User and Administrative Access for CSM-Based Systems: Network Architecture Evolution and Access Control Mechanics in Shasta v1.4 and Shasta v1.5

Alex Lovell-Troy
Hewlett Packard Enterprise
Bloomington, MN USA
alt@hpe.com

Sean Lynn
Hewlett Packard Enterprise
Bloomington, MN USA
sean.lynn@hpe.com

Abstract—**This paper describes the implications for user and administrative access when adopting Cray System Management (CSM). Customers have choices for how deeply to integrate a CSM-managed system with their broader datacenter networks and how to manage access across the boundaries between system and datacenter. While the primary goal is to present options for versions of CSM delivered with Shasta v1.4 and v1.5 for HPE Cray EX Systems, this paper also provides guidance on future direction.**

Keywords—Cray System Management; access control; user access; administrative access; network architecture

I. Introduction

CSM provides a cloud-like experience for administrators and users alike. This paper outlines the considerations that sites need to make in network planning and explores common configuration options. Readers will benefit from a basic understanding of the lower levels of the OSI model, in particular, the Data Link layer (layer 2) and Network layer (layer 3). Additionally, because of the many possible network options available in CSM going forward, this paper will not cover all options. HPE HPC Service and Support personnel are available for detailed review of individual use cases and site-planning.

Traditional HPC system management architectures rely on hardware-based network isolation to separate user and administrative access. There is an implicit assumption that HPC systems can be treated as appliances, safe within protected networks. Increasingly, customers seek to integrate their HPC systems with their datacenter networks in order to address new workloads, not only for management, but also for HPC and AI workload interaction. Further, the Slingshot interconnect offers Ethernet connectivity which allows sites to connect the high-speed fabric directly to the datacenter as well. As customers choose these connections, the appliance assumption becomes invalid.

The CSM installation prepares and deploys a distributed system across a group of management nodes organized into a Kubernetes cluster which uses Ceph for utility storage. These nodes perform their function as Kubernetes Master nodes, Kubernetes Worker nodes, or utility Storage nodes with the Ceph storage.

System services on these nodes are provided as containerized microservices orchestrated by Kubernetes to be scheduled on Kubernetes Worker nodes with horizontal scaling to increase or decrease the number of instances of some services as demand for them varies, such as when booting many compute nodes or application nodes.

As network architectures evolve, the software interactions on those networks need to evolve as well. All management functions in CSM are available through REST APIs that can be accessed from both inside the system and externally. A root account on a management node is not sufficient to manage a Shasta system. Users and Administrators are credentialed by the site through existing LDAP directories (or similar) and permissions can be set at that level and enforced within CSM. Access to the APIs is secured via encryption following the same standards that ensure private communication on the broader Internet and access is controlled through an expressive policy engine similar to public cloud Identity and Access Management (IAM). Workload Managers from companies like SchedMD and Altair already have API integration modules and are ready for CSM.

Together, the network architectures and cloud-native software services act in concert to ensure Administrators can manage the system and Users can use the system through common cloud-like interactions. Users can be assured that results are not compromised by unauthorized changes. Administrators can tune the security options to fit the needs of the site. Both groups can rely on a secure audit trail to match actions to people even when authorization is temporarily delegated. In CSM, we have invested heavily in applying the cloud security and interaction model to all parts of the management system. This investment is a continued evolution from the traditional HPC systems management model towards a more mature cloud-based systems management that fits today's customer use cases with room to grow in the future. The following analysis highlights concepts necessary for working with the new paradigm, considers current feature status, and discusses some of the implementation detail that should inform sites with CSM-based systems.

II. Connecting CSM-Based Systems to the Datacenter Network

A CSM-based supercomputer is designed to operate cleanly on a datacenter network. It offers a set of connection options for users and administrators to access CSM services as well as several connection options to allow parts of the supercomputer to access the rest of the datacenter and internet at large.

A. Basic Network Architecture

For the purposes of this paper, the CSM management network architecture is predominantly a layer 3 architecture with layer 2 features used to reduce complexity. As such, the many IP subnets involved in managing a supercomputer come together in routers where they can be segregated and routed as needed. To simplify the routing tables inside the system as well as at connection points with the site networking, layer 2 VLANs are used

B. Site Connection Options

Depending on the expected usage of the compute infrastructure in a Shasta system with CSM, customers may choose several options for site backbone connections to the overall Shasta system. In keeping with the leaf/spine architecture of the physical network, Shasta system architecture favors leaf connections to the site networks with spine ports reserved for internal communication, see Fig. 1. Ideally, the spine switches should be dedicated to Layer 3 routing. Because of port density, some smaller systems utilize spine ports rather than leaf ports for exit traffic which is an acceptable exception. With larger systems, the need for overall bandwidth on the spine switches increases along with the number of ports on those spine switches. Once a system has grown beyond the capacity for a single pair of spine switches, the architecture required to maintain balanced site connectivity through the spine switches becomes overly complex and leaf switch connectivity is clearly preferred. It is important to note that the edge routers that provide layer 3 routing for Slingshot in a Shasta system can be thought of as management leaf switches in the leaf/spine architecture and provide a reasonable attachment point for site networking. Whether the site chooses to allocate ports on the edge router(s) for site connections or to allocate ports on a management leaf switch, the overall architecture remains the same. The installation defaults in Shasta v1.4 assume that a management leaf is acting as an "exit-leaf", but altering this default is relatively trivial after installation is complete.

C. Customer Access Networking

CSM is designed to integrate the HPC system with the datacenter. The concept of the Customer Access Network (CAN) is to provide customers with a single Layer 3 point of integration between the datacenter and HPC system without tying it to any particular management node. Through the CAN, the site delegates a set of IP addresses to be managed by CSM. And, the site can choose what networking equipment is involved in that delegation. Default Shasta v1.4 configuration assumes that one of the Aggregation switches is the physical point of contact. The CAN IP addresses, which are managed by CSM are allocated to services and DNS resolution supplied by CSM maps hostnames to IP address. Sites do not need to maintain an independent map of those hostnames and IP address because one of the services is the DNS service. CSM manages the DNS of all services and nodes on the CAN. DNS server on a predictable IP address through which all CAN services can be looked up. For the integration, it does not actually matter which IP address is associated with the DNS address "shasta.<system-name>.<site-domain>". What matters is that every administrator should be able to resolve that name from their administrative workstation. Connections flow the other direction as well to keep the integration clean.

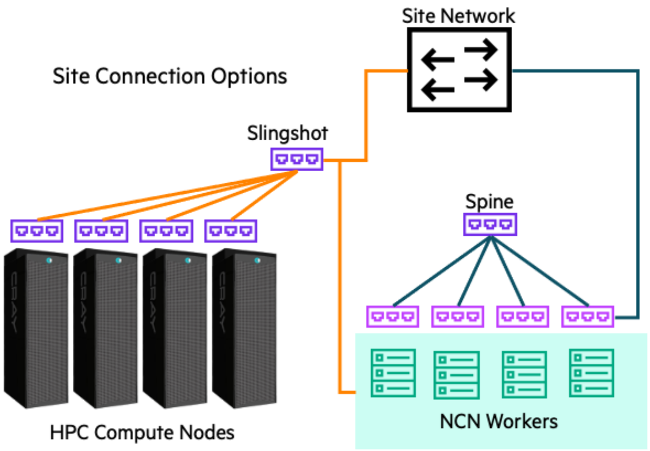

Fig. 1. Site Connection Options

We recommend that sites provide upstream DNS services so that all nodes in the Shasta can resolve other <system-name>.<site-domain> addresses and potentially Internet addresses like google.com. We also recommend upstream NTP services to ensure consistent time across the whole datacenter. This is particularly important because all TLS certificates have time-based validity. If the clocks are off by a significant margin, TLS will fail. Finally, while CSM does provide an authentication service in Keycloak and it is possible to manage users there, we recommend LDAP federation for user and group information. That should simplify the jobs of the admins. When they disable a user in their central LDAP, that gets reflected within the Shasta system without further action. For that reason, we recommend that sites provide an upstream LDAP server for federation.

Site Services accessed by CSM

- NTP for time synchronization
- DNS for name resolution beyond the system itself
- LDAP for authentication and authorization

CSM Services available to the site

- shasta.<system-name>.<site-domain> – the main API gateway
- auth.<system-name>.<site-domain> – the identity provider for all OpenID Connect authentication
- vcs.<system-name>.<site-domain> – the internal Version Control Service (git server)

- nexus.<system-name>.<site-domain> – the internal package repository server
- sma-grafana.<system-name>.<site-domain> – Grafana Metrics Dashboard
- sma-kibana.<system-name>.<site-domain> – Kibana Logs Dashboard
- <static ip address> - Internal DNS service

Customers access CSM services from their site networks through a Customer Access Network (CAN) which can be customized as needed by each site. In versions of CSM up to and including Shasta v1.4, the CAN exists solely on the management network, rather than the high speed slingshot interconnect. In Shasta v1.5, the CAN IP addresses will be accessible from the high speed network. Where sites support connections to the internet, the default egress route follows the CAN path as well. It is important to note that all compute nodes access datacenter services through this same pathway. At scale, using the Shasta v1.4 CAN architecture, it is possible for CSM management switches to become a bottleneck for access to datacenter services. There are mitigation features available for data streaming and storage access that shift traffic away from the management switches. These mitigations are beyond the scope of this paper.

1) How does CSM access site services: CSM can operate in a completely air-gapped fashion where necessary, but it is designed to integrate with site services when they are available. The three most notable site services that CSM can leverage are DNS, NTP, and LDAP. For each of these, all management Kubernetes Worker nodes need an IPv4 route to the upstream services. Because CSM does not implement Network Address Translation(NAT) to route between private IP addresses and public ones, each Worker node needs an IPv4 Address on the customer network to use for access to site services. With the CAN in place, each Worker node receives a static IP address within the /24 of space allocated to the system and a route to site services. By default in Shasta v1.4, the Worker nodes use the IP address of the switch directly connected to the site network as their default route. If sites choose not to implement the CAN, they will need to provide each Worker node with an IPv4 address and dedicated NIC for accessing site networks.

a) Integrating with site NTP and DNS: NTP and DNS are the most common site services for CSM. Both services support an underlying protocol that can cluster internally. By default, that is the configuration for air-gapped systems. It ensures that DNS will function for all internal lookups and that all nodes in the system will have an internally consistent representation of time. It is important to note that without an upstream reference clock, the internal system time may be significantly off from the time recognized by the rest of the world. While this will work for most things, there are time-based certificates in the system certificate bundle. If the internally consistent time is too far removed from the time reference of the certificates, some TLS services will fail with certificate errors. Once site DNS and NTP services are available, CSM will use standard protocol features to integrate the upstream services with the internal services. In this case, and under normal operation, CSM acts as a buffer or cache. Every internal request is answered by an internal service which will request answers from an upstream service infrequently in order to maintain its own cache.

b) Integrating with site LDAP: LDAP is a bit less common and there are several integration options. The first use of LDAP in a CSM-based system is to provide authentication for administrators to access the CSM REST APIs and Web Dashboards. For some systems, this will be the extent of the integration. In this case, a service running within CSM called Keycloak is configured at install-time to connect to an LDAP directory to download information about each user through a federation setup. This can include, but is not limited to, names, email addresses, login IDs, SSH public keys, home directories, default shells, and UID/GID mappings. It is possible to configure all of this information manually within Keycloak as well. However, support for federation allows sites to control all of this information centrally. Changes in LDAP will be reflected in Keycloak within a configurable refresh time limit. If the site chooses to extend API access to users, the federation also allows for differential permissions based on what LDAP groups a user is part of. Keycloak provides the actual support for Role Based Access Control (RBAC) within a CSM system. LDAP federation provides a convenient method for central management of admin/user information. Integrating LDAP for authentication of users for access to management nodes, User Access Nodes (UANs), and Compute Nodes is also possible, but a less common request from CSM customers. The installation procedure provides options for this parallel integration. Of note, LDAP does not generally support federation for checking passwords. It is bad security practice to allow multiple systems to store the same password. Consequently, every time a password needs to be verified, the node involved must securely hash the user input and send it to the upstream authentication server. In the case of CSM, that cannot be an intermediate server and must be the official site LDAP. This means that each individual node that wants to check user passwords must have an independent route to the site LDAP server.

D. Accessing CSM Services

As outlined above, CSM also makes several services available to the site through resilient Virtual IP addresses (VIPs) that are not tied to any single Worker node. The implementation of these VIPs relies on a software load balancer within CSM called MetalLB and the standard Border Gateway Protocol (BGP) implementation on the management switches. The software and switches communicate through a peering arrangement to exchange information about the optimal routes for packets. Site network hardware does not need to peer with CSM hardware or software to support these VIPs. Within Kubernetes, routing and security rules operate at the API Gateway level rather than the VIP level. Multiple VIPs can be established for each Gateway without issue.

1) Using the High Speed Network for the default route: By default in Shasta v1.4, the default route for all traffic exiting the system is through the management network and the CAN. Nodes without a public IP address cannot access the Internet directly. Sites that wish to use the edge router IP address(es) as the default routes for the compute nodes may do so by adjusting the compute image. There are static routes in the base image that will provide compute nodes with access to NMN services over the NMN interface.

E. High Speed and Management CAN (CSM 1.5 Roadmap Features)

In CSM 1.5, the CAN network architecture is evolving to provide additional connection options as shown in Fig. 2. A new network pathway through the Slingshot interconnect is being added to provide CAN connectivity at much higher speeds. The High Speed CAN (CHN) can be optionally configured at installation time. It allows access to the same CAN IP subnet with routing provided by the edge router(s). Adding the CHN does not disable the traditional CAN on the management network. It is up to the site network administrators to choose the preferred routes external to CSM. Additionally, customers may choose to separate administrative access from user access through a new dedicated management network on the CAN. The Management CAN (CMN) is only available on the management network and is configurable at install time. If desired, the CMN is configured within the management network as a dedicated VLAN with its own dedicated subnet, allocated by the site. If it is configured, access to system administration tasks will be disallowed on the CAN and only permitted on the CMN.

F. Routing IP Traffic over the CAN, CHN, and CMN

A fundamental assumption of the overall Shasta network architecture is that packets should remain on the network on which they are initiated. Following from that, traffic from the compute nodes to the management nodes via the Node Management Network will never cross over to another network. To aid this, if users want jobs to target CHN services, they will need to use the CHN domain name of that service to ensure correct routing. For example, to obtain a JWT for use with the Shasta APIs from a compute node, the correct hostname will be formed from the service name (auth) along with the subdomain for the CHN (chn) and the system name and site domain. For an example system "sn-2024" with the site domain "dev.cray.com" the CHN domain name would be auth.chn.sn-2024.dev.cray.com. By default, CSM will configure a CNAME for the authentication service without a network subdomain that points to the CAN address of the authentication service. This is adjustable post installation.

G. Accessing Nodes and Services through CSM networking

The CAN architecture and routing described above provide IP connectivity to CSM-managed systems. This section will describe the ways CSM allocates and uses the delegated IP addresses to provide access. Each IP address assigned by CSM also receives forward and reverse name services through the CAN DNS services. CSM manages the default egress, API connectivity, and Node access through the CAN.

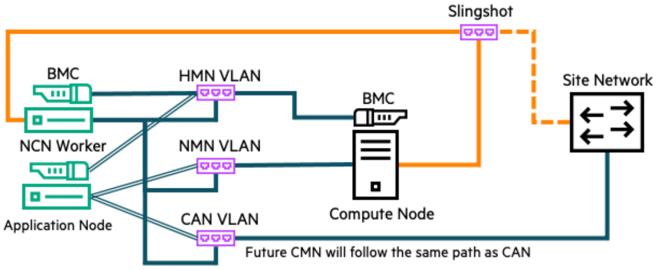

Fig. 2. Detailed VLAN diagram

The first use of the CAN is to allow packets destined for the datacenter network to traverse the management switches for routing at the datacenter level. At install time, the site can choose which IP subnet and VLAN should be used for the default egress of the system.

The second use of the CAN is to allow admins and users to access CSM's REST APIs. Admins need administrative control but users also need access to some of these APIs for certain services and other information. For this purpose, CSM allocates one or more virtual IPs on the CAN which it registers in DNS and allocates TLS certificates. Each of these virtual IPs can float among the management NCNs as needed and even be used in parallel for load balancing through automatic route management with BGP. The route management is purely internal to CSM.

The third use of the CAN subnet is for providing access to User Access Instances (UAI), User Access Nodes (UAN), and other Application Nodes (AN). UAIs are containerized virtual machines, useful for submitting jobs and daily work on the supercomputer. They are private to individual users and created on-demand, providing access to compilers, debugging tools, and workload managers. Because of the way they are managed within CSM, several UAIs can share a single IP address which may change over time. However, users are assured to get the same environment each time through the internal SSH dispatcher. The dedicated UANs and ANs each consume a single IP address on the CAN for each node. Unlike UAIs, the IP addresses of the UANs and ANs should not change from one boot to the next.

III. Access Control for Management APIs

The design of CSM as expressed in both the network and software architecture puts policy enforcement at the center of access control and security. With policy expressed in network Access Control Lists (ACLs), firewalls, Identity and Access Management (IAM), and Attribute-based Access Control (ABAC)[1], CSM provides layers of overlapping security access controls. Each layer builds on the one before it with more targeted access control and policy expressed with more detail.

[1] CSM supports Attribute-based and Role-Based access control policies as appropriate. See external reference for more details on both. https://www.okta.com/identity-101/role-based-access-control-vs-attribute-based-access-control/

ACLs keep network traffic separate. Firewalls restrict IP access at the host level. IAM requires strong authentication of a person. ABAC enforces authorization based on the intent of an action. Without the implied security of dedicated physical networks, together these layers secure the system.

A. Access Control Lists and Firewalls

In Shasta v1.4, the management switches are not configured with detailed ACLs impacting customer access as part of the installation. Site service and support personnel can work to customize what ACLs exist after installation is complete according to the needs of the site. With Shasta v1.5, ACLs and route-maps will be key to keeping the various parts of the CAN in their appropriate places. Notably, the exit-leaf switches and edge routers will have ACLs to prevent unauthorized access. In particular, the CMN configuration will use ACLs to reject administrative traffic from unauthorized sources. Further customization in v1.5 and beyond will improve the segregation between networks inside the system as well. Host firewalling on the management nodes is managed by Kubernetes on each node where it protects access to various services from various networks. Firewalls on the UANs and compute nodes are customizable by the site to fit their requirements.

B. Identity and Access Management with Attribute-based Access Control

In addition to traditional POSIX permissions on compute nodes, CSM supports cloud-like authentication and authorization patterns for access to CSM services. As part of the installation, CSM asks for LDAP directory credentials to bind to a site LDAP server. This integration allows site administrators to identify which LDAP groups are permitted to access the system and manage their group permissions. For instance, at the time a user logs in to the system through the API, his or her group memberships are synced to CSM and available for policy actions. Administrators can customize how that access is managed from the Keycloak dashboard within CSM. Users who are deactivated centrally will be barred from system access once their session expires. Session expiration time is also customizable within the Keycloak dashboard. Using LDAP group membership to permit or deny access through Keycloak is how RBAC is supported within CSM and should be familiar to datacenter administrators.

C. Policy Based Access Control to Microservices

Cray System Management (CSM) is functionally a single set of APIs that span several microservices with a distributed and centralized model for resiliency and consistency. Primary access control logic rests within a centralized/distributed[2] service called an API Gateway. An API Gateway permits or denies access to microservices through policy-based access control logic. CSM uses open policy agent (OPA) to distribute and process policies. OPA is a cloud-native tool that is rapidly becoming a standard within the Kubernetes community. It is based on the Rego policy language from Styra which is very clear and expressive for writing policy as well as fast and efficient for processing. Decisions made by OPA are highly cacheable and do not substantially affect the time to process an API call.

Depending on the policy, some users can be limited to a subset of the APIs or a subset of the functions on those APIs. In Shasta v1.4, a single API gateway and set of authorization policies are included. Future versions of CSM may support the deployment of multiple, parallel API gateways on different network endpoints with different policies. For sites that require strict separation of access without deploying multiple copies of the same microservices, the parallel API gateways may offer a compelling solution in v1.5 and beyond

A small set of policies are included with Shasta v1.4. Where needed, the policy can take into account the source network and many other characteristics in deciding whether to permit the API call to continue. A simplified example of the appropriate policies would look something like this:

1. Reject requests to shut down the system from users that do not have an Administrative Role.
2. Reject requests to shut down the system from any network except the Administrative Network.

Policy Based Access Control is a fundamental cloud-native building block allowing replacement of the traditional hardware-based physical separation of user and administrative concerns.

IV. Authentication Options for Nodes

Authentication of users and administrators for SSH-based access to nodes within an HPE Cray EX system running CSM follows a different path than API authentication and authorization. Site administrators have several integration options for Linux authentication, all based on traditional Linux authentication mechanisms. If sites choose to use an upstream LDAP directory for managing all admins and users for both API access and for standard logins to NCNs, UANs, UAIs, and Compute nodes, the considerations in the above section on "accessing site services" applies. Site admins have other options as well. Through modifications to the system images and/or through post-boot customizations using CFS, UANs and Compute nodes can be configured to use any Linux authentication supported at the site. This includes distribution of flat files as well as more complex integrations like Kerberos or other NSS-based authentication systems. The integration of one authentication system for administrative APIs has no limiting effect on choices available for node login options.

V. Summary

CSM's network and software architecture are targeted at datacenter integration and evolving to meet datacenter needs. The network architecture roadmap evolves the Shasta v1.4 Customer Access Network to enable better network segregation and higher speed connections. The layered authentication and authorization frameworks provide multiple decision points for assessing user intent, with tooling to manage each layer improving in each release. POSIX interactions, which are key to most HPC jobs are not impacted by these changes. With

[2] The configuration is centrally managed, but the service itself can scale horizontally without any practical limit on throughput. For more on this topic, refer to Istio documentation

datacenter integration, CSM enables the future of cloud-like HPC interactions and streaming AI workflows.