National Supercomputing Centre

Kubernetes Platform in ASPIRE2A

M Teguh Satria

Assistant System Manager, NSCC

26 July 2022

Kubernetes Platform in ASPIRE2A

Outline

- Background / Motivation
- Cluster Configuration
- Supported Workloads
- User and Security Policy
- Future Improvement

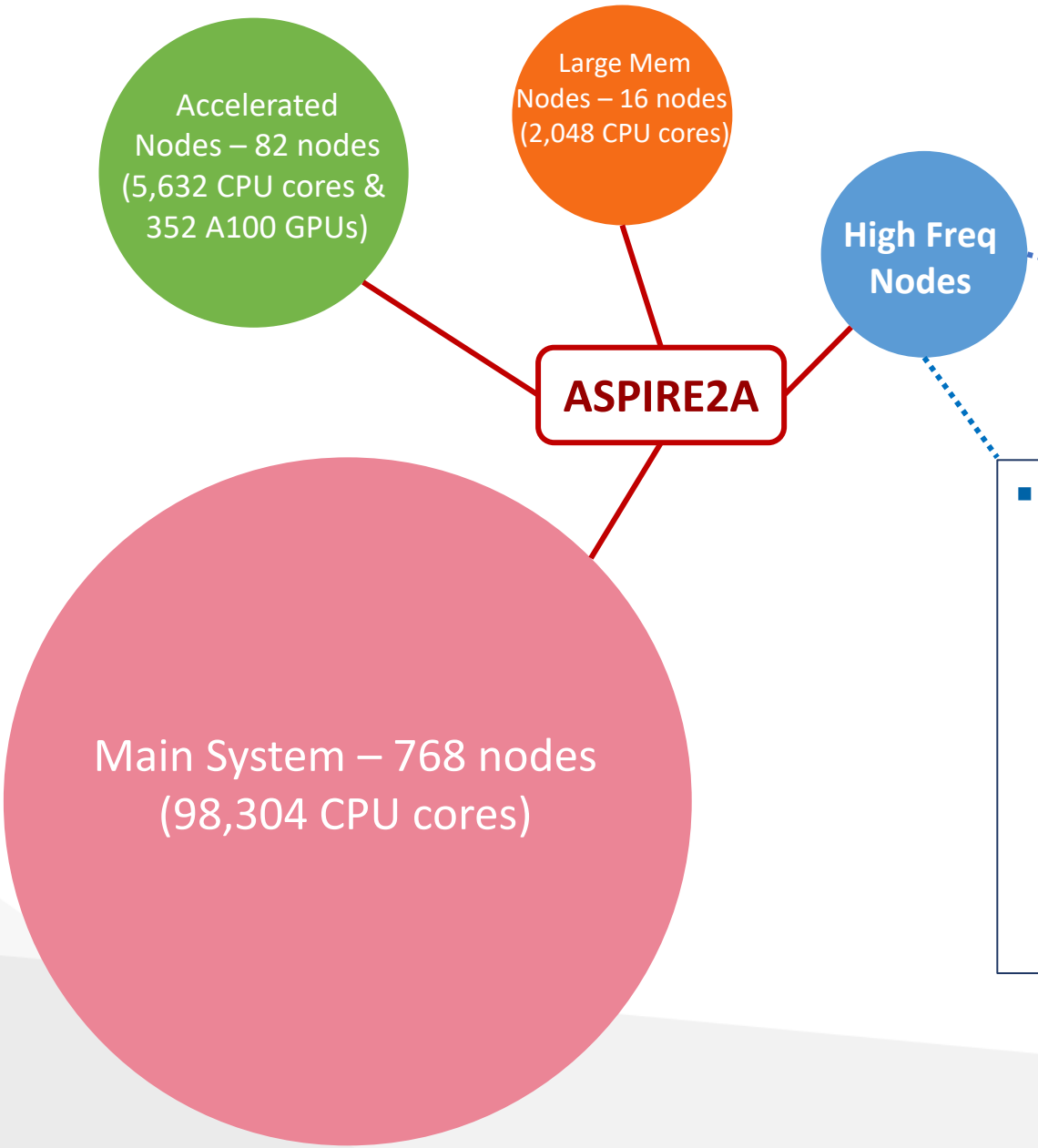
Background / Motivation

- To provide more option to run containerized applications.
- To handle supporting application for the main HPC applications i.e. small-size database.
- To support more type of workloads.

Environment

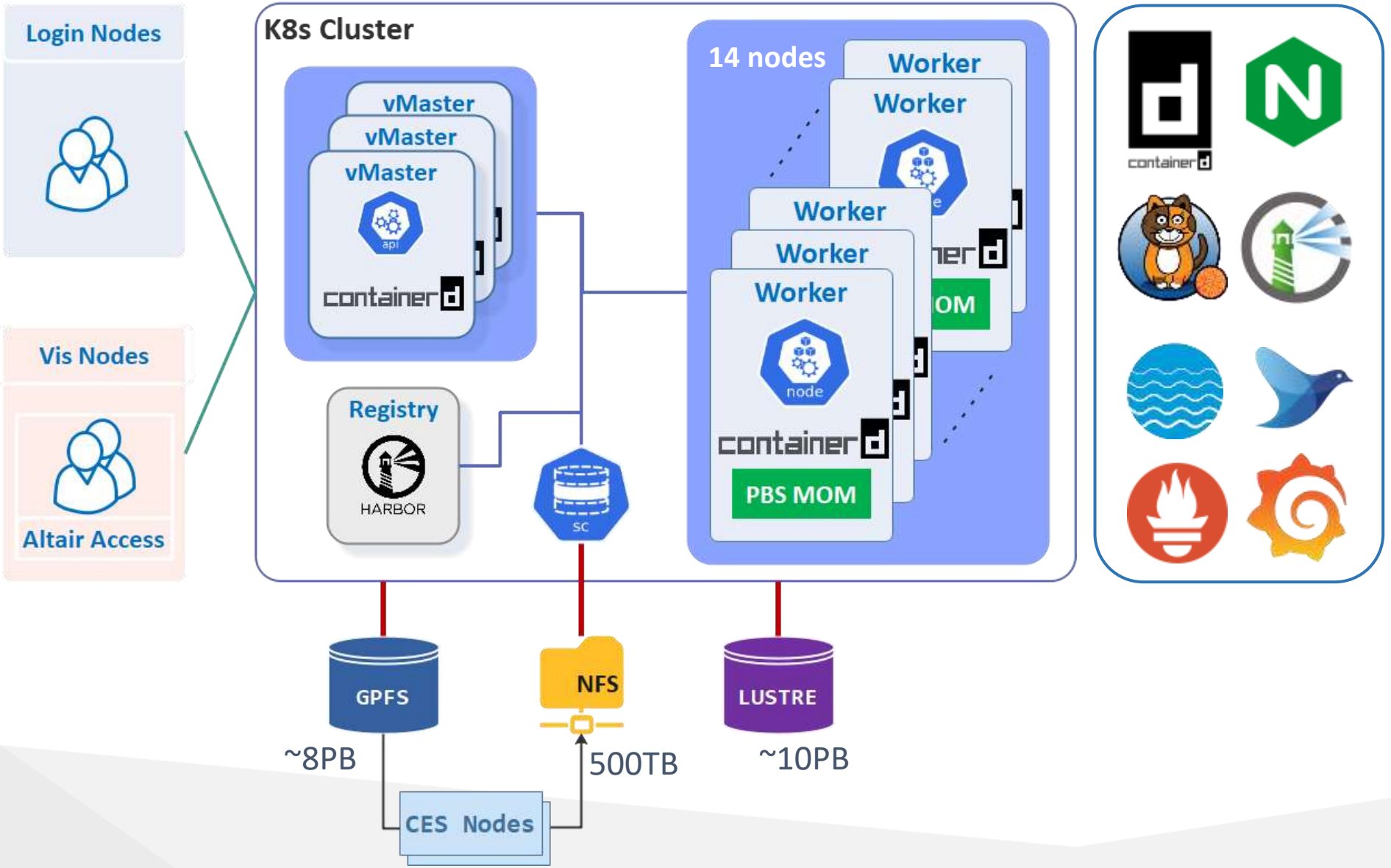
- The kubernetes cluster is deployed within ASPIRE2A system architecture – an on-premise HPC system.
  - Login nodes and Viz nodes are the main working spaces for users.
  - Altair PBS Pro and Altair Budget are the main scheduler and accounting system.
  - Parallel file system storage – Lustre and GPFS.
  - Security policy for HPC system – jobs must run as user's uid.

Cluster Configuration



- 16 nodes (1024 CPU cores) with specs :
  - 2x AMD Milan 75F3 CPU 32-core, up to 4GHz
  - 512 GB RAM
  - 2x 480GB SSD
  - 50GbE interconnect
  - Slingshot 100Gb interconnect.
  - RHEL 8.4 OS

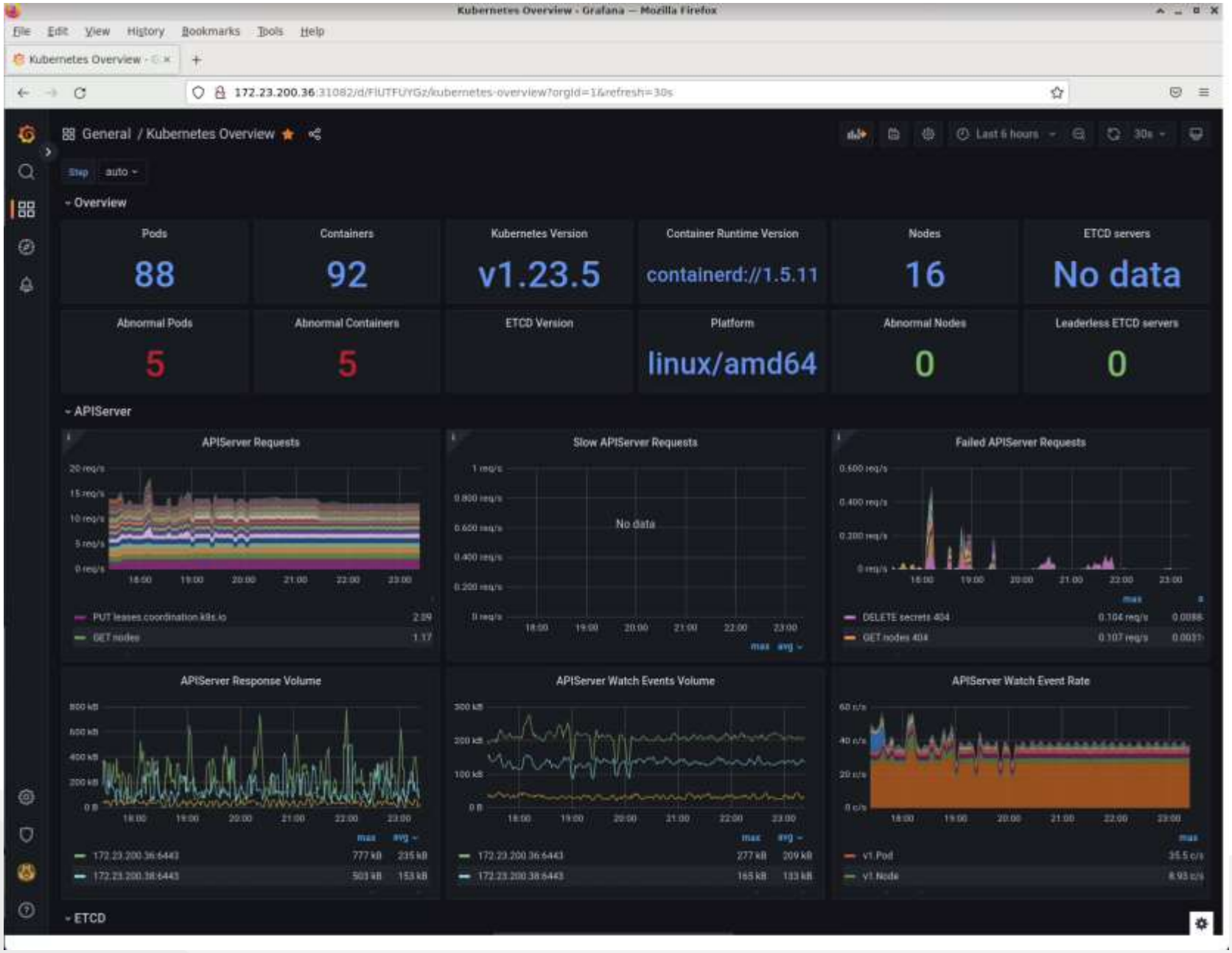
Cluster Configuration



Cluster Configuration

```
[root@asp2a-k8s-vmaster01 ~]# kubectl get nodes
NAME                STATUS                        ROLES                  AGE   VERSION
asp2a-hfn003        Ready,SchedulingDisabled      <none>                 18d   v1.23.5
asp2a-hfn004        Ready,SchedulingDisabled      <none>                 19d   v1.23.5
asp2a-hfn005        Ready,SchedulingDisabled      <none>                 18d   v1.23.5
asp2a-hfn006        Ready,SchedulingDisabled      <none>                 18d   v1.23.5
asp2a-hfn007        Ready,SchedulingDisabled      <none>                 18d   v1.23.5
asp2a-hfn008        Ready,SchedulingDisabled      <none>                 18d   v1.23.5
asp2a-hfn009        Ready,SchedulingDisabled      <none>                 18d   v1.23.5
asp2a-hfn010        Ready,SchedulingDisabled      <none>                 18d   v1.23.5
asp2a-hfn011        Ready,SchedulingDisabled      <none>                 18d   v1.23.5
asp2a-hfn012        Ready,SchedulingDisabled      <none>                 18d   v1.23.5
asp2a-hfn013        Ready,SchedulingDisabled      <none>                 18d   v1.23.5
asp2a-hfn014        Ready,SchedulingDisabled      <none>                 19d   v1.23.5
asp2a-hfn015        Ready                         <none>                 21h   v1.23.5
asp2a-k8s-vmaster01 Ready                         control-plane,master   23d   v1.23.5
asp2a-k8s-vmaster02 Ready                         control-plane,master   23d   v1.23.5
asp2a-k8s-vmaster03 Ready                         control-plane,master   23d   v1.23.5
[root@asp2a-k8s-vmaster01 ~]#
```

Cluster Configuration

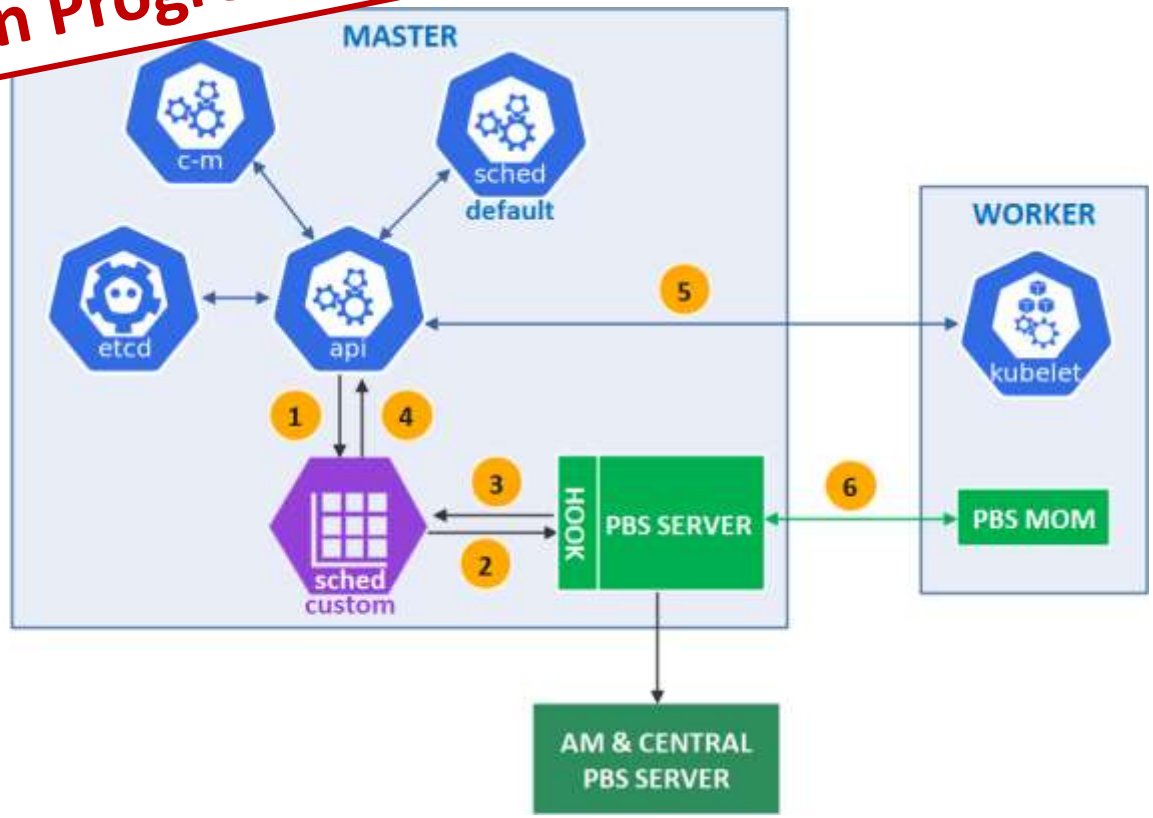


Cluster Configuration

```
[root@asp2a-k8s-vmaster01 ~]# kubectl get svc -A
NAMESPACE       NAME                                                 TYPE           CLUSTER-IP       EXTERNAL-IP     PORT(S)                      AGE
default         kubernetes                                           ClusterIP      10.96.0.1        <none>          443/TCP                      23d
ingress-nginx   ingress-nginx-controller                             LoadBalancer   10.108.90.185    172.23.200.52   80:30637/TCP,443:32425/TCP   6d5h
ingress-nginx   ingress-nginx-controller-admission                   ClusterIP      10.97.222.37     <none>          443/TCP                      6d5h
kube-system     kube-dns                                             ClusterIP      10.96.0.10       <none>          53/UDP,53/TCP,9153/TCP       23d
kube-system     stable-kube-prometheus-sta-coredns                   ClusterIP      None             <none>          9153/TCP                     7d19h
kube-system     stable-kube-prometheus-sta-kube-controller-manager   ClusterIP      None             <none>          10257/TCP                    7d19h
kube-system     stable-kube-prometheus-sta-kube-etcd                 ClusterIP      None             <none>          2379/TCP                     7d19h
kube-system     stable-kube-prometheus-sta-kube-proxy                ClusterIP      None             <none>          10249/TCP                    7d19h
kube-system     stable-kube-prometheus-sta-kube-scheduler            ClusterIP      None             <none>          10259/TCP                    7d19h
kube-system     stable-kube-prometheus-sta-kubelet                   ClusterIP      None             <none>          10250/TCP,10255/TCP,4194/TCP 7d19h
monitoring      alertmanager-operated                                ClusterIP      None             <none>          9093/TCP,9094/TCP,9094/UDP   7d19h
monitoring      prometheus-operated                                  ClusterIP      None             <none>          9090/TCP                     7d19h
monitoring      stable-grafana                                       NodePort       10.107.165.233   <none>          80:31082/TCP                 7d19h
monitoring      stable-kube-prometheus-sta-alertmanager              ClusterIP      10.104.221.80    <none>          9093/TCP                     7d19h
monitoring      stable-kube-prometheus-sta-operator                  ClusterIP      10.96.56.187     <none>          443/TCP                      7d19h
monitoring      stable-kube-prometheus-sta-prometheus                NodePort       10.107.114.224   <none>          9090:31710/TCP               7d19h
monitoring      stable-kube-state-metrics                            ClusterIP      10.102.86.196    <none>          8080/TCP                     7d19h
monitoring      stable-prometheus-node-exporter                      ClusterIP      10.105.74.161    <none>          9100/TCP                     7d19h
[root@asp2a-k8s-vmaster01 ~]#
```

Integration with PBS Pro

Work in Progress !



Supported Workloads

	Job	CronJob	Deployment	StatefulSet
Purpose	Run pod(s) once until completion.	Run Jobs on repeating schedule.	Applications that are persistent and stateless (does not care which network it is using, and it does not need permanent storage). E.g. webui-based application.	Applications that are persistent and stateful (ensure that pods can be reached through unique identities – like hostname, IP – that will not change). E.g. databases.
Features / Additional Parameters	parallelism – run multiple pods in parallel. activeDeadlineSeconds – time limit (walltime). suspends – to holds the Job (delete active pods if any). completions – number of successful pods in order to mark the job is completed.	Similar like Jobs. schedule – repeating time, i.e. "*/1 * * * *"	replicas – to guarantee the availability of a specified number of identical pods. Rolling updates and Rollback.	Name replicas with ordinal index. volumeClaimTemplates – template for PVC. Unlike the Deployment, each pod in StatefulSet will has its own PVC.

User and Security Policy

- Kubernetes service is available on request basis.
  - SSL cert and kubeconfig will be provided upon on-boarding.
- Interfaces to users are via Login nodes (command line) and Viz nodes (remote desktop).
- Role-based Access Control (RBAC) is enabled.
  - Projects/Groups have their own respective namespaces.
  - Project/Group lead allows to control the role of project's member in their own namespace.
  - Namespaces will have resource limit.
- Pod Security Policy (PSP) is enabled.
  - Default – run as user's uid and gid, allow to mount all storage (GPFS, Lustre, PV).
  - Restricted – allowed to run as root, can only mount PersistentVolume.
  - Privileged – designed for applications managed by Administrator.

Future Improvement

- Adding GPU nodes.
- Bare metal load balancer.
- More options on storage:
  - Object storage.
  - CSI-based StorageClass.

National Supercomputing Centre

Thank You

contact@nscc.sg