



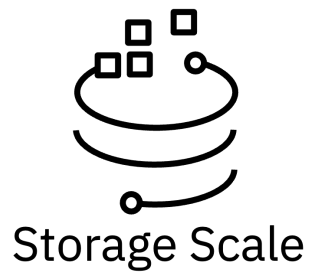
HUK Coburg – Proof of technology OpenShift workloads accessing IBM Storage System via InfiniBand

IBM Storage Scale Days 2024

March 5-7, 2024 | Stuttgart Marriott Hotel Sindelfingen

Renar Grunenberg, HUK Coburg
Alexander Saupp, IBM Client Engineering
Harald Seipp, IBM Client Engineering

Disclaimer



- IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
- IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

The motivation

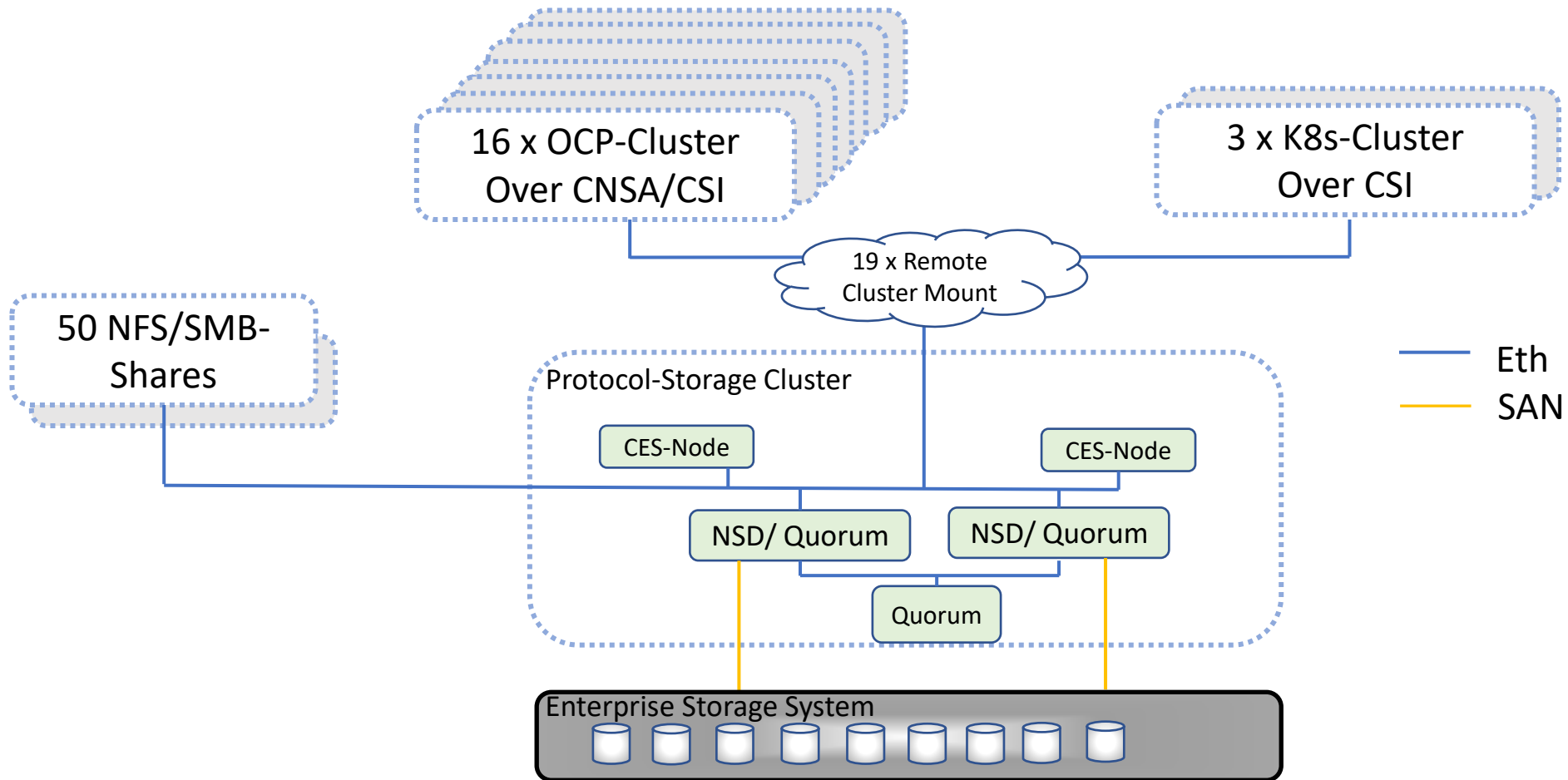
HUK design points

- Storage Scale as common data layer, leveraging and expanding to new ESS infrastructure
- Stretched IB setup as private high-speed network to reduce latency and optimize resource utilization at max
- Consolidation of disaggregated Scale Clusters to implement the goal to a real global data platform
- Shared access to data, from traditional and containerized setups
- Optimize usage of GPU datapath in the upcoming data-science Projects
- use of new technology like NVMeoF

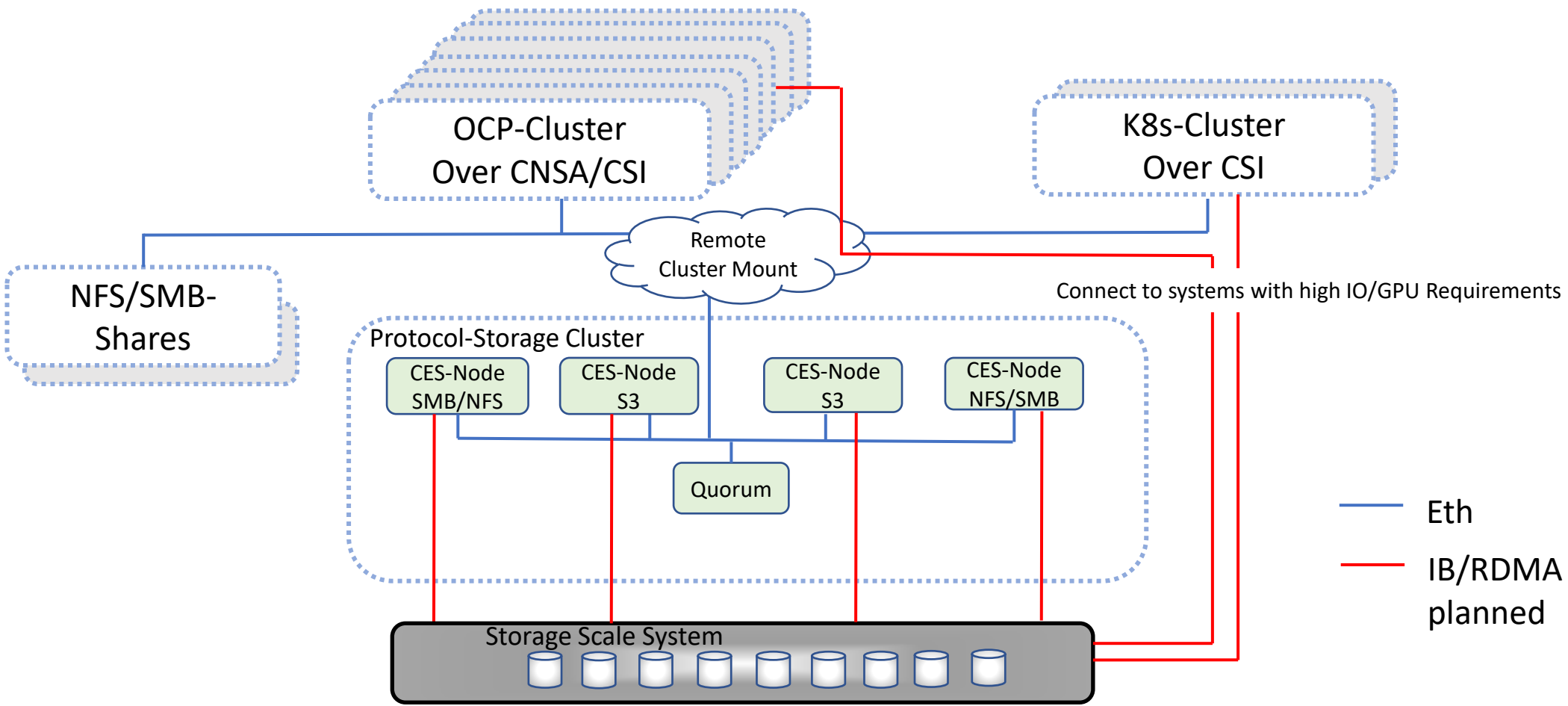
Challenges

- no IBM Support for CNSA via Infiniband Networks
- limited experience with Infiniband in OpenShift
- Permission management for shared access with rootless containers

IBM Scale design (current)



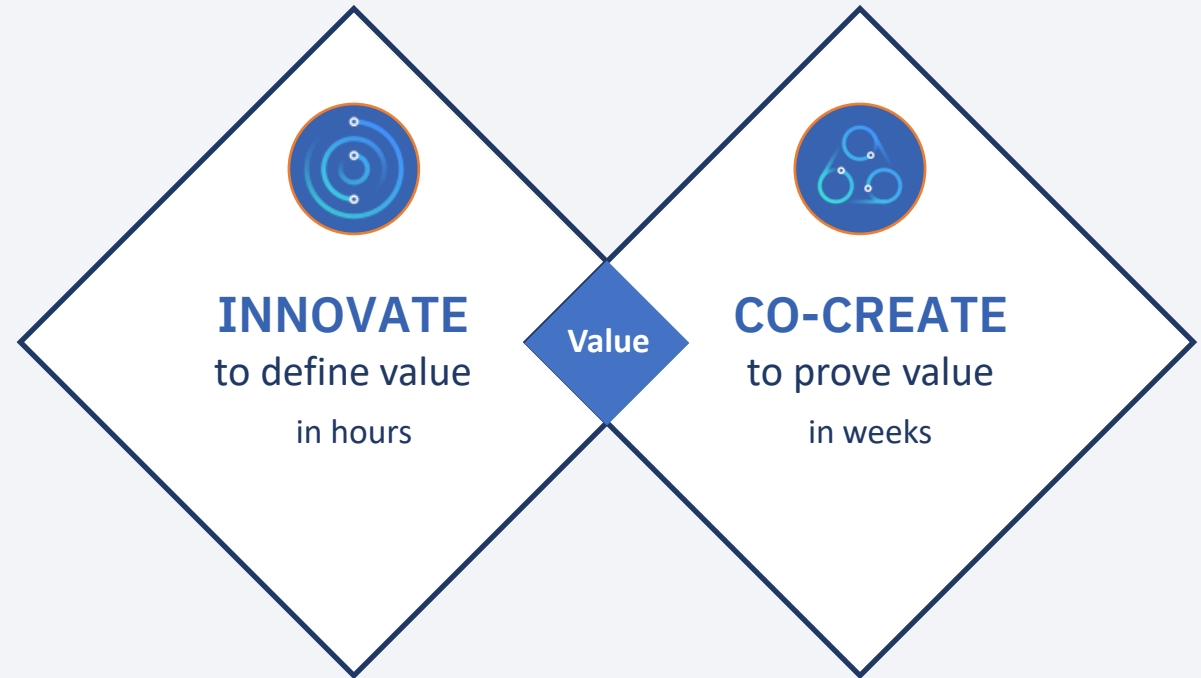
IBM Scale design (planned)



Let's create ↻
value together

Client Engineering is an investment by IBM to jointly innovate and rapidly prove solutions to your business opportunities by leveraging IBM hybrid cloud and AI technologies.

Investment in your success



You bring your business and technology context, sponsorship, subject matter experts, and data. We bring a deeply skilled multi-disciplinary squad, technical accelerators, proven method, and a memorable experience.



Lab Results

Smoke tests

The screenshot shows the Red Hat OpenShift console interface. The top navigation bar includes the Red Hat OpenShift logo and a notification: "You are logged in as a temporary administrative user. Update the cluster". The breadcrumb trail indicates the current location: "Home > Pods > Pod details".

The left sidebar contains a navigation menu with the following items: Operators, Workloads, Pods (selected), Deployments, DeploymentConfigs, StatefulSets, Secrets, ConfigMaps, CronJobs, and DaemonSets. The "Pods" menu item is expanded, showing sub-items: Administrator, Home, Operators, Workloads, Pods, Deployments, DeploymentConfigs, Secrets, ConfigMaps, CronJobs, and Jobs.

The main content area displays the details for the pod "c83f2-app1" in the "ibm-spectrum-scale" project. The pod is in a "Running" state. Below the pod name, there are tabs for "Details", "Metrics", "YAML", "Environment", "Logs", "Events", and "Terminal". The "Terminal" tab is active, showing a shell prompt "sh-4.4#" and the output of the command "mmdiag --network | tail -n 10".

```
sh-4.4# mmdiag --network | tail -n 10
GPFS Send-Queue: 0      GPFS Recv-Queue: 0
Socket Send-Queue: 0   Socket Recv-Queue: 0
Device details:
devicename  speed  mtu    duplex  rx_dropped rx_errors tx_dropped tx_errors
net1        100000 2044   full    0          0         3          0
RDMA Connections between nodes:
Fabric 0 - Device mlx5_2 Port 1 Width 4x Speed EDR lid 10
hostname  idx  CM state VS buff RDMA_CT(ERR) RDMA_RCV_MB RDMA_SND_MB VS_CT(ERR) VS_SND_MB VS_RCV_MB WAIT_CON_SLOT WAIT_NODE_SLOT
c83f0ib-ess-a-hs 0 N RTS (Y)256 0 (0) 0 0 7953 (0) 2 0 0 0
c83f0ib-ess-b-hs 0 N RTS (Y)256 0 (0) 0 0 8452 (0) 2 0 0 0
sh-4.4#
```




Installation & Configuration








Requirements

- Spin up baremetal OCP 4.12.26 via assisted installer
- Install Node Feature Discovery Operator + NVIDIA Network Operator.
Mellanox IB Switches + network cards (no GPU) - the operator is still labeled NVIDIA while not related to their GPUs
<https://docs.nvidia.com/networking/display/public/sol/rdg+for+accelerating+ai+workloads+in+red+hat+ocp+with+nvidia+dgx+a100+servers+and+nvidia+infiniband+fabric#sr-c-99399137> RDGforAcceleratingAIWorkloadsinRedHatOCPwithNVIDIADGXA100ServersandNVIDIInfiniBandFabric-Post-installationConfiguration
- Ensure to have RedHat subscriptions as per Nvidia requirements for MOFED builds
<https://docs.nvidia.com/datacenter/cloud-native/openshift/23.9.0/appendix-ocp.html#cluster-entitlement>
- Each worker node requires at least one IB port 'up'
Must be same device on all nodes (cannot be node1:ib0 and node2:ib1)
- Nvidia: No stacked master/worker nodes supported!

NVIDIA Operator

Installed Operators

Installed Operators are represented by ClusterServiceVersions within this Namespace. For more information, see the [Understanding Operators documentation](#) or create an Operator and ClusterServiceVersion using the [Operator SDK](#).

Name	Namespace	Managed Namespaces	Status	Last updated	Provided APIs
 Node Feature Discovery Operator 4.12.0-202310241244 provided by Red Hat	 openshift-nfd	All Namespaces	 Succeeded Up to date	🕒 13 Nov 2023, 11:14	NodeFeatureDiscovery NodeFeatureRule
 NVIDIA Network Operator 23.5.0 provided by NVIDIA	 nvidia-network-operator	 nvidia-network-operator	 Succeeded Up to date	🕒 11 Oct 2023, 11:25	HostDeviceNetwork IPoIBNetwork MacvlanNetwork NicClusterPolicy

Two NVIDIA CRs:

- The NicClusterPolicy will trigger MOFED pods per node, which will build against the running kernel (after node reboot)
 - The IPoIBNetwork (namespace context) defines the subnet to use (auto assign IP from range vs. definition of static IPs via labels (preferred))
-
- Troubleshooting the Network Operator / labeling / MOFED kernel builds failing / NicClusterPolicy staying in "NotReady" state
-> solved with NVIDIA expertise & advise

IBM Storage Scale CNSA installation

Scale Preparations – business as usual

- MCO <https://www.ibm.com/docs/en/scalecontainernative?topic=premise-red-hat-openshift-configuration>
- IBM Registry Pull secret <https://www.ibm.com/docs/en/scalecontainernative?topic=cluster-image-pull-secrets-optional>
- Configure Scale cluster <https://www.ibm.com/docs/en/scalecontainernative?topic=cluster-install>
- Create Rest API / GUI users on storage cluster <https://www.ibm.com/docs/en/scalecontainernative?topic=cluster-premise>
- Apply quota, .. to storage cluster as per above link

[..]

Annotate nodes to use the IB network (available default option in CNSA)

```
oc annotate node c83f2-app1 'scale.spectrum.ibm.com/daemon-network={ "name": "ipoibnetwork-scale", "ips": [ "192.168.0.31/24" ] }' -overwrite
oc annotate node c83f2-app2 'scale.spectrum.ibm.com/daemon-network={ "name": "ipoibnetwork-scale", "ips": [ "192.168.0.32/24" ] }' -overwrite
oc annotate node c83f2-dan4 'scale.spectrum.ibm.com/daemon-network={ "name": "ipoibnetwork-scale", "ips": [ "192.168.0.33/24" ] }' -overwrite
```

business as usual [..]

Hack (still): enable RDMA within OCP Scale Cluster [..]

```
mmchconfig verbsRdma=enable -N c83f2-appX
mmchconfig verbsRdmaCm=disable -N c83f2-appX
mmchconfig verbsRdmaSend=yes -N c83f2-appX
mmchconfig verbsPorts=mlx5_2 -N c83f2-appX
```

```
mmshutdown; mmstartup # or destroy pods
```

Outlook

Testing

scheduling of MOFED rebuild / Scale starts

OpenShift versions

..

5.2.0 (April):

Tech Preview

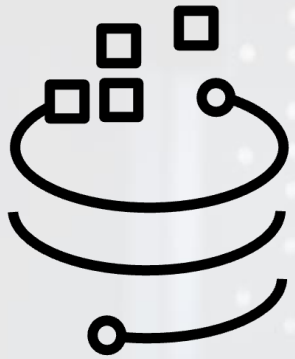
verbs parameters via CR

5.2.1 or 5.2.2

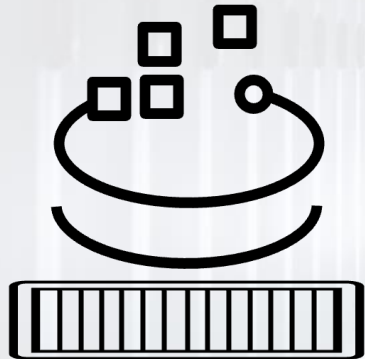
GA

TBD: additional features such as port autodetection / multiple adapters per Node

Thank you for using



Storage Scale



Storage Scale
System