LA-UR-22-23449

Approved for public release; distribution is unlimited.

Title:	Deploying Cray EX Systems with CSM at LANL
Author(s):	Stradling, Alden Reid Johnson, Steven Lee Van Heule, Graham Knox
Intended for:	Cray User Group, 2022-05-02/2022-05-06 (Monterey, California, United States)
Issued:	2022-05-03 (Rev.1) (Draft)

Los Alamos NATIONAL LABORATORY

Deploying Cray EX Systems with CSM at LANL

Steve Johnson, Alden Stradling, Graham Van Heule

May 4, 2022

LA-UR-22-20724

Talk Structure

- We’re not going to “wall-of-text” in this presentation. That’s the paper’s job.
  - Slides are starting points for discussion.
  - Please feel free to interrupt and ask questions!
  - Any of these topics could lead to a long discussion

Deploying Cray EX Systems with CSM at LANL

01

Early Days

- Shasta 0.8.6 through 1.2. The dark times.

02

The Road to Prod

- Network and Slurm and authentication and what to do without filesystems

03

Early User Period

- When the cluster needs to run but security needs work. 1.2 to 1.4

04

Quality of Life

- Image management becomes manageable, cluster health gets healthy, and more

05

Sleeping Well

- Training, config management beyond internal Gitea, and making the cluster resilient

06

Status and Planning Ahead

- Dreams of diminishing downtime, and what makes our admins happy

01 The Early Days

1. Deployment of 0.8
    a. So HARD
    b. So FRAGILE
    c. So UNSTABLE
2. Kubernetes
    a. Learning curve
    b. Uncertainty in what was Cray and what was normal k8s
    c. Strangers in a strange land
3. AuthN/AuthZ
    a. Adapting to LANL's… unique LDAP structure

02 The Road to Prod

1. Chicoma (production…) and Guaje (TDS)
    a. Shasta 1.2:
        i. PersistentVolume issues
        ii. Completely unmanageable image build process (recurring!)
        iii. Installation process was very fraught
        iv. etcd!!!
        v. Concerns about viability
    b. Shasta 1.3:
        i. Major concerns addressed
        ii. Installation was made more ***reliable*** and ***fast***
        iii. Major help from CSM team

Major concerns remained, but at least we were confident that we had good support and that major improvements were inbound.

03 Early User Period

1. Challenges (not all laid to HPE/Cray's account!)
    a. Vendor-accessible enclave, and therefore…
    b. Not able to use production filers (security concerns)
    c. Auth distribution to nodes
    d. CPU throttling
    e. Some MPI concerns
2. Addressed with
    a. LDAP user lists distributed through Keycloak and S3 to nodes (since deprecated)
    b. ZFS carve-out on Ceph filer (and zpool backups!)
    c. No good answer on CPU issue yet
    d. Local builds of MPI rather than using PE versions
3. 6 weeks to go from 1.3 to 1.4 (!)
    a. Improvements were important, but made us cautious and risk-averse

04 Quality of Life

1. Image Management
    a. Overwhelmingly large and confusing command output
    b. Lots of commands to run to perform simple tasks
    c. No defaults in cluster management
    d. Fixed for now with scripting, awaiting SAT module
2. Node/Cluster Health Monitoring Issues Solved
    a. No node health validation and repair built in
        i. Need to validate lid is valid on nodes
        ii. Need to validate that cfs completed successfully
        iii. Need to validate node has correct hsn ip per dns
        iv. Check Fabric Health
    b. No Mechanism to report NCN health (UAN, Lnet, management)
    c. Provides cluster level health built in via Prometheus
3. AuthN/Z – moved to nssdb basis from standard LANL sourcing

05 Sleeping Well

1. Storage (Ceph) resiliency
2. Training
    a. The new model of Cray's software is nothing like its predecessor
    b. No general in house knowledge of tools like Kubernetes
    c. Initial training completed for all of our affected team and beyond (40 people)
3. Config Management
    a. Configuration source for all clusters in a centralized location
    b. Configuration generic to all clusters
        i. Submodules? AdditionalInventory?
    c. Reconstitute whole cluster from repos
4. WLM/Slurm
    a. Centralized DB, configless slurmd, config changes through git
    b. External slurmctld as well? Networking?

06 Status and Planning Ahead

1. Status: Operating and stable
    a. Nvidia software complications
    b. Slingshot 1.5->1.6->1.7 ups and downs
    c. Cooling loop issues present but minimal
    d. Vulnerable to CDU vagaries!
2. Prospects:
    a. Looking into Prometheus as a replacement to nhc for cluster level health.
    b. Kubernetes security training
    c. Building gitlab runner pipelines for image build and deployment
    d. Reliable reboot and rebuild
    e. Upgrades to 1.5 and beyond with growing confidence (please!)
3. Pressures:
    a. 5 new Shasta systems showing up by mid-2023!

Conclusions

- Shasta has been a challenge, for both intrinsic (k8s) and maturity reasons
- The CSM team and others have done remarkable things for us in crisis
- Site adaptation is still a challenge, may be irreducible
- Evolving CSM – better install, better tested, more resilient
- Cluster resiliency and downtime in general – far better!
  - Ceph story, etcd encryption story

Paying down technical debt is never painless, and never a linear process.