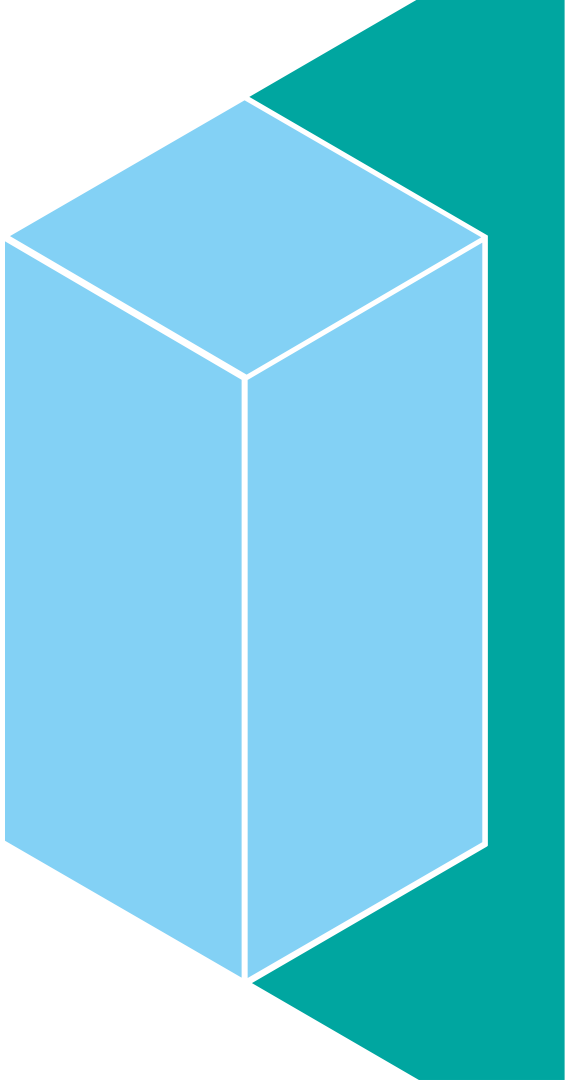
IBM
Spectrum
MPI

IBM
Spectrum
LSF



Running InfiniBand and GPU Accelerated MPI Applications in a Kubernetes Environment

Joshua Hursey, Scott Miller
Spectrum MPI
jhursey@us.ibm.com, scott.miller1@ibm.com

High Level Requirements

- **Goal:**
  - Run a “traditional” HPC batch script that uses mpirun in Kubernetes with minimal/no changes.
- **MPI Requires:**
  - Direct access to GPUs and Infiniband network devices
  - Must be able to ‘ssh’ between pods to start daemons and launch processes
  - The pods have stable names that can be used in a hostfile
  - All pods must be running when the HPC batch script is executed
  - The job runs as the user, and the processes in the pods run as the user
- **Container Requirements:**
  - Nvidia CUDA user space drivers
    - Confirm version compatibility with the host.
  - Mellanox MOFED
    - Confirm version compatibility with the host.
  - Open MPI or Spectrum MPI installed
- **Kubernetes Configuration Requirements:**
  - Mellanox Plugin for InfiniBand support
  - Nvidia Plugin for GPU support

```
shell$ kubectl describe node 172.16.3.18
Name:                172.16.3.18
Roles:               worker
...
Capacity:
 cpu:                160
 ephemeral-storage:  956642692Ki
 hugepages-1Gi:      0
 hugepages-2Mi:      0
 memory:             598005824Ki
 nvidia.com/gpu:     4
 pods:               500
 rdma/hca:           1k
...
```

Vanilla Kubernetes (K8s) Environment
Prefer fewer big pods instead of many little pods

Compute Nodes = K8s StatefulSet

- Pods carry state with a unique, predictable name

```
apiVersion: apps/v1
kind: StatefulSet
metadata:
  name: my-cnode
spec:
  podManagementPolicy: Parallel
  replicas: 4
... # Pod dnsConfig: searches
```

Two versions:
- **ssh based**
  - Runs as root
- **kubectl based**
  - Runs as the user
  - kubectl used to move between pods

DNS = K8s Headless Service

- DNS between pods. No external IP or load balancing.

```
apiVersion: v1
kind: Service
metadata:
  name: my-hpc-cluster
spec:
  clusterIP: None
  selector:
    app: my-hpc-compute-nodes
```

$ ping my-cnode-1.my-hpc-cluster

Hostlist = K8s ConfigMap

- Dynamically created & mounted into each container.

```
apiVersion: v1
kind: ConfigMap
metadata:
  name: my-mpi-hostfile
data:
  hostfile: |-
    my-cnode-0
    my-cnode-1
    my-cnode-2
    my-cnode-3
```

OMPI_MCA_orte_default_hostfile=
/opt/mpi/etc/hostfile

```
volumeMounts:
- name: mpi-hostfile
  mountPath: /opt/mpi/etc
volumes:
- name: mpi-hostfile
  configMap:
    name: my-mpi-hostfile
```

Launch Node = K8s Job + initContainers

- `initContainers`: Waits for StatefulSet pods to become running & DNS visible before starting the job

```
kind: Job
      initContainers:
      - name: my-mpi-waiter
        image: jjhursey/k8s-waitfor
        command: [ ... ]
      containers:
      - name: my-mpi-launcher
        image: my_mpi:test_with_k8s
        command: ["./my-batch.sh"]
...
```

Deadlock is possible!

LSF Connector for Kubernetes ParallelJob Controller/CRD

https://github.com/IBMSpectrumComputing/lsf-kubernetes

```
spec:
  name: pj-mpi-kubexec
  description: This is a parallel job for MPI.
  schedulerName: lsf
  priority: 100
  resizable: false
  mpiOptions:
    # (*) Enable MPI feature set. Fix up environment for MPI
    enable: true
    # (*) Image to use when waiting for Pods to become 'Ready'
    waiterImage: "jjhursey/k8s-waitfor:ppc64le"
    # 'false' : mount in user credentials from the host
    # 'true'  : use the user credentials inside the container image
    containedUser: false
    # 'false' : create ConfigMap with hostnames
    #           OMPI_MCA_orte_default_hostfile=/opt/mpi/etc/hostfile
    hostfileDisable: false
    # A Service is required for the waiterImage
    # 'false' : Setup Headless Service for all pods (expose port 22 internally)
    # 'true'  : User must then setup their own Service
    serviceDisable: false
    # 'false' : Enable DNS configuration in the "StatefulSet" (pods managed by LSF)
    # 'true'  : Do not insert DNS config to pod definition
    # Note that this is not needed for a 'kubexec' style launch
    dnsDisable: true
    # 'false' : Setup Open/Spectrum MPI to use kubexec instead of ssh
    #           OMPI_MCA_plm_rsh_agent=/opt/mpi/bin/kubexec.sh
    kubexecDisable: false
```

- LSF as the HPC aware scheduler for Kubernetes
  - Does not schedule any pod until they can all be scheduled (resizable: false)
  - Robust batch queuing and job management
  - Seamlessly connect the various k8s components
- Tech. Preview: "**mpiOptions**" to activates the k8s components needed to support MPI apps.

```
apiVersion: ibm.com/v1alpha1
kind: ParallelJob
metadata:
  name: pj-mpi-kubexec
  namespace: default
  annotations:
    lsf.ibm.com/queue: "priority"
    lsf.ibm.com/user: "myuser"
    lsf.ibm.com/gpu: "num=1"
```