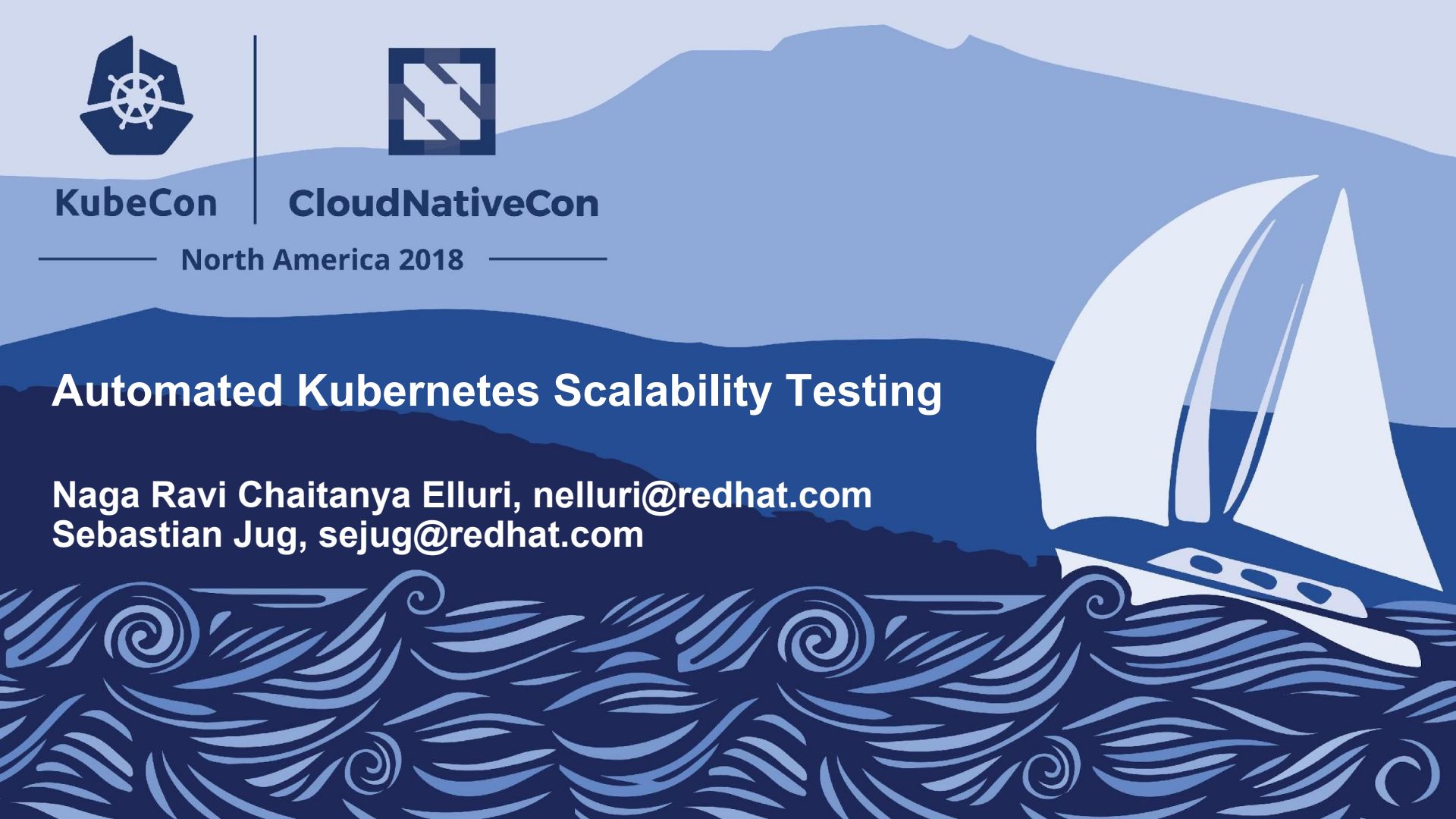


Automated Kubernetes Scalability Testing

Naga Ravi Chaitanya Elluri, nelluri@redhat.com
Sebastian Jug, sejug@redhat.com

Who are we?

- Performance and Scalability Team at Red Hat working on OpenShift.
- Pushing the limits of OpenShift Scalability.

OpenShift Scalability

- Does OpenShift support running applications at scale?
- What are the cluster limits?
- How can we tune a cluster to get maximum performance?
- Challenges?

INFRASTRUCTURE

Scale Lab

- Operate Red Hat's products at scale.
- Shared on-demand resources.

Cluster Availability

- 2000 nodes cluster every release.
- 250 nodes cluster every sprint.

AUTOMATION PIPELINE AND TOOLING

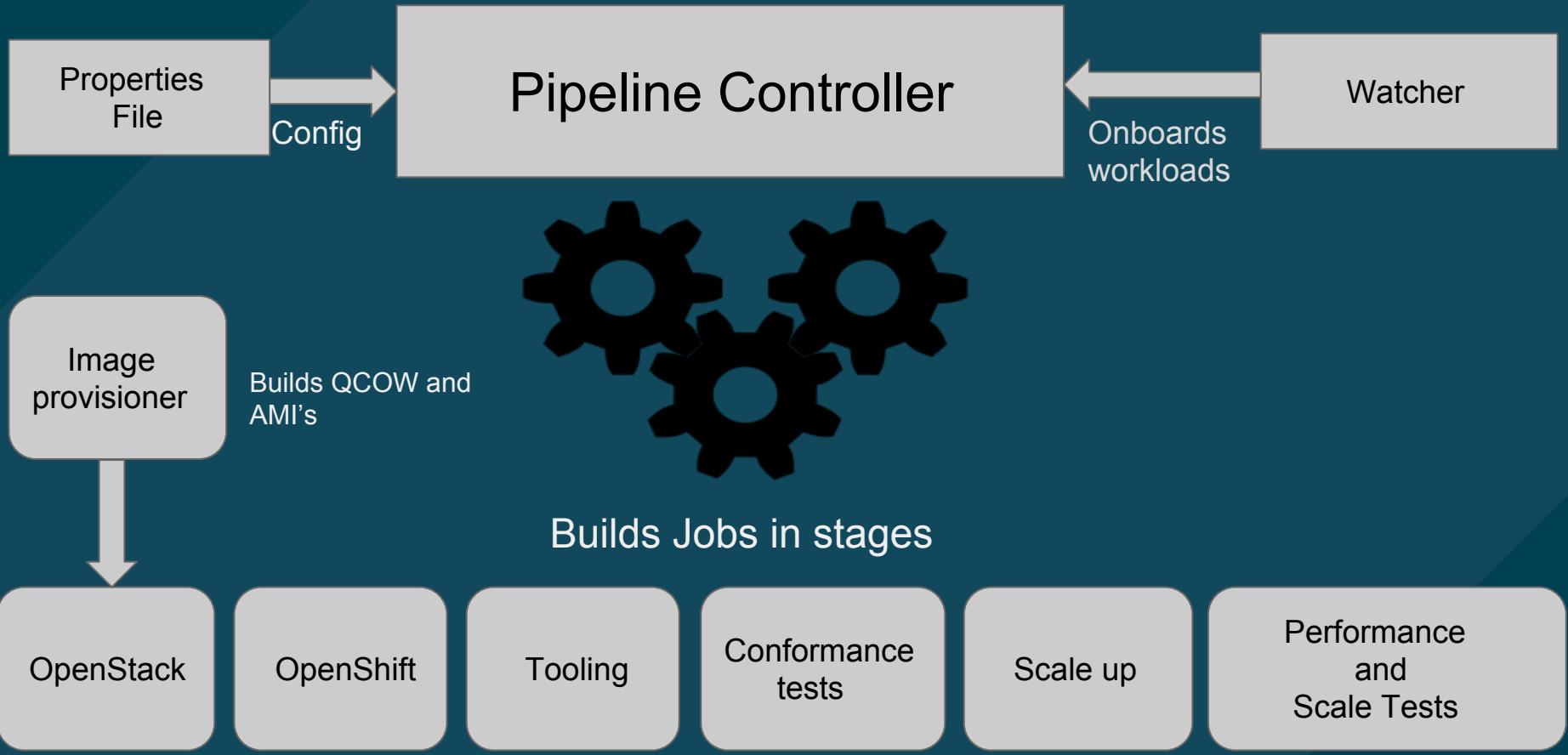

Properties File
Config
Pipeline Controller
Watcher
Onboards workloads
Image provisioner
Builds QCOW and AMI's
Builds Jobs in stages
OpenStack
OpenShift
Tooling
Conformance tests
Scale up
Performance and Scale Tests

Image Provisioner

- Watches for new OpenShift code bits.
- Builds AMI’s and QCOW images.
- Reduces the install time.

Image Validator

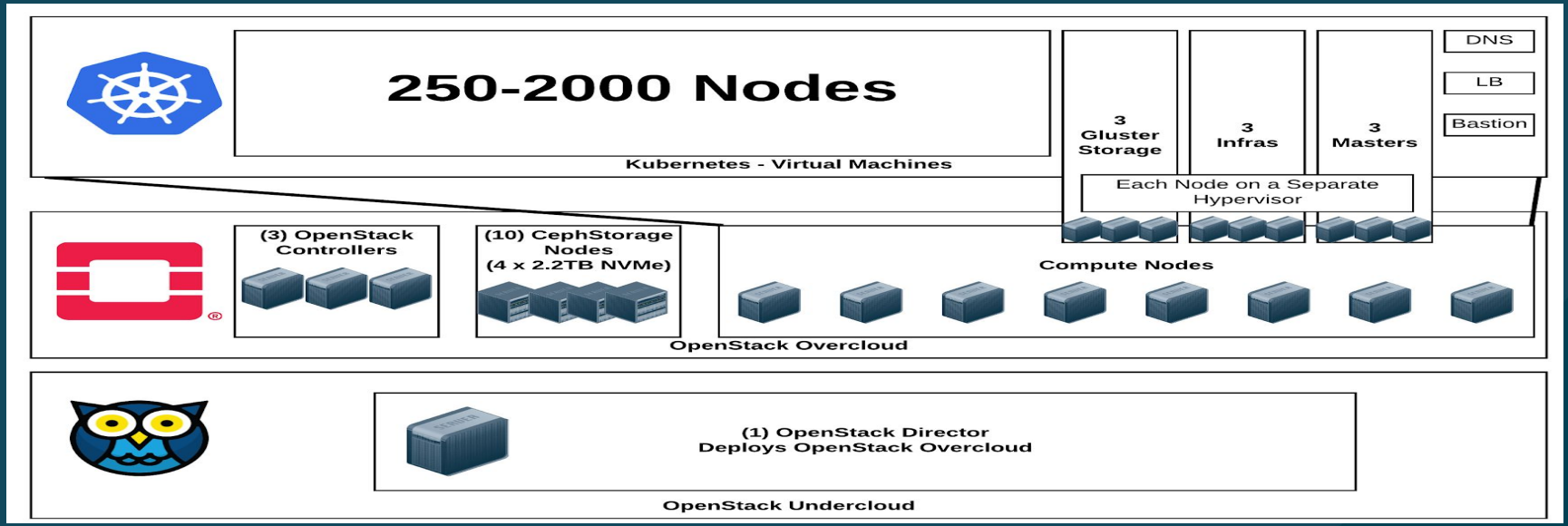
- Validates the images by installing an all-in-one node OpenShift.

Pipeline Controller and Watcher

- Parses the properties file and builds the Jobs.
- Concurrent Jobs on both public and private clouds.
- Watcher creates and/or updates the Jobs when new templates are checked in.

OpenStack

- Allows us to scale the cluster to a higher node count.
- Support for OpenShift-on-OpenStack.



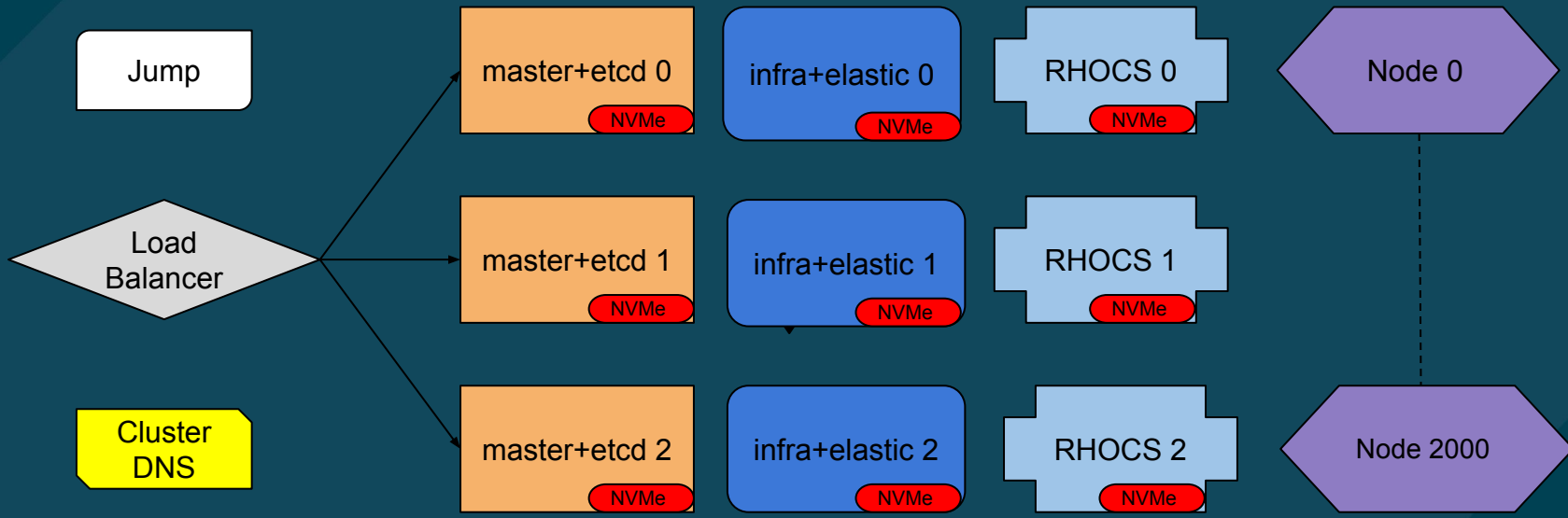
Physical Hardware	Threads	RAM (MiB)	Disk (GB)	Machines	
1029P	64	256000	480	74	
1029U	72	262144	63	10	
Total available resources:	**4480**	**17920000**	**33600**	**84**	

Scale-CI Hardware

VM Role	vCPUs	RAM (MiB)	Disk	Instances	Flavor name
Cluster DNS	1	1740	71	**1**	m1.small
Load Balancer	4	16128	96	**1**	m4.xlarge
Master + etcd	16	124672	128	**3**	r4.4xlarge-pci
Container Native Storage	16	65280	200	**3**	m4.4xlarge-pci
Infra + Elastic	40	163584	256	**3**	m4.10xlarge-pci
Jump node	16	65280	200	**1**	m4.4xlarge
Node (small)	2	7936	96	**2000**	m4.large
Total cloud resources:	**4237**	**17015756**	**194119**	**2012**	

OpenShift Overview

Master = r4.4xlarge (16/122)
Infra = m4.10xlarge (40/160)
CNS = m4.4xlarge (16/64)
Nodes = m4.large (2/8)



redhat.

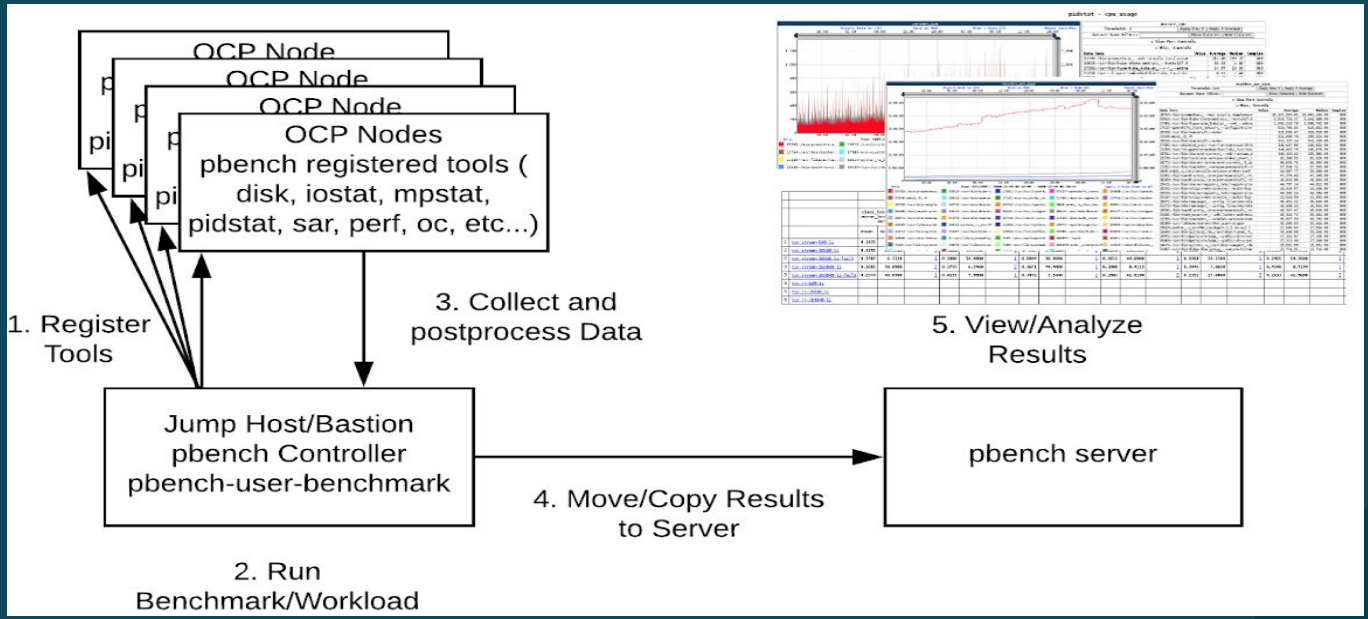
Tooling

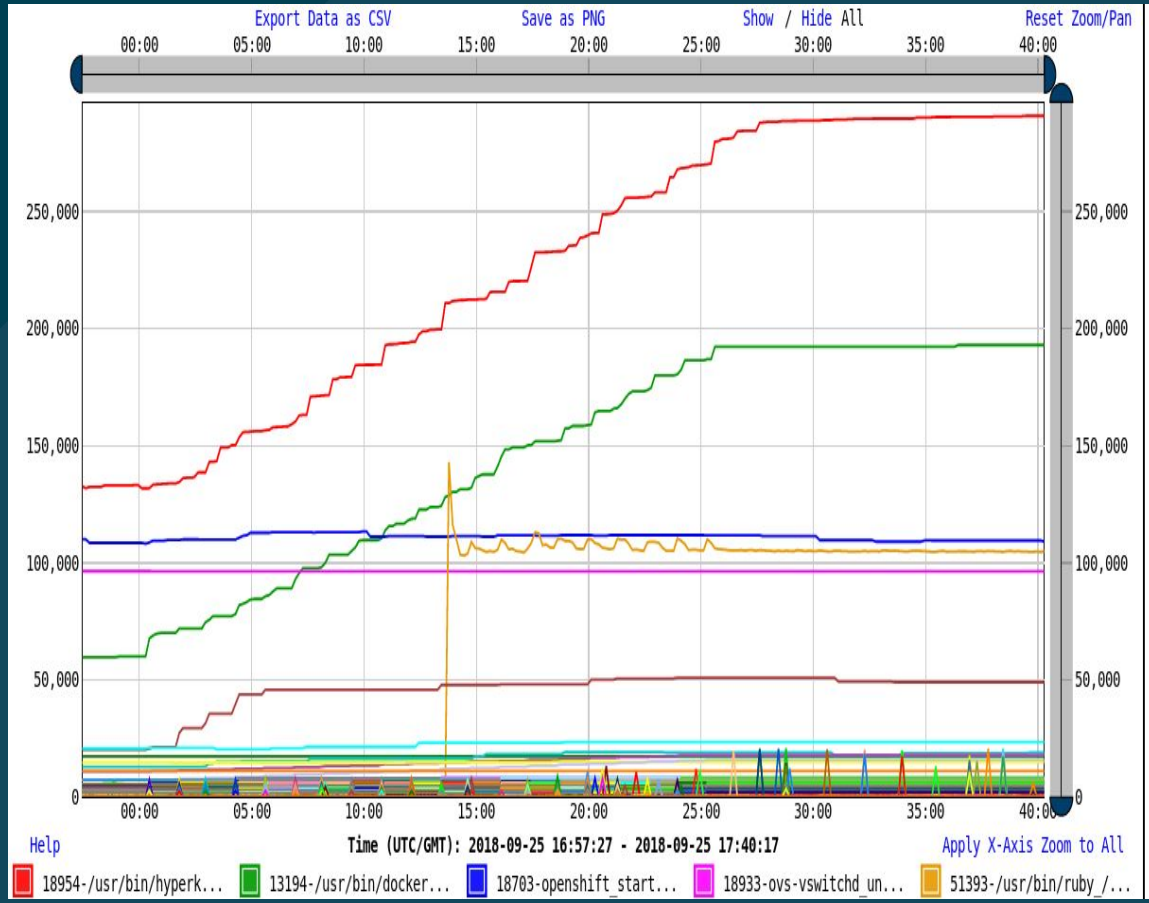
Cluster Loader - Deploys large numbers of various objects to a cluster.

```
provider: local
ClusterLoader:
  cleanup: true
  projects:
    - num: 1
      basename: clusterproject
      tuning: default
      ifexists: delete
      pods:
        - num: 1000
          image: gcr.io/google_containers/pause-amd64:3.0
          basename: pausepods
          file: pod-pause.json
  tuningsets:
    - name: default
      pods:
        stepping:
          stepsize: 50
          pause: 60
        ratelimit:
          delay: 0
```

Pbench

- A Benchmarking and Performance Analysis Framework.
- Consists of three sub-systems: pbench-agent, pbench-server and pbench-webserver.





Threshold: 100 | Apply Max Y | Apply Y Average

Dataset Name Filter: | Show Datasets | Hide Datasets

+ View Port Controls

+ Misc. Controls

Data Sets	Value	Average	Median	Samples
18954-/usr/bin/hyperkube_kubelet_--v=2_--addre		226,253.27	233,150.00	258
13194-/usr/bin/dockerd-current_--add-runtime_d		143,885.12	154,960.00	258
18703-openshift_start_network_--config=/etc/or		110,939.46	111,408.00	258
18933-ovs-vswitchd_unix:/var/run/openvswitch/d		96,485.47	96,484.00	258
51393-/usr/bin/ruby_/usr/bin/fluentd_--no-supe		65,939.88	104,904.00	258
13205-/usr/bin/docker-containerd-current_-l_un		44,780.11	48,148.00	258
41350-/bin/node_exporter_--web.listen-address=		22,614.33	23,384.00	258
17027-/usr/bin/python2_-Es_/usr/sbin/tuned_-l_		17,388.00	17,388.00	258
4262-/usr/libexec/docker/rhel-push-plugin		17,216.40	18,464.00	258
15134-/usr/sbin/NetworkManager_--no-daemon		15,435.77	16,224.00	258

Conformance tests and Scale up

- Runs kubernetes e2e tests.
- Scales up the cluster if the conformance is green.
- Ansible forks is our friend during scale up.

Performance and Scale tests

- Control plane, kubelet density focused, Networking, HAProxy and storage tests.

- Tests to validate and push the cluster limits.

REPEAT

Onboarding other teams

- Eliminates the need for huge infrastructure.
- No need to install and maintain a cluster.
- Reuse the tooling and automation.
- Enables them to test their workloads at scale.

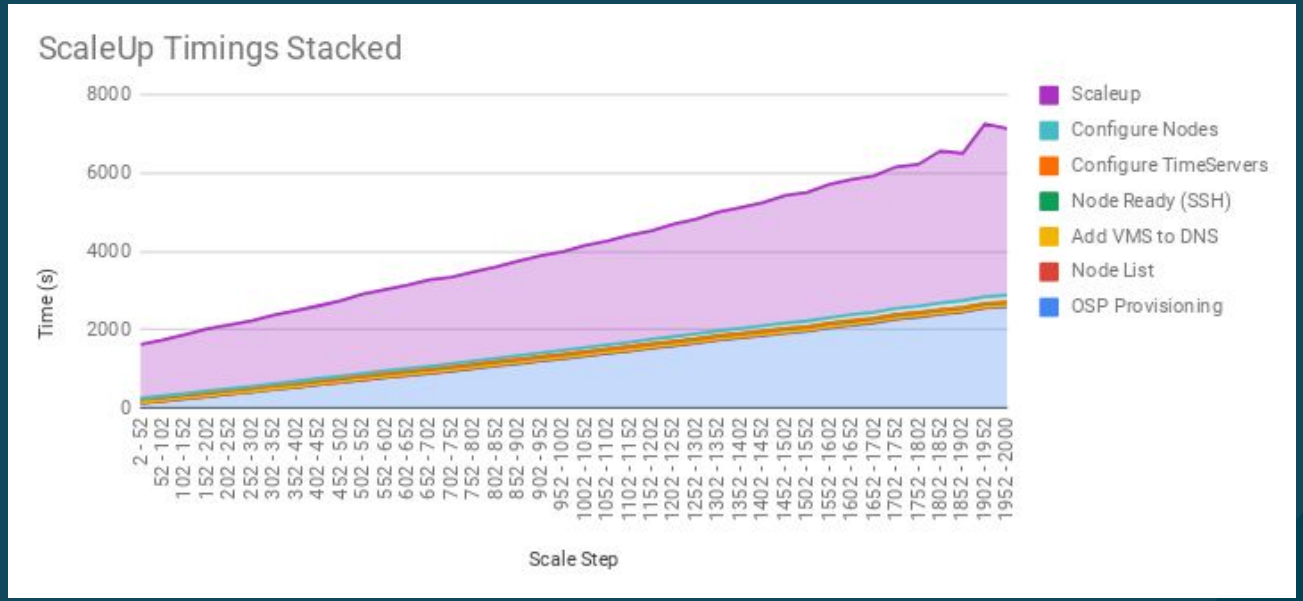
Large scale cluster results and challenges

Performance and Scale tests

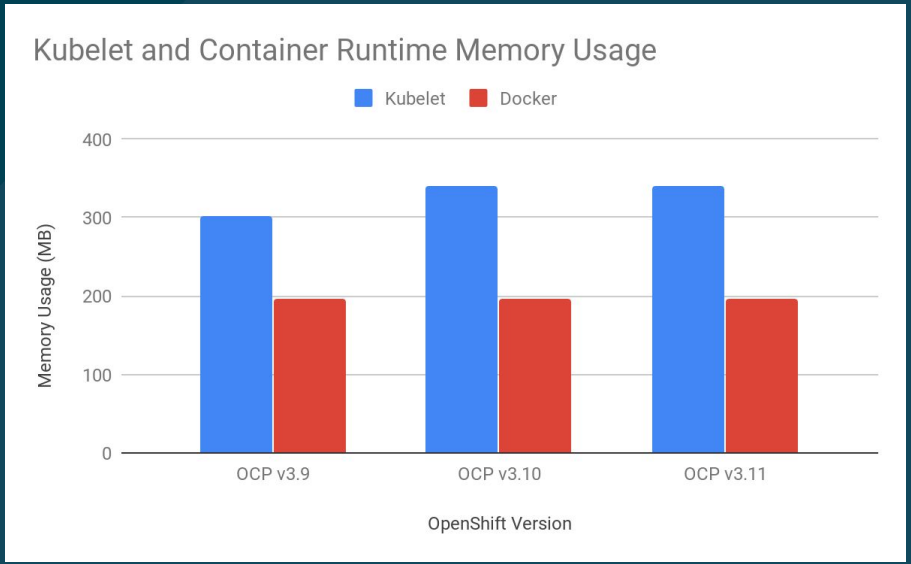
- Scale Up
- Node Vertical
- Master Vertical
- HTTP/Router
- Networking
- Storage
- Logging
- Cluster Limits

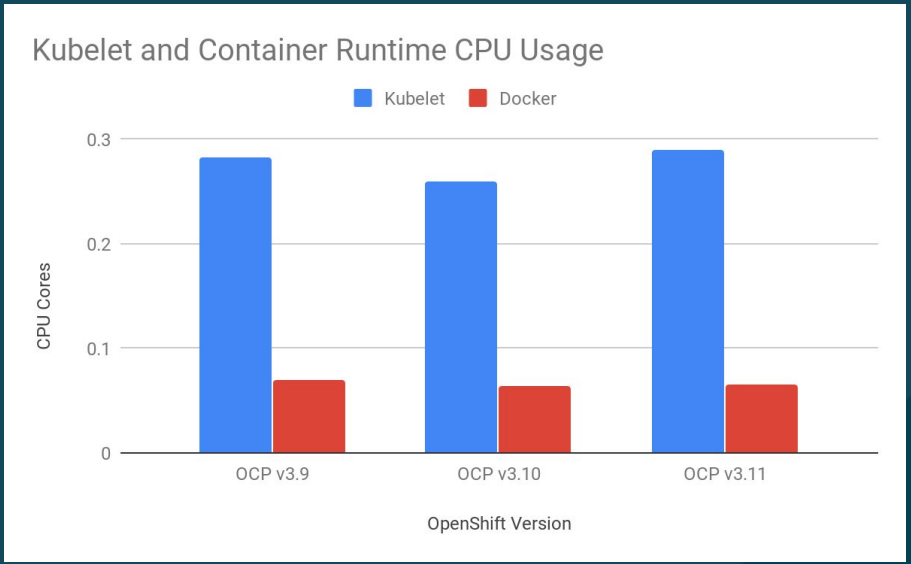
Scale Up

- Measure how long process takes to go from core cluster to complete 250/2000 node cluster



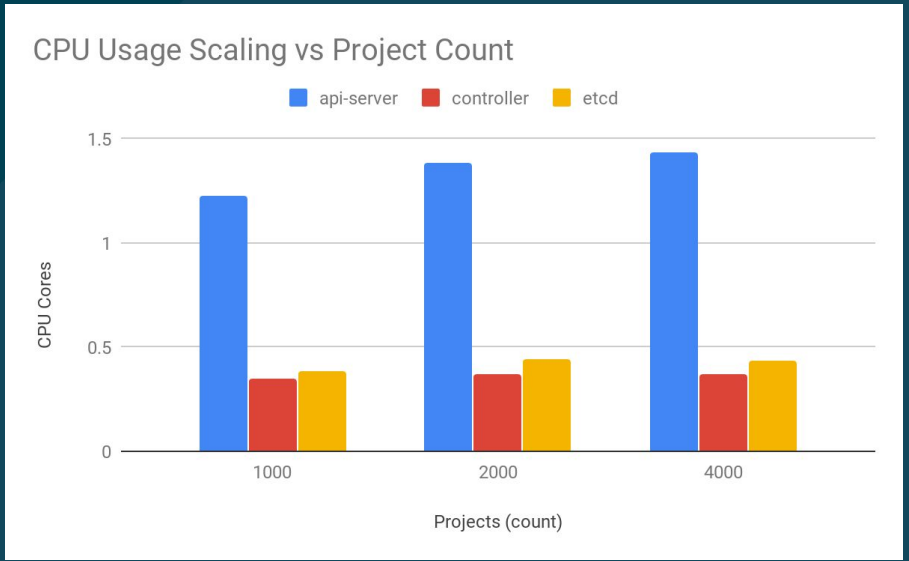
Node Vertical

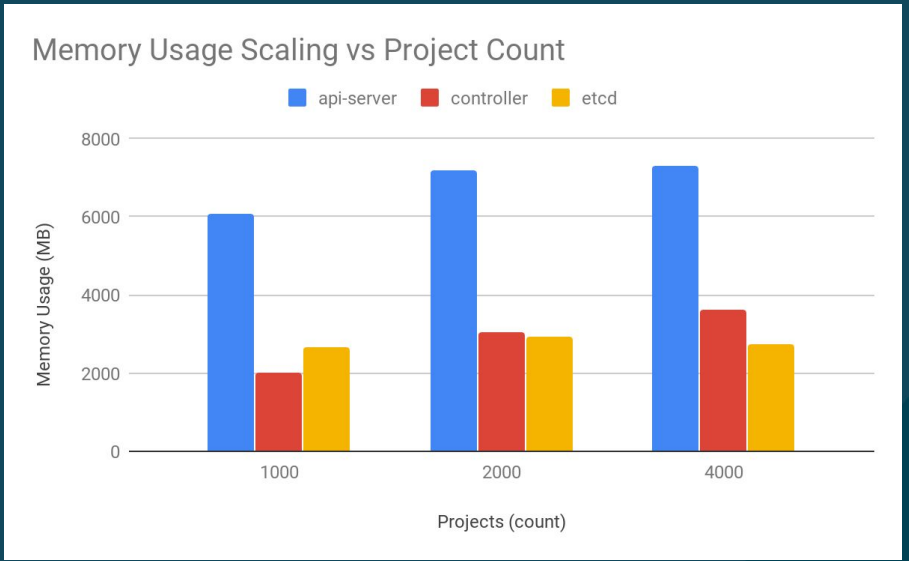




redhat.

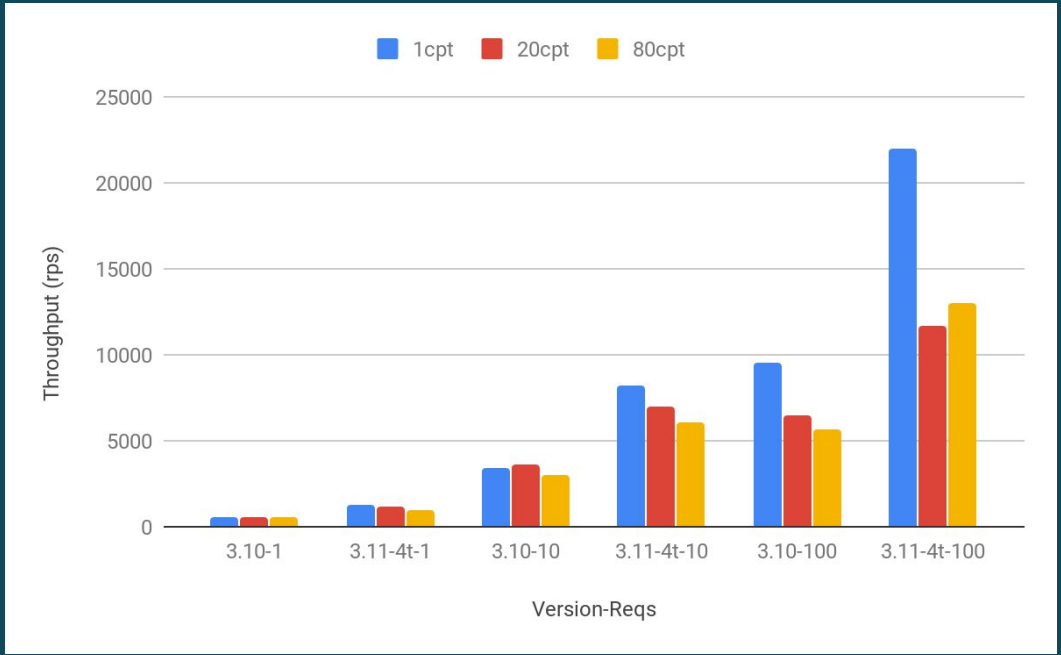
Master Vertical





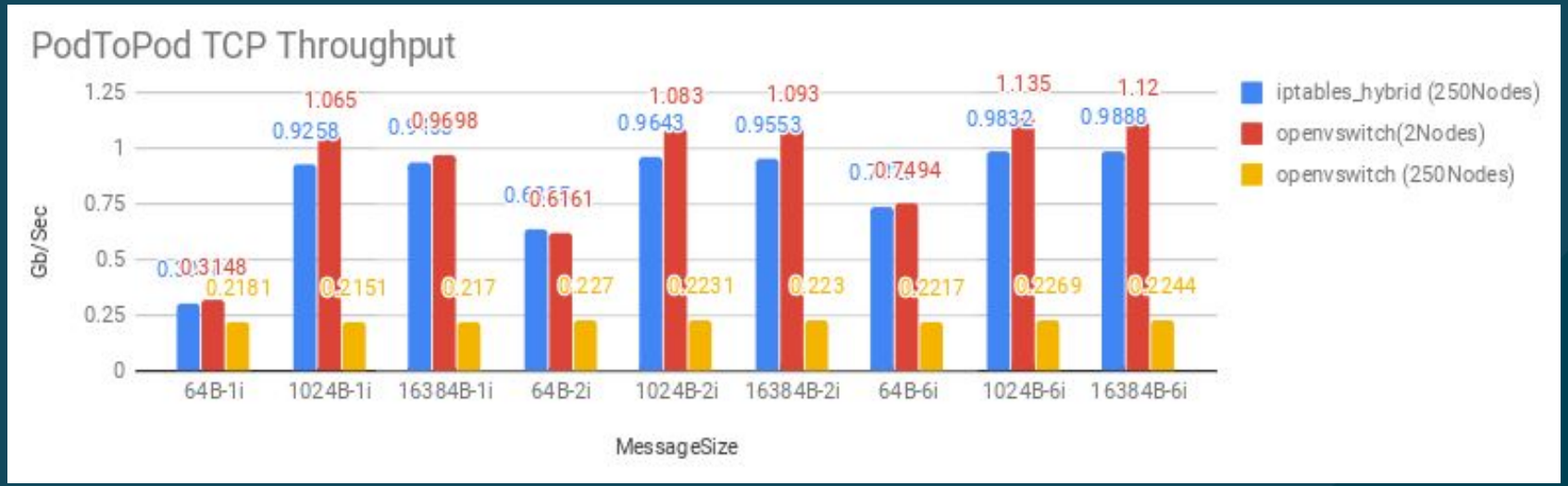
HTTP/Router

- HAProxy 1.8 in OCP 3.11 shows substantial gains due to ROUTER_THREADS



Networking

- Double encapsulated SDN: OpenStack Kuryr
- OpenStack Network configuration can highly affect performance
  - Example: Neutron Firewall driver in OpenStack



Storage

- Pgbench with Postgresql / MongoDB
  - Backed by CNS/RHOCS

- Gluster-block improvements over 3.10

Prometheus

- Capacity Planning
  - Research and develop control-plane resource usage docs.
  - Prometheus needs more than 1TB as storage space to accomodate large scale metrics for 15 days retention.

- Synthetic large scale workloads

Config Mirror

- Replicates other Clusters.
- Reproduce customer environments in support situations.

Cluster Limits

Limit Type	3.7 Limit	3.9 Limit	3.10 Limit	3.11 Limit
Number of nodes [1]	2,000	2,000	2,000	2,000
Number of pods [2]	120,000	120,000	150,000	150,000
Number of pods per node	250	250	250	250
Number of pods per core	10 is the default value. The maximum supported value is the number of pods per node.	10 is the default value. The maximum supported value is the number of pods per node.	There is no default value. The maximum supported value is the number of pods per node.	There is no default value. The maximum supported value is the number of pods per node.
Number of namespaces	10,000	10,000	10,000	10,000
Number of builds: Pipeline Strategy	N/A	10,000 (Default pod RAM 512Mi)	10,000 (Default pod RAM 512Mi)	10,000 (Default pod RAM 512Mi)
Number of pods per namespace [3]	3,000	3,000	3,000	3,000
Number of services [4]	10,000	10,000	10,000	10,000
Number of services per namespace	N/A	N/A	5,000	5,000
Number of back-ends per service	5,000	5,000	5,000	5,000
Number of deployments per namespace [3]	2,000	2,000	2,000	2,000

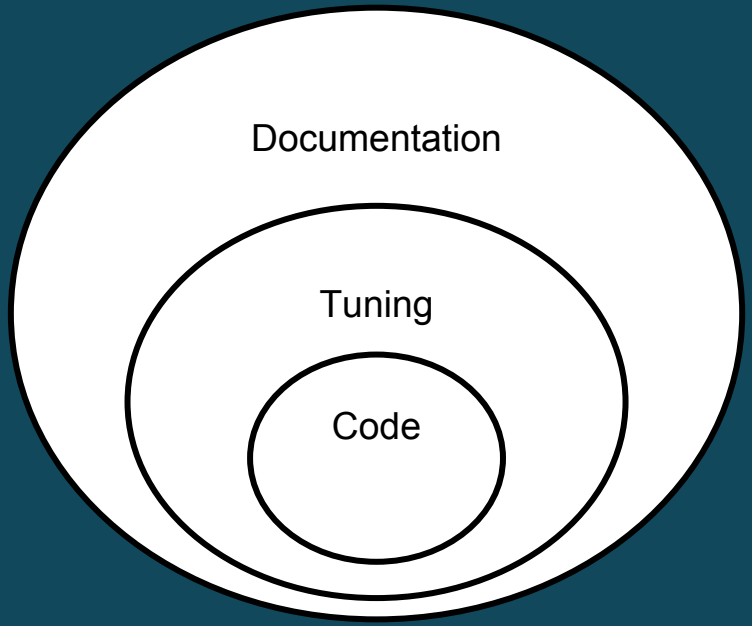
Higher Limits

- Kubernetes cluster limits were lowered.
- Designed tests to validate and push the cluster towards higher limits.
- Discussed with sig-scalability about the results.

Is it possible to run 500 pods per node?

Tuning

- Layered approach to addressing performance issues.



Tuned Daemon

- Automatic profile based host tuning

Tune ApiServer QPS and Burst rates

- Default limits might be low for large and/or dense clusters.
- Double or quadruple the rates depending on the available resources.

Scaling and Performance Guide

https://docs.openshift.com/container-platform/3.11/scaling_performance/index.html

All elements of cluster optimization and tuning:

- Installation (forks, pipelining)
- API Server Overrides (burst & QPS)
- Kubelet Config (pods per node/core)
- Network & Routing Optimization
- Workload tuning features (CPU manager, Huge pages)

Code

- Tooling and Scale Tests - https://github.com/openshift/svt
- Cluster Loader - https://github.com/openshift/origin
- Pbench - https://github.com/distributed-system-analysis/pbench
- Pipeline - https://github.com/openshift/aos-cd-jobs

What next?

- Test operators at scale.
- Onboard more teams to take advantage of the infrastructure and tooling.
- Onboard new tests for the features added upstream.
- Continue to push towards higher cluster limits.
- Support pipeline in various public clouds.

Thank you!

- Sebastian Jug - sejug@redhat.com
- Naga Ravi Chaitanya Elluri - nelluri@redhat.com


KubeCon
CloudNativeCon
North America 2018