CRAY EX SHASTA V1.4
SYSTEM MANAGEMENT OVERVIEW

Harold Longley, CSM User Experience Solutions Architect
CUG 2021, May 3-5, 2021
AGENDA

- Cray System Management (CSM) Architecture
- New in Shasta v1.4
CSM ARCHITECTURE
**CSM OVERALL ARCHITECTURE**

- Kubernetes as Platform-Building-Platform
- Kubernetes, Istio, and Operators for infrastructure
- Layered microservices for managing HPC system
- HPC-enablement only in the upper layers
- Northbound APIs for Users and Admins
- Southbound APIs for interacting with Compute hardware

All User/Admin interactions protected by TLS 1.3 and OIDC authentication

---

Interactive Services

Infrastructure Services

Hardware Management

RBAC

Security Policy

Audit

Horizontal Scaling

Secure Secrets

Resiliency

Automatic Backup

Kubernetes and Object Storage
### CSM FEATURE LAYERS

<table>
<thead>
<tr>
<th>On-Demand Interactive Access Containers</th>
<th>LDAP integration for User/Group Mapping</th>
<th>Cooperative Multitenancy / Multischeduling</th>
<th>Monitoring Visualization/Event Log Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flexible Mapping of Boot Configuration to Node Groups</td>
<td>On-system Compute Image Customization</td>
<td>Scalable Configuration Management</td>
<td>Telemetry and Log Aggregation</td>
</tr>
<tr>
<td>Automatic DNS and DHCP with Geolocation</td>
<td>Fast System Boot with PXE and DVS</td>
<td>Compute Image Registry</td>
<td>Slingshot Fabric Management</td>
</tr>
<tr>
<td>Automatic Redfish Discovery</td>
<td>Blade and Cabinet Control</td>
<td>Logical Node Grouping</td>
<td>Power Control</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPIFFE for Attestation</td>
</tr>
</tbody>
</table>

#### Interactive Services
- Interactive Services
- Infrastructure Services
- Hardware Platform Management

#### CUG 2021
**MICROSERVICE SECURITY LAYERS**

- Pod to Pod Traffic is secured by Istio with mTLS and Kubernetes Policy
- Ingress and Egress traffic is regulated by OPA
- Istio provides gateway services to expose collections of services
- MetalLB allocates Virtual IP addresses that pass traffic to Istio Gateways

- Keycloak handles authentication and issues refreshable bearer tokens, required for API Access
- Keycloak federates with upstream LDAP or Kerberos for user directories
Both images and recipes are delivered as part of the installation media.

BSS: Boot Script Service
BOS: Boot Orchestration Service
BOA: Boot Orchestration Agent
CFS: Configuration Management Service
IMS: Image Management Service
USER ACCESS OPTIONS

- Power Users
  - Compile and Run

- Standard Users
  - Run and Monitor

User Access Node

Slingshot Interconnect

User Access Instances

Management NCNs

Filesystems
USER ACCESS SERVICE AND BROKER

- On-Demand containerized SSH environment “serverless”
- SSH is the only User-Facing API
- Templated UAI Pods launched and destroyed as-needed
- User state persisted only in cross-mounted filesystems (like /home)
- Internal SSH relies only on single-use SSH keys
- Broker consumes a single IP regardless of how many users
- Multiple brokers can be used to handle different user types and user groups
PRODUCT STREAMS IN SHASTA V1.4

- New installation process with CSM and other product streams each having their own install.sh
  - CSM – Cray System Management
  - SDU – System Dump Utility
  - SAT – System Admin Toolkit
  - Slingshot – High speed network fabric management
  - SMA – System Monitoring Application including monitoring, telemetry and log aggregation
  - OS – rpms from SUSE
  - COS – Cray Operating System for compute nodes
  - UAN – User Access Nodes
  - CPE – Cray Programming Environment
  - WLM – Slurm or PBS Pro workload management
  - Analytics – AI and Analytics software
- Enables delivery of product stream updates on varying release schedules
INSTALLATION OVERVIEW
INSTALLATION IN SHASTA V1.4

- Moved installation bootstrap from ncn-w001 to ncn-m001
  - After first time installation, ncn-m001 is no longer special node (like BIS node was)
- Cray site initialization (CSI) toolkit
  - Gather data from site survey to feed into the CSM installation process
    - System name, system size, site network information for CAN, site DNS, site NTP, bootstrap node network information
- New cabling and management network switch configuration guide
- Image based NCN installs of SLE 15 SP2
  - Management nodes boot over faster PCIe NICs instead of onboard NIC
- CSM installation has pre/post flight checks at various points during installation
  - CSM validation suite can be used for system management health check during normal operation of system
- Artifact storage in Nexus
  - RPM repositories, container images, Helm repositories, firmware content
- Software updates for CSM
  - Developed process to patch a release
  - New process to deliver rpms for late-breaking workarounds and minor documentation updates
- UAN uses separate image recipe from COS (for compute nodes)
USING GIT FOR MANAGING CFS CONFIGURATION

- Stores Ansible to apply to nodes at lifecycle events
- All Ansible in git repositories with branches to allow site customization
- Ordered configuration management across multiple repositories
- CFS sessions as part of pre-boot Image Customization as well as post-boot Node Personalization
CFS FOR POST-BOOT CUSTOMIZATION

Managing All Nodes with System Images

$ git clone

$ ansible-playbook...
CFS FOR IMAGE CUSTOMIZATION

Managing Compute and Application Nodes with System Images

$ git clone

$ ansible-playbook...

IMS
BOS/BOA
CFS
CRUS
**Infrastructure**

- Scalable DHCP with Kea
- Scalable DNS with CoreDNS
- Monitoring/Alerting additions
  - Postgres cluster monitoring and alerting dashboards
  - Etcd cluster monitoring and altering dashboards
  - Alerts for failed or degraded NCN disks
- Procedures for NCN reboot or rebuild

**Security**

- Migration to trusted base OS for container images
- SPIRE/SPIFFE token service
- Certificate Management Tooling improved
- Vault moved from etcd to raft for key/value store
- OPA (Open Policy Agent) policies replace PSPs (Pod Security Policies)
- RSA Multi-factor Authentication (in v1.3.1)
FAS (Firmware Action Service) can update firmware for Management nodes, Compute nodes, Application nodes, Slingshot switches, and Mountain cabinet components

- Procedure for NIC firmware updates, but not orchestrated by FAS
- Locking API enables locking of NCNs/CNs before using FAS or power up/down (CAPMC)
- Boot reliability and scaling improvements
  - BOS (boot orchestration), CFS (configuration), CAPMC (power control), HBTD (node heartbeats), HMNFD (fanout), SPIRE (token service)
  - Tuned critical services for Kubernetes resource requests and limits
  - Moved several services from singleton pods to multiple instances
- All node console logs gathered by cray-conman to SMA logging infrastructure
  - Cray-conman can be used for interactive console access for all node types
- UAI SSH Broker
SDU, SMA, SAT CHANGES IN SHASTA V1.4

• SDU
  • Runs in container under podman on Kubernetes master nodes

• SMA
  • ElastaAlert – log alerting feature
  • Conversion of LDMS to V4
  • Support for external rsyslog

• SAT
  • Runs in container under podman on Kubernetes master nodes
  • sat hwinv supports more types
    – node enclosure power supplies, node accelerators (GPUs), node accelerator risers, node HSN NICs
  • Monasca alarms for Redfish Events with sma-monasca-translator
    – Sensor readings exceeding thresholds
    – Removal or addition of drives
    – Power events
  • sat swap works with Slingshot fabric controller
  • SAT logfile moved to /var/log/cray/sat/sat.log
  • Removed sat cablecheck
    – Instead use “show cables” in Slingshot Topology Tool (STT)
RELATED PRESENTATIONS AND PAPERS

- CUG 2021
  - Managing User Access with UAN and UAI
  - User and Administrative Access Options for CSM-Based Shasta Systems
- CUG 2020
  - Advanced Topics in Configuration Management
  - HPE Cray Supercomputers: System User Access; User Access Node or User Access Instance, Which is Right for Me?
- CUG 2019
  - Shasta Software Technical Workshop
  - Shasta System Management Overview
  - Reimagining Image Management in the New Shasta Environment
  - Hardware Discovery and Maintenance Workflows in Shasta Systems
THANK YOU

harold.longley@hpe.com