

NET2: a first example of OpenShift/OKD for Tier 2 provisioning and cluster management in US ATLAS

Eduardo Bach (UMass)

Will Leight (UMass), Rafael Coelho Lopes de Sá (UMass), Verena Martínez (UMass)
Fernando Barreiro Megino (UTA)

HEPiX Autumn 2023, University of Victoria, Canada



The Northeast Tier 2 (NET2) is a new Tier 2 cluster at the Massachusetts Green High Performance Computing Center (MGHPCC), a zero-carbon data center near UMass Amherst

- NET2 is built as a native Kubernetes site using OKD/OpenShift
- We present the first results of NET2 in production
- We discuss the advantages of using OKD/OpenShift to deploy a Tier 2 site
- We describe the essential steps for achieving production readiness and the next steps.



NET2 is a US-ATLAS site with a pure OKD cluster

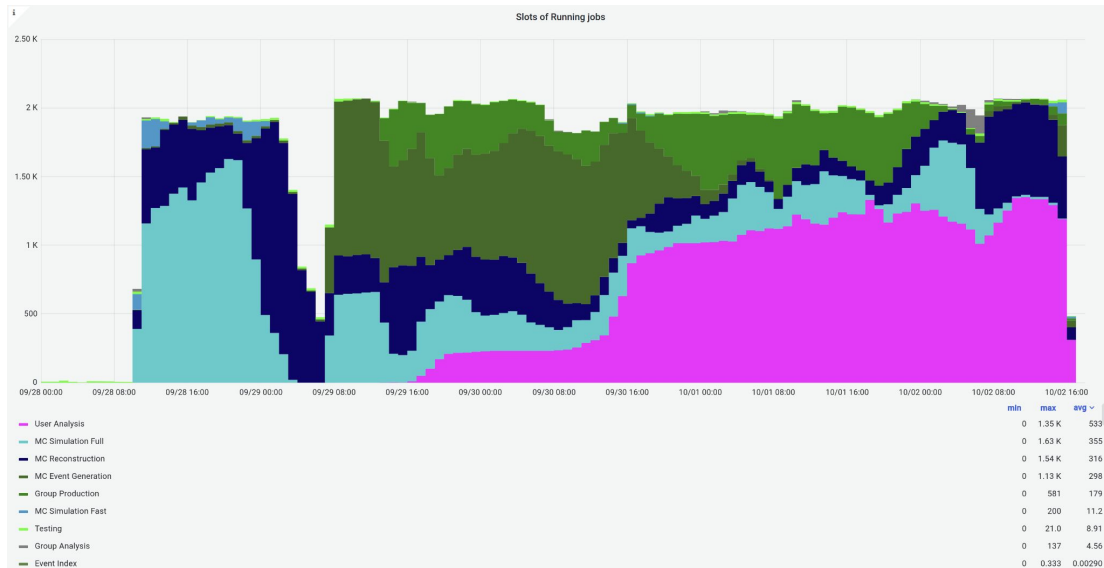
Not a traditional system

- No CE, Condor, PSB, or Slurm
- **OKD: the community distribution of Kubernetes that serves as the upstream project for Red Hat OpenShift.**
- Direct submissions to Kubernetes API;

Now in production

- Request for new queue: Sep/20
- First HammerCloud jobs: Sep/27
- Different types of jobs: Sep/28
- Began production: Sep/29

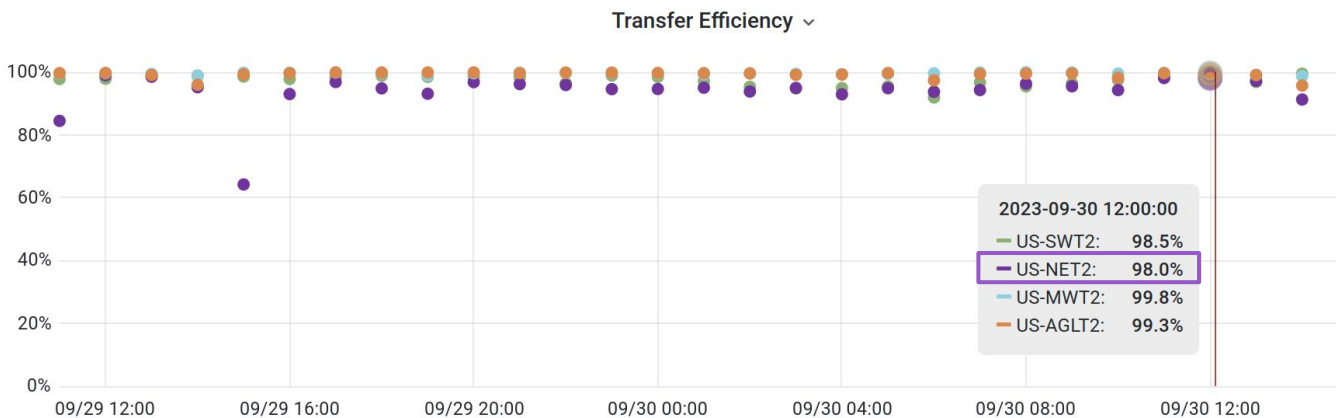
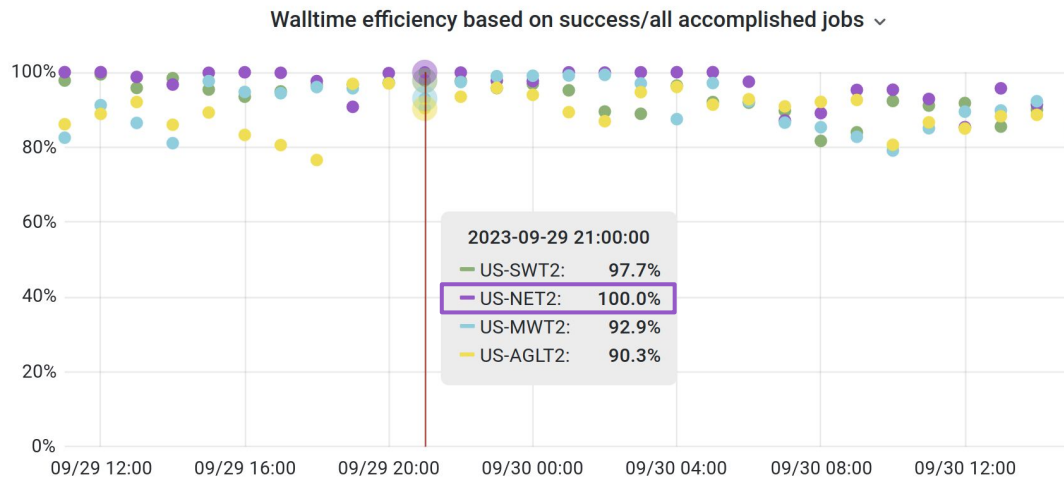
From opening the queue to production in less than 2 weeks.



NET2 operates completely as an OKD cluster. No subclusters!

Preliminary results

The new Tier 2 has started operation with good efficiency, comparable to other Tier 2 in the US



Preliminary results

Queue name	Queue type	Region	Status	Job type	Resource type	N running slots	N jobs total	% failure	assigned	activated	running	transferring	merging	finished	failed
AGLT2_MERGE 🚩 ⚙️ 🚩 🚩	production	US	online	prod	all	20	323	0.7	0	4	13	3	0	298	2
BNL_PROD_INTEL 🚩 ⚙️ 🚩 🚩	unified	US	online	analy	all	16	657	0.8	0	0	2	0	0	650	5
MWT2_VHMEM_UCORE 🚩 ⚙️ 🚩 🚩	production	US	online	prod	all	642	847	1.1	0	4	82	8	0	743	8
NET2_Amherst 🚩 ⚙️ 🚩 🚩	unified	US	online	prod	all	720	2059	1.4	0	973	265	382	0	429	6
NET2_Amherst 🚩 ⚙️ 🚩 🚩	unified	US	online	analy	all	1,334	3472	1.5	0	1585	648	0	0	859	13

WallClock Consumption of Successful and Failed Jobs - Pie Chart



Panda is reporting comparable results.

All the problems after launch are unrelated to the use of Kubernetes/OKD.

Container-native batch computing

- NET2 employs Kubernetes for streamlined batch computing.

Harvester & Kubernetes API

- Harvester submits ATLAS grid jobs directly to Kubernetes API.
- Jobs run as containerized pods.

Kubernetes job resource

- Utilizes the "job" resource type for pod control.
- Ensures successful pod termination, offloading tracking responsibility from Harvester.

Resource allocation

- Harvester sets specs like memory and CPU for each job.
- Kubernetes handles scheduling based on resource availability.

Credential management

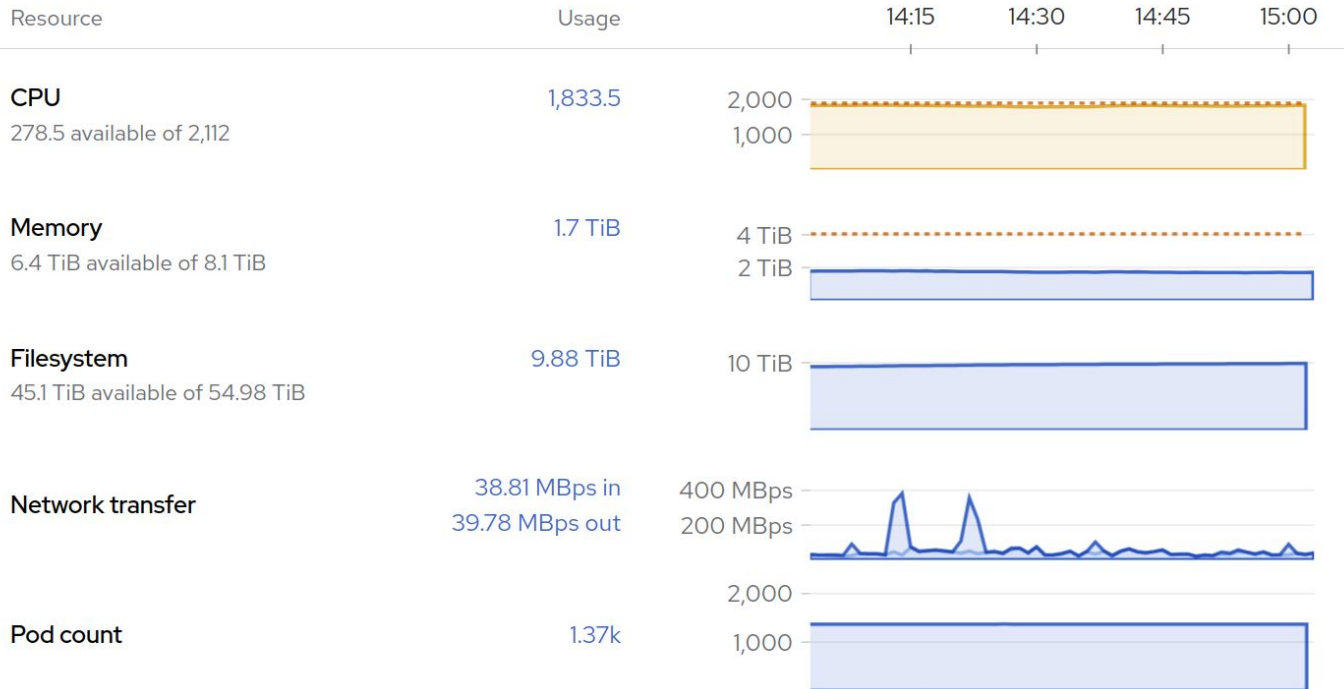
- Regular updates of X509 proxies as Kubernetes secrets.
- Enables pilot jobs to authenticate with grid services.

NET2 in numbers (so far)



Cluster utilization

Filter by Node type ▾ 1 hour ▾



This week NET2 is currently providing approximately 2000 cores.

After S&C week, new racks are being added to the cluster, which will quickly ramp up to a size comparable to other US Tier 2 sites.

OKD: Tool to manage it all, including bare metal



```
apiVersion: v1
baseDomain: net2.mghpcc.org
metadata:
  name: compute
networking:
  machineNetwork:
    - cidr: 69.16.44.0/24
      networkType: OVNKubernetes
compute:
- name: worker
  replicas: 30
controlPlane:
  name: master
  replicas: 3
  platform:
    baremetal: {}
platform:
  baremetal:
    bootstrapOSImage: http://69.16.44.252:8080/fedora-coreos-37.20221127.3.0-gemu.x
    bootstrapExternalStaticIP: 69.16.44.32
    bootstrapExternalStaticGateway: 69.16.44.1
    bootstrapExternalStaticDNS: 8.8.8.8
    provisioningDHCPRange: 172.20.174.33,172.20.174.254
    provisioningNetworkCIDR: 172.20.174.0/23
    clusterProvisioningIP: 172.20.175.252
    bootstrapProvisioningIP: 172.20.174.2
    externalBridge: baremetal
    provisioningBridge: provisioning
    apiVIPs:
      - 69.16.44.30
    ingressVIPs:
      - 69.16.44.31
    provisioningNetwork: "Managed"
    hosts:
      - name: node034
        role: worker
        bmc:
          address: redfish://172.20.173.34/redfish/v1/Systems/System.Embedded.1
          disableCertificateVerification: True
        bootMACAddress: B0:7B:25:D4:E0:AA
        rootDeviceHints:
          deviceName: "/dev/sda"
```

The cluster is defined using a single yaml file

Single node OKD cluster will bootstrap the control planes (using Ironic/Terraform)

node069	Externally provisioned	node069	control-plane, master	redfish://172.20.173.69/redfish/v1/Systems/System.Embedded.1
node070	Externally provisioned	node070	control-plane, master	redfish://172.20.173.70/redfish/v1/Systems/System.Embedded.1
node071	Externally provisioned	node071	control-plane, master	redfish://172.20.173.71/redfish/v1/Systems/System.Embedded.1

BMH node044

✔ Provisioned

N node044

worker

redfish://172.20.173.44/redfish/v1/Systems/System.Embedded.1

BMH node045

✔ Available

-

-

redfish://172.20.173.45/redfish/v1/Systems/System.Embedded.1

BMH node046

! Inspection error

-

-

redfish://172.20.173.46/redfish/v1/Systems/System.Embedded.1

Bare Metal Hosts > Bare Metal Host details

BMH node044

Actions ▾

Overview

Details

YAML

Network Interfaces

Disks

Events

Details

[View all](#)Host name
node044Role
workerNode
N node044

Inventory

[199 Pods](#)

1 Disk

7 NICs

64 CPUs

Status

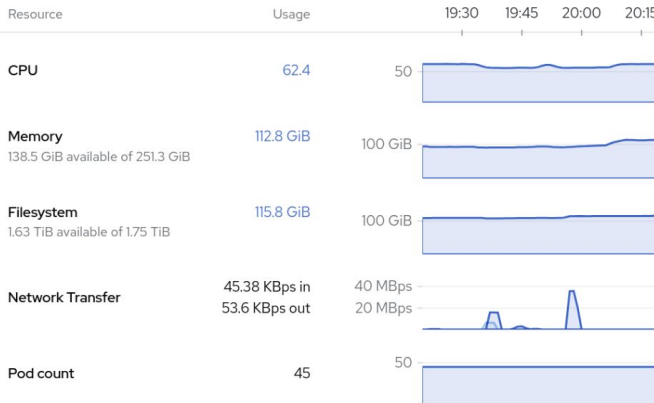
✔ Provisioned

✔ Hardware

⚡ Powered on

Utilization

1 hour ▾



Activity

[View events](#)

Ongoing

There are no ongoing activities.

Recent events

[Pause](#)

There are no recent events.

Advantages of OKD for a Tier 2 cluster



Single tool to manage all the aspects of a Tier 2 cluster

- Provision directly from bare metal (using the embedded Ironic features)
- Used to deploy resources (Squid, CVMFS) needed for grid computing
- Provisioning/recovering from disasters fully automated: from erasing the nodes disks to starting to receiving jobs in less than 3 hours!

Potential to save a considerable amount of admin labor

It can unify computing resources from different systems under the same administration: often important in large data centers (like the MGHPCC, where the NET2 cluster is located).

Preparing for production: squid proxy

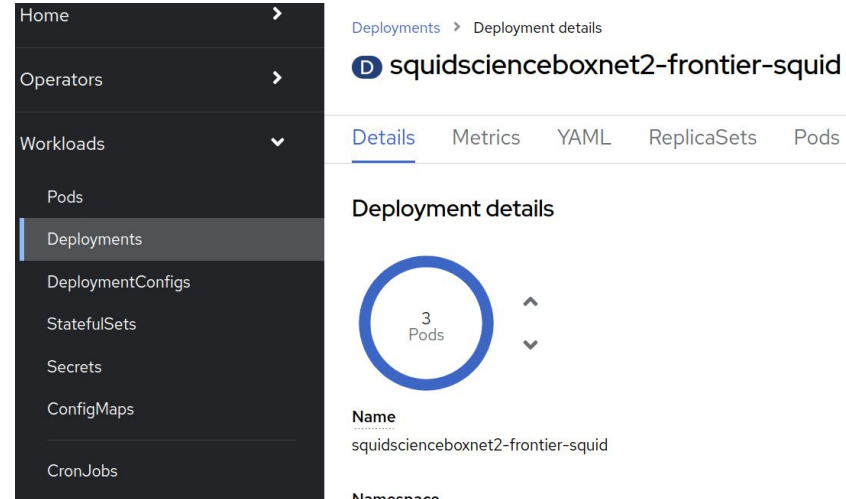
Helm chart by CERN's ScienceBox:

Ryan's (UVic) contributions to the image made it nearly working out of the box

Setup is (easily) highly scalable

We plan to suggest modifications to facilitate adoption across other OKD clusters

WIP: exploring the new CVMFS proxy sharding feature (different namespace)



The screenshot shows the OpenShift console interface. On the left is a dark sidebar menu with options: Home, Operators, Workloads (expanded to show Pods, Deployments, DeploymentConfigs, StatefulSets, Secrets, ConfigMaps, and CronJobs), and Deployments. The main content area shows the path 'Deployments > Deployment details' for the deployment 'squidscienceboxnet2-frontier-squid'. Below this are tabs for 'Details', 'Metrics', 'YAML', 'ReplicaSets', and 'Pods'. The 'Details' tab is active, displaying a circular gauge with '3 Pods' and a 'Deployment details' section with fields for 'Name' (squidscienceboxnet2-frontier-squid) and 'Namespace'.

```
helm repo add sciencebox
https://registry.cern.ch/chartrepo/sciencebox
```


Tasks to be accomplished in the near future

Setting up (K)APEL (<https://github.com/rptaylor/kapel>) for accounting.

Write documentation so that other sites can use the same system

Ramp up the system with more servers

Acknowledgements



Special thanks for the help received from **ATLAS colleagues**:

Fernando Harald Barreiro Megino, Ryan Taylor, Fabio Luchetti, Rodney Walker, Doug Benjamin, Hironori Ito, Petr Vokac, Frederick Luehring, Ofer Rind, Lincoln Bryant, Rob Gardner

And also from the **RedHat research colleagues**:

Heidi Dempsey, Lars Kellogg-Stedman, Christopher Tate, Michael Zink

Thank you!



Questions?