28	GPU Server with 3x GPU PCIe Gen 4 adapter cards
27	GPU Server with 3x GPU PCIe Gen 4 adapter cards
26	
25	
24	AFM Node
23	AFM Node
22	KVM
21	32-Port 100 GbE Ethernet spine switch
20	32-Port 100 GbE Ethernet spine switch
19	48-Port 1 GbE Management Ethernet Switch
18	48-Port 1 GbE Management Ethernet Switch
17	Storage / Compute server
16	Storage / Compute server
15	Storage / Compute server
14	Storage / Compute server
13	Storage / Compute server
12	Storage / Compute server
11	Storage / Compute server
10	Storage / Compute server
9	Storage / Compute server
8	Storage / Compute server
7	Storage / Compute server
6	Storage / Compute server
5	Storage / Compute server
4	Storage / Compute server
3	Storage / Compute server
2	Storage / Compute server
1	Reserve 1U Space at Bottom

Optional components



# IBM Fusion HCI

## version 2.8.x – available services

The screenshot displays the IBM Storage Fusion web interface. The top navigation bar includes a close button, the text 'IBM Storage Fusion', the instance name 'hciocp1', and icons for help, user, notifications, and a menu. The left sidebar contains a list of navigation items: Quickstart, Events, Applications, Backup & restore, Disaster Recovery, Cloud Satellite, Infrastructure, Storage, Services (highlighted), and Settings. The main content area is titled 'Services' and includes a subtitle 'Discover and manage available services. [Learn more](#)'. It is divided into two sections: 'Installed' and 'Available'. The 'Installed' section lists two services: 'Backup & Restore v2.8.1' and 'Global Data Platform v5.2.0.1', both with a 'Healthy' status indicator. The 'Available' section, with the subtitle 'View list of [supported services](#).', shows two service cards. The first card is for 'Data Cataloging' (IBM • Metadata), which 'Provides rapid automated data discovery and robust metadata capture, curation and enrichment.' The second card is for 'Data Foundation MCG Only' (IBM • Storage), which 'Provides object storage service based on existing storage.'

IBM Storage Fusion hciocp1

Quickstart  
Events  
Applications  
Backup & restore  
Disaster Recovery  
Cloud Satellite  
Infrastructure  
Storage  
Services  
Settings

### Services

Discover and manage available services. [Learn more](#)

#### Installed

Service	Version	Status
Backup & Restore	v2.8.1	Healthy
Global Data Platform	v5.2.0.1	Healthy

#### Available

View list of [supported services](#).

#### Data Cataloging

IBM • Metadata

Provides rapid automated data discovery and robust metadata capture, curation and enrichment.

#### Data Foundation MCG Only

IBM • Storage

Provides object storage service based on existing storage.

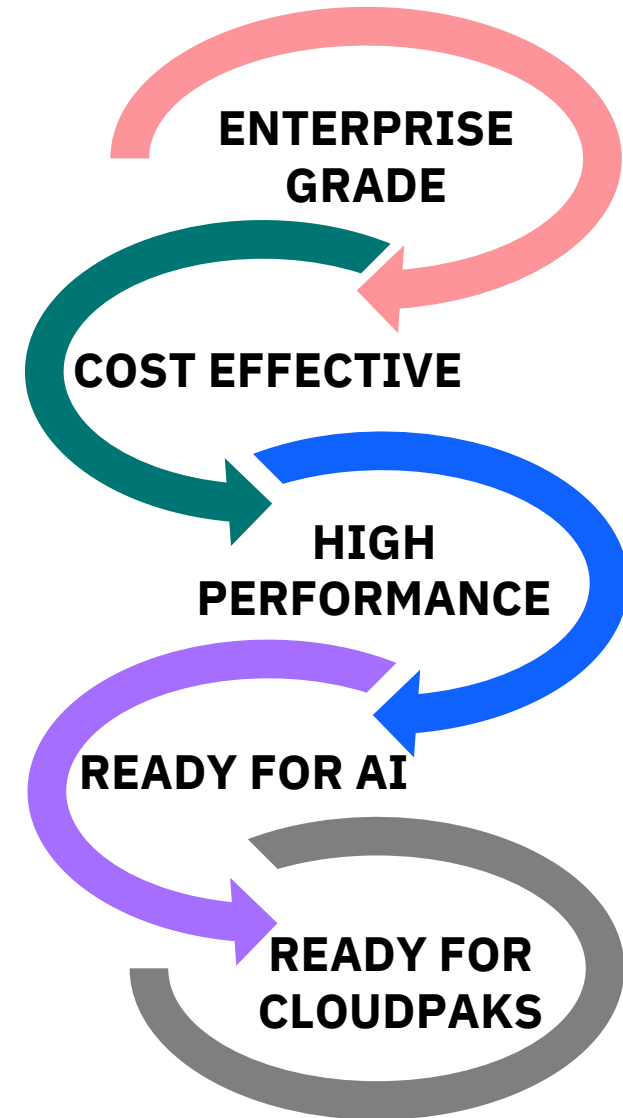


# IBM Fusion HCI



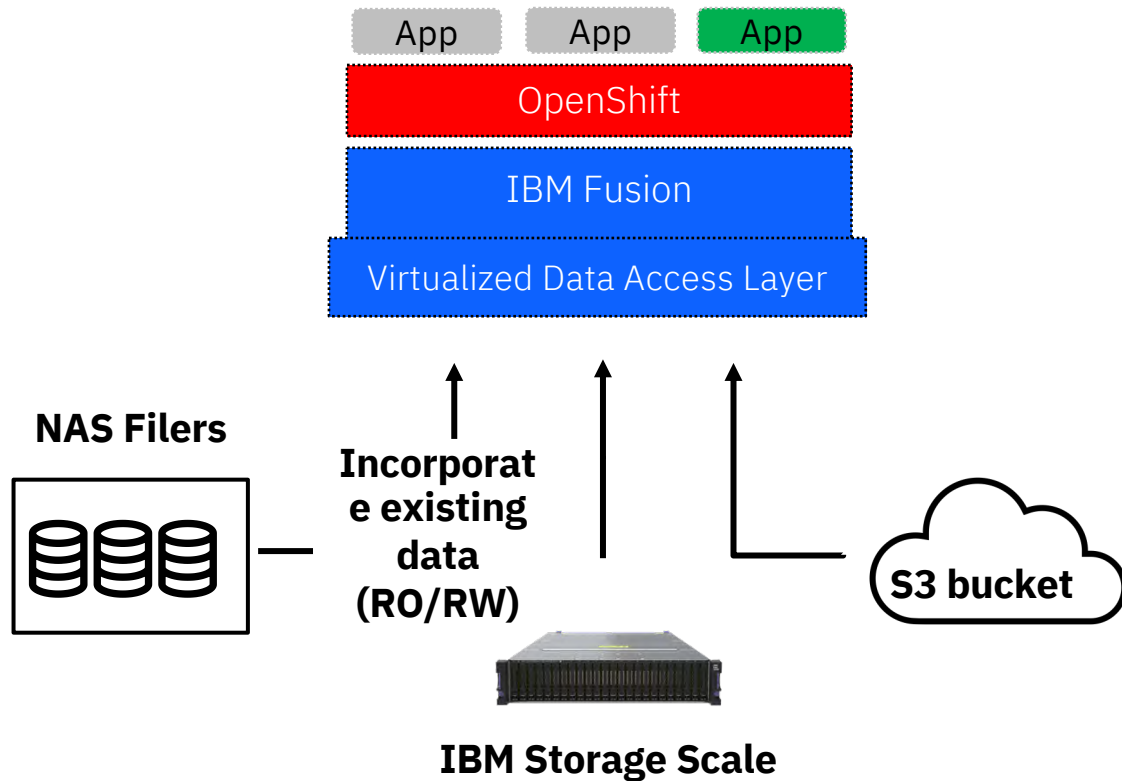
**Factory integrated x86  
bare-metal OpenShift  
Cloud-in-a-Box/Cloud on  
prem**

Designed for AI, for  
CloudPaks  
and other container-native  
workloads



# IBM Fusion

Access to any data, anywhere from edge to core to cloud



## Access data anywhere, local and remote

Connect to NAS filers or S3 object stores

Connect to Storage Scale

Protect investments – Add data capacity from existing resources

Choose the performance required with flash or capacity drives

## High performance access to remote data

Cache remote data locally in a performance tier

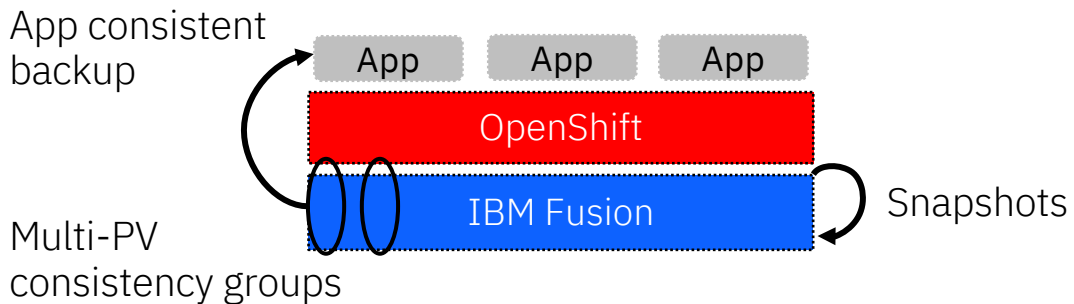
Maintain two-way data consistency between the cache tier and the remote data

Allow one-way read only access

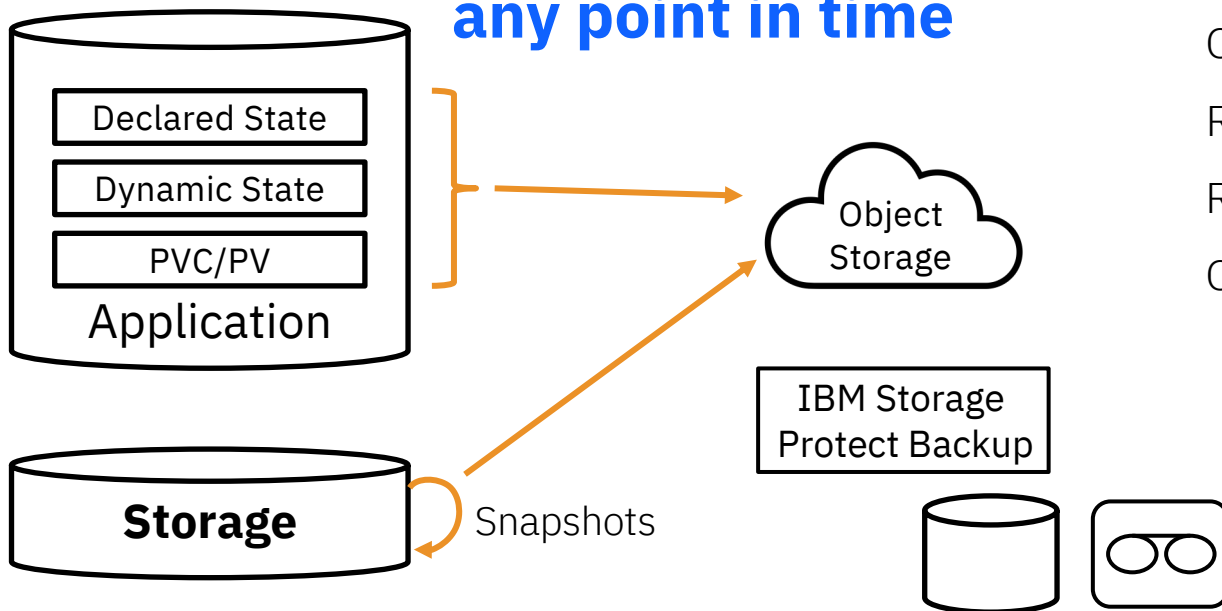
Control caching with policies (pre-fetch, cache size, access mode, etc.)

# IBM Fusion

## Protect application data & migration



**Recover to  
any point in time**



## Enterprise data protection to backup and restore container applications

Policy driven orchestration to meet range of SLAs

Application consistency groups for **crash-consistent backups**

Application hooks for **application-consistent backups**

Offload of backups to **Object Storage - anywhere**

Restore to a previous **point-in-time copy**

Restore across OpenShift clusters

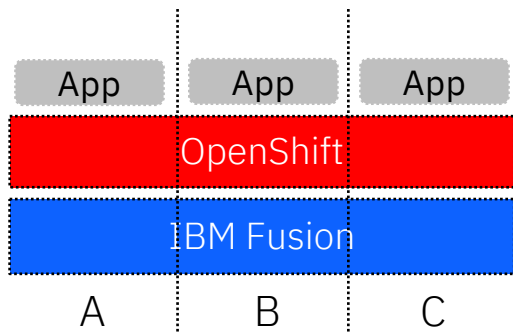
Object backup to Tape with IBM Storage Protect

**Application backup is more than  
just backing up storage**

# IBM Fusion

Ensure application availability – High Availability/Disaster Recovery

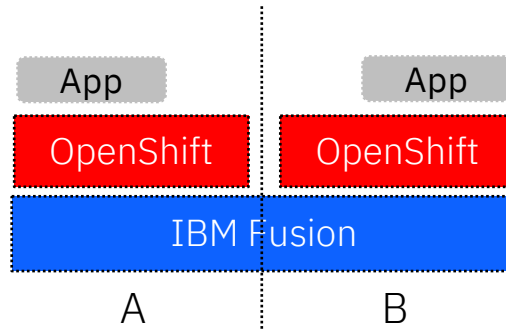
## Stretched Clusters



- Sync replication across rack/zone
- RPO of zero

## Cross Zone HA

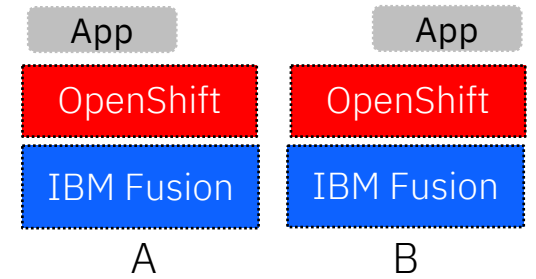
## Separate OCP, Stretched Fusion



- Sync replication across zone
- RPO of zero

## Metro DR

## Separate OCP, Separate Fusion

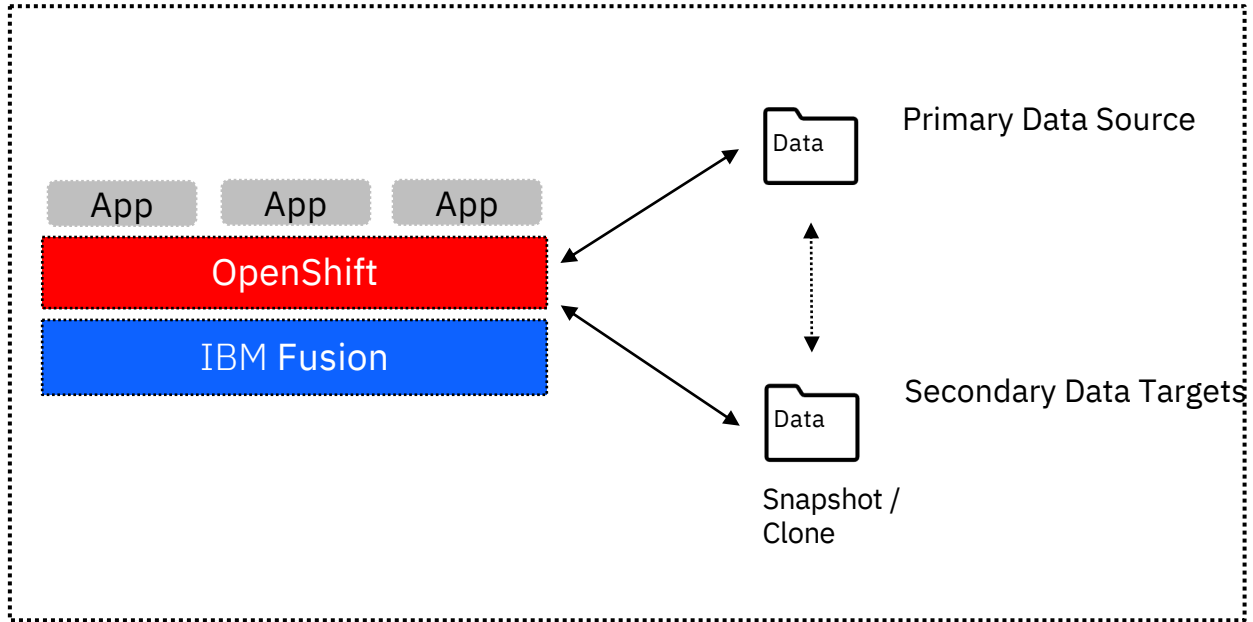


- Async replication across zone
- RPO of minutes

## Regional DR

# IBM Fusion

Ready for DevOps use cases



## DevOps workflows and management

- Integrate snapshots and clones into CI/CD pipelines
- Use production data in test/development to validate everything works when pushed to production
- Test patches before new images are pushed to production



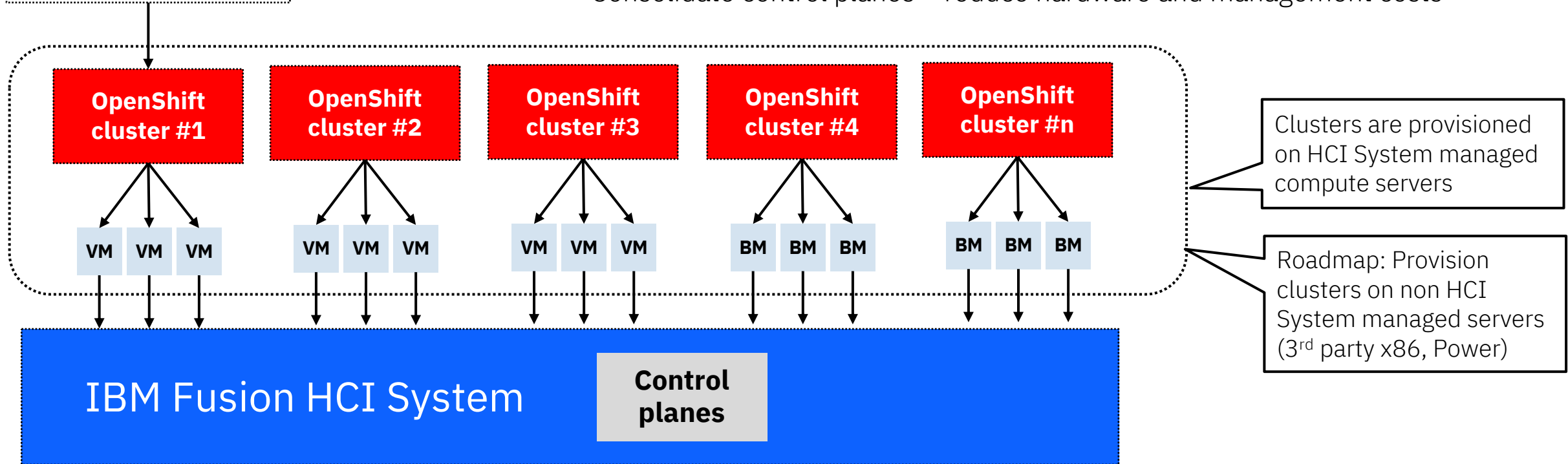
# IBM Fusion HCI

## Hosted Control Plane - multiple clusters on single appliance

### Rapidly deploy OpenShift clusters

- ✓ General-purpose applications
- ✓ Dev/test environments

- Increase cluster-to-appliance ratio, improve TCO
- Provision fresh clusters in under an hour
- Run multiple versions of OpenShift
- Improve security – isolate management from applications
- Consolidate control planes – reduce hardware and management costs



# IBM Fusion HCI

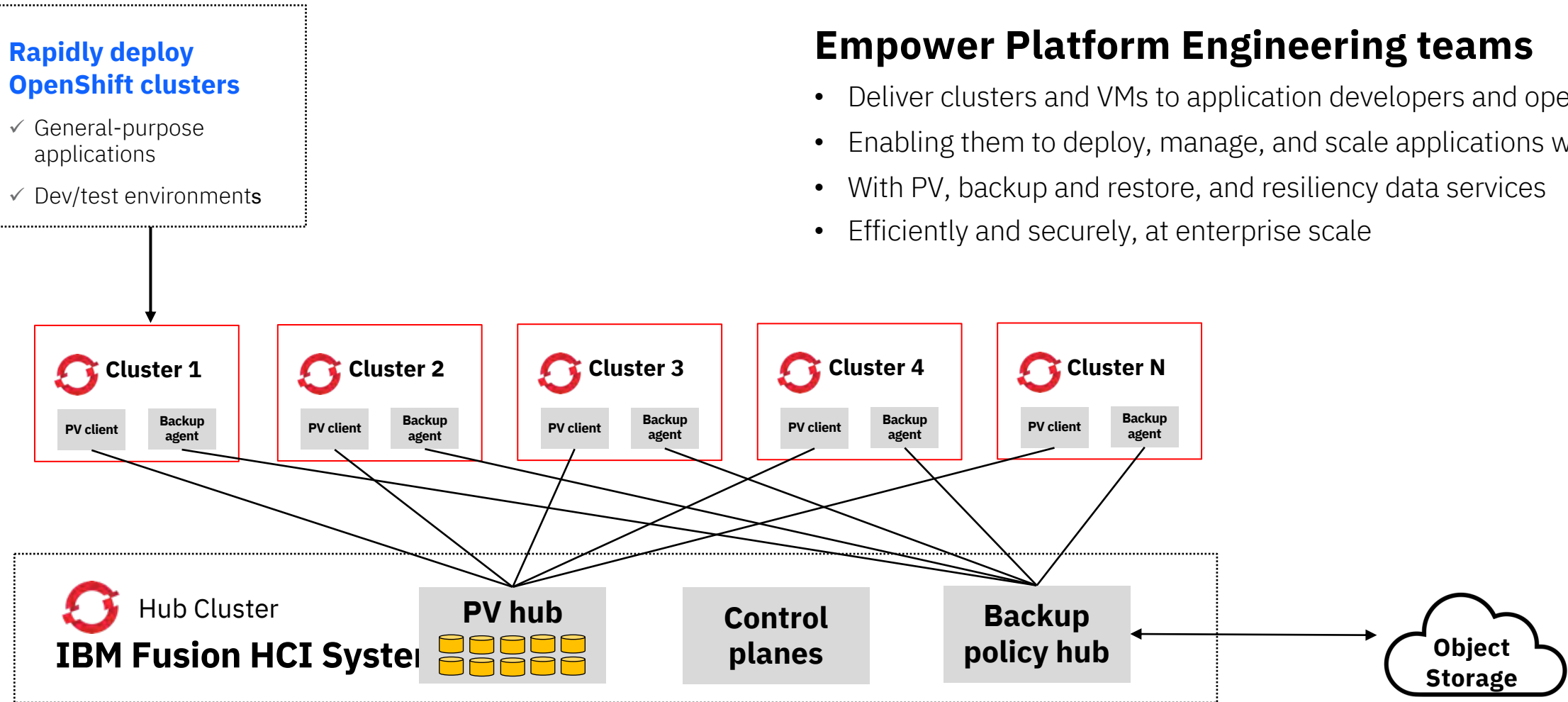
## Hub and Spoke managed clusters

### Rapidly deploy OpenShift clusters

- ✓ General-purpose applications
- ✓ Dev/test environments

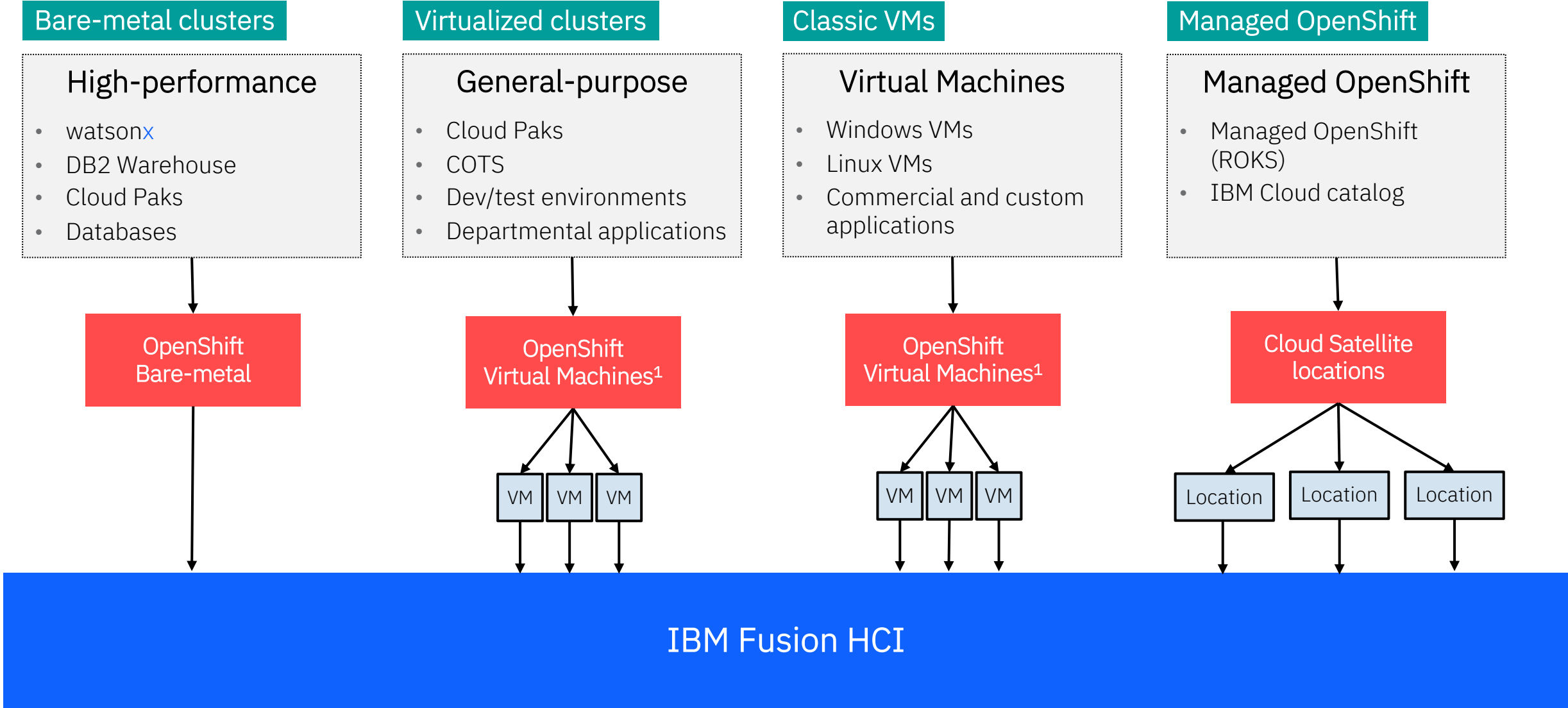
### Empower Platform Engineering teams

- Deliver clusters and VMs to application developers and operations teams
- Enabling them to deploy, manage, and scale applications with ease
- With PV, backup and restore, and resiliency data services
- Efficiently and securely, at enterprise scale



# IBM Fusion HCI

## Flexible workload patterns



<sup>1</sup> OpenShift Virtualization. KVM plus KubeVirt

# IBM Fusion

Closes the gap between expectations and fulfilment



+



**Application portability**



**Security, resiliency, and backup**



**Integration with existing storage**



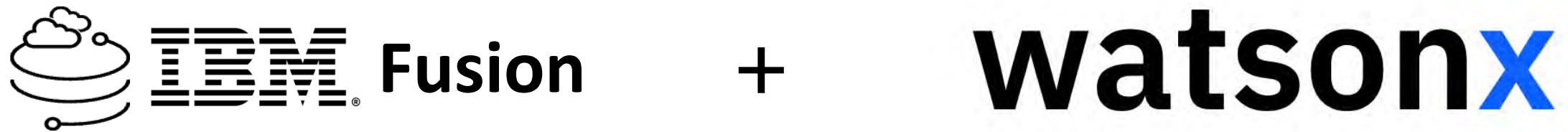
**Enterprise hardened**



**Infrastructure elasticity, agility**



**Data discovery and sharing**



Full GenAI and Foundational Model Platform running within  
your data center. From network to AI model.

## IBM POV: Four core principles to tailor generative AI for enterprise

### Open

---

→ Based on the best AI and cloud technologies available

→ Facilitating access to the innovation of the open community and multiple models

### Targeted

---

→ Designed for targeted business use cases, that unlock new value at optimal cost

→ Including curated models that can be tuned to proprietary data and company guidelines

### Trusted

---

→ Built with AI and data governance, transparency, and ethics that support increasing regulatory compliance demands

→ Providing guidance on appropriate models to leverage to create real business value with trust

### Empowering

---

→ Leveraging a platform that enables clients to customize models with their data and integrate into complex environments to move from experimentation to

→ Running anywhere, designed for scale and widespread adoption to truly create enterprise value

The platform  
for AI and data

**watsonx**

## **watsonx.ai**

Train, validate, tune,  
and deploy AI models

A next generation enterprise studio for AI builders to train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models. It enables you to build AI applications in a fraction of the time with a fraction of the data.

## **watsonx.data**

Scale AI workloads, for  
all your data, anywhere






Fit-for-purpose data store, built on an open lakehouse architecture, supported by querying, governance and open data formats to access and share data.

## **watsonx.governance**

Accelerate responsible, transparent, and explainable AI workflows

End-to-end toolkit for AI governance across the entire model lifecycle to enable responsible, transparent, and explainable AI workflows.

# IBM’s generative AI tech stack

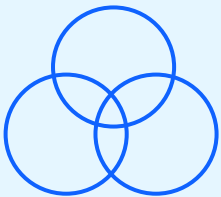
<b>AI assistants</b> 	Empower individuals to do work without expert knowledge across a variety of business processes and applications.	<b>watsonx</b> Orchestrate <b>watsonx</b> Assistant <b>watsonx</b> Code Assistant <b>watsonx</b> Orders
<b>SDKs and APIs</b> 	Use programmatic interfaces to embed watsonx platform capabilities in assistants and applications.	<b>Ecosystem integrations</b>
<b>AI and data platform</b> 	Leverage generative AI and machine learning — tuned with your data — with responsibility, transparency and explainability.	<div><b>watsonx</b> watsonx.ai watsonx.governance watsonx.data</div> <div><b>Foundation models</b> Open Source Llama 2   <i>Hugging Face</i> Geospatial   <i>Meta AI</i> Granite   <i>IBM + NASA</i> ...   <i>IBM</i></div>
<b>Data services</b> 	Access data fabric services to define, organize, manage, and deliver trusted data to train and tune models.	<b>Data fabric services</b>
<b>Hybrid cloud AI tools</b> 	Build on a consistent, scalable foundation based on open-source technology.	<b>Red Hat</b> OpenShift AI (e.g., Ray, Pytorch)



What IBM offers

watsonx assistants

Purpose-built to  
increase productivity



Tailored  
Automated  
Integrated

watsonx Orchestrate

Harness the power of AI  
and automation to free up  
individuals from tedious tasks

watsonx Assistant

Build better virtual agents,  
to deliver consistent and  
intelligent customer care

watsonx Code Assistant

Accelerate development,  
application modernization,  
and assist with IT operations

watsonx Assistant for Z

Use generative AI to transform  
engagement and interaction  
with the mainframe

watsonx BI Assistant

Get AI-powered insights in  
seconds from your personal  
business analyst and advisor

Reinventing how work gets done |  
+AI to AI+

IBM is actively  
engaging with  
enterprise clients  
across a broad  
set of business  
domains

Non-exhaustive

Customer-facing functions and experiences	HR, Finance, and Supply chain functions	IT development and operations	Core business operations
<b>Customer service</b> Empower customers to find solutions with easy, compelling experiences  Automate answers with 95% accuracy	<b>HR automation</b> Reduce manual work and automate recruiting, sourcing and nurturing job candidates  Reduce employee mobility processing time by 50%	<b>App modernization, migration</b> Generate code, tune code generation response in real time  Deliver faster development output	<b>Threat management</b> Reduce incident response times from hours to minutes or seconds  Contain potential threats 8x faster
<b>Marketing</b> Increase personalization, improve efficiency across the content supply chain  Reduce content creation costs by up to 40%	<b>Supply chain</b> Automate source to pay processes, reduce resource needs and improve cycle times  Reduce cost per invoice by up to 50%	<b>IT automation</b> Identify deployment issues, avoiding incidents, optimize application demand to supply  Reduce mean time to repair (MTTR) by 50%+	<b>Asset management</b> Optimize critical asset performance and operations while delivering sustainable outcomes  Reduce unplanned downtime by 43%
<b>Content creation</b> Ex. Enhance digital sports viewing with auto-generated spoken AI commentary  Scale live viewing experiences cost effectively	<b>Planning and analysis</b> Make smarter decisions, focus on higher value tasks with automated workflows and A.  Process planning data up to 80% faster	<b>AIOps</b> Assure continuous, cost- effective performance and connectivity across applications  Reduce application support tickets by 70%	<b>Product development</b> Ex. Expedite drug discovery by inferring structure with AI from simple molecular representations  Faster and less expensive drug discovery
<b>Knowledge worker</b> Enable higher value work, improve decision making, and increase productivity  Reduce 90% of text reading and analysis work	<b>Regulatory compliance</b> Support compliance based on requirements / risks, proactively respond to regulatory changes  Reduce time spent responding to issues	<b>Data platform engineering</b> Redesign the approach for data integration using generative AI  Reduce data integration time by 30%+	<b>Environmental intelligence</b> Provide intelligence to proactively plan and manage impact of severe weather and climate  Increase manufacturing output by 25%

# Thank you.

IBM and the IBM logo are trademarks of IBM 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 [ibm.com/trademark](https://www.ibm.com/trademark).

THIS DOCUMENT, INCLUDING BUT NOT LIMITED TO, LOSS OF DATA, BUSINESS INTERRUPTION, LOSS OF PROFIT OR LOSS OF OPPORTUNITY.

IS DISTRIBUTED “AS IS” WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IN NO EVENT, SHALL IBM BE LIABLE FOR ANY DAMAGE ARISING FROM THE USE OF THIS INFORMATION

Client examples are presented as illustrations of how those clients have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

Not all offerings are available in every country in which IBM operates.

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

© 2024 International Business Machines Corporation

