HPE Cray Supercomputers: System User Access

User Access Node or User Access Instance, which is right for me?

Jeff Keopp and Alan Mutschelknaus
Hewlett Packard Enterprise
Bloomington, MN USA
keopp@hpe.com, alanm@hpe.com

***Abstract* -- User Access Nodes (UANs) and User Access Instances (UAIs) represent the primary entry point for end users of the new HPE Cray Supercomputers. While UANs align closely with the eLogin nodes on prior Cray systems, UAIs offer a new cloud-like approach with dynamic, single user containers which provide portability, and user isolation. Together, UANs and UAIs offer complementing feature sets that benefit different sets of users. This paper will discuss the features of UANs and UAIs to help customers choose which implementation best suits their needs. Differences from the eLogin nodes used by previous Cray systems will also be discussed.**

Keywords: User Access, User Access Node, User Access Instance, HPE Cray Supercomputers

I. Abbreviations used in this document

Abbreviation	Meaning
CAN	Customer Access Network
CFS	Configuration Framework Service
CMS	Cray Management System
COS	Cray Operating System
CPE	Cray Programming Environment
CPS	Content Projection Service
IMS	Image Management Service
HMN	Hardware Management Network
HMS	Hardware Management Service
HSM	Hardware State Manager
HSN	High Speed Network
NMN	Node Management Network
UAI	User Access Instance
UAN	User Access Node
UAS	User Access Service
WLM	Workload Manager

II. Introduction

User Access Nodes (UANs) and User Access Instances (UAIs) represent the primary entry point for end users of the new HPE Cray Supercomputers. They provide the environment for users to develop, compile, launch jobs on the compute nodes, and analyze the results. While UANs align closely with the eLogin nodes which performed this role on prior Cray systems, UAIs offer a new cloud-like approach with dynamic, single user containers, which provide portability, and user isolation. Together, UANs and UAIs offer complementing features that benefit different sets of users. This paper discusses the features and limitations of UANs and UAIs in the current v1.3 release to help customers choose which implementation best suits their needs. Differences from the eLogin nodes used by previous Cray systems will be discussed along with UAN/UAI image management, lifecycle management - including the use of the switchboard utility for UAIs, user access control and security, networking, user workflows, and sizing UAN and UAI hosts for the number of desired concurrent users.

III. User Access Overview

A. User Access Node (UAN)

User Access Nodes are multiuser physical nodes running the software necessary for end users to develop, build, and launch their applications on the new HPE Cray Supercomputers. They are in many ways like the eLogin nodes used on previous Cray systems. Like eLogin nodes, all users of a UAN share the same OS image, tools, and process space.

Key improvements of the UAN over the eLogin are the implementation of the workload manager (WLM) client commands, the Cray Programming Environment (CPE), and a direct network connection to the high-speed network (HSN) which connects to the compute nodes and shared storage.

WLM (SLURM, PBS Pro, etc.) client commands are now installed and executed on the UAN. On eLogin, the WLM commands were not actually installed on the eLogin node. They were symbolic links to “eproxy” which executed them remotely on a network gateway node. This led to possible issues where the environment on the eLogin was different than the environment on the network gateway node where they ran.

The Cray Programming Environment (CPE) is now projected to the UAN instead of being built into the image

as they were on eLogin nodes. The eLogin images would grow quite large, to hundreds of gigabytes, when multiple versions of CPE were installed. These improvements greatly reduce the size of the UAN image and simplifies updating CPE to a process which does not require rebooting the UANs.

UANs do contain local disk for scratch and swap space, but the UAN is network boot only, in the current release, so any changes not saved to a shared persistent filesystem, such as ClusterStor or remotely mounted home directories, will be lost on a reboot of the UAN.

The UAN may be connected to customer networks in multiple ways. See the networking description in section VI.

B. User Access Instance (UAI)

User Access Instances are in many respects a single user containerized version of a UAN. Each UAI is created for, and accessible by, a specific user. Users may run a UAI image that is customized to their individual needs and may run one or more UAIs. Each UAI may use the same image or different images. This allows a single physical host node to run multiple images simultaneously. Images are registered and made available for use by the system administrator in the Cray User Access Service (UAS). UAS manages the lifecycle of UAIs and is discussed later in this paper.

UAIs do not have local disk and any changes not saved to a shared persistent filesystem, such as ClusterStor or remotely mounted home directories, will be lost when the UAI is deleted. It is important to note that UAIs do not have any local scratch or swap space and are only connected to the customer networks via the Customer Access Network (CAN). It is possible to run a UAI out of memory. However, because the UAI is a single user container, only that user is affected by an out of memory error.

The workload manager (WLM) client commands are natively installed in the UAI image. The Cray Programming Environment (CPE) is volume mounted on the UAI from the UAI host server. It is not installed in the UAI image. Updating CPE requires mounting the desired version(s) of CPE on the UAI host servers and creating a new UAI that is configured to volume mount the CPE version desired. UAIs are hosted on designated non-compute nodes of the HPE Cray Supercomputers.

IV. Image Management

The UAN and UAI images are built by the HPE Cray Image Management Service (IMS). This is the same service that builds and manages the Compute node images. Images are built using Kiwi-NG recipes and configured using the HPE Cray Configuration Framework Service (CFS). CFS operates in two modes, image configuration and node personalization. Node personalization is performed on UAN nodes after the image has booted.

The UAI image goes through an additional process after the image is configured by CFS. It is containerized and must be registered with the Cray User Access Service Manager (UAS). UAN images are automatically registered with IMS.

The default UAN and UAI images are very closely related to the Compute node image (COS). This ensures compilations on UANs and UAIs are done with libraries that match those of the Compute nodes where the jobs will execute.

V. Lifecycle Management

A. User Access Node (UAN)

User Access Nodes are booted, rebooted, and shut down by the HPE Cray Boot Orchestration Service (BOS) in a similar manner to the compute nodes. BOS uses session templates to define which nodes to act upon, which image to boot, whether to run the Configuration Framework Service (CFS) after booting the image, and which CFS branch to use for configuration data.

B. User Access Instance (UAI)

User Access Instances are managed by the HPE Cray User Access Service (UAS). Users and system administrators primarily interact with the UAS using the Cray CLI ('cray uas' command) which calls the UAS APIs. Like all of the management services, UAS is a RESTful service, so the API documentation could be used as a guide to writing your own tools to manipulate the UAS instead of using the provided Cray CLI method.

Cray CLI supports the following user commands for UAS:

- **create** – Create a UAI for the current user using the default UAI image or a user-specified image.
- **delete** – Delete one or more UAIs
  - Only the admin can delete UAIs that belong to other users – see administrative commands.
- **list** – List all UAIs for the current user and their status
- **images list** – List available UAI images.

Cray CLI supports the following administrative commands for UAS:

- **uais list** – List all UAIs or those for a given user.
- **uais delete** – Delete all UAIs or those for a given user.

1) Cray Switchboard

In addition to the Cray CLI, users can use the Cray Switchboard utility. The switchboard automates much of the Cray CLI workflow for managing a user's UAIs. When run interactively as a CLI, the switchboard command supports the following operations:

- **start** – authenticate the user, start a UAI, and connect the user to it.

- **delete** – delete the UAI
- **list** – list all the user's UAIs

The system administrator can configure the Cray Switchboard as a ForceCommand in sshd_config. This allows users to ssh to a node running switchboard and be authenticated and then automatically create a UAI and connected them to it, if they don't have one running, be connected to their existing UAI, or presented a list of their UAIs if they have more than one. The user then chooses one and will be connected to it. The Cray Switchboard is installed on UANs.

VI. Networking

User Access Nodes and User Access Instances would not be very useful if the users couldn't access them and the system administrators couldn't manage them. This section discusses the networking configuration of UAN and UAI.

A. User Access Node (UAN)

The standard networking configuration of a UAN is the following:

- Hardware Management Network (HMN) – connects to the dedicated Baseboard Management Controller (BMC) port to provide remote power control and console access plus other functions of the Hardware Management Service (HMS). Console messages logged by the Cray Conman Service. The HMN is restricted to admins and those services that are authorized.
- Node Management Network (NMN) – This network is used for network booting and other Cray Management Services (CMS) and the System Monitoring Framework (SMF). System administrators also have access to the UAN over this network from the management and worker nodes of the HPE Cray Supercomputer. The NMN is restricted to authorized users and services.
- High Speed Network (HSN) – 100GbE connection used for ClusterStor shared storage and compute node traffic.
- Customer Access Network (CAN) – 2x 40GbE bonded interfaces. This network is connected to the customer end user network for user logins and moving data to/from the UAN and the customer data center networks.
- Additional networks may be added by the customer site to address specific networking needs.

B. User Access Instance (UAI)

The standard networking configuration of a UAI is the following:

- Cluster Network – This is an internal network used by the Kubernetes cluster for communication among cluster members.
- Customer Access Network (CAN) – 1Gb connection to the customers end user network via a macvlan bridge on the UAI host.
- Additional networks may **not** be added by the customer.

VII. Workflows

The following examples show a simple workflow example where a user needs to build an application, queue that application in the WLM to run on the compute nodes, and then analyze the results of the application job.

A. User Access Node (UAN)

- User logs into the UAN via SSH.
- User builds their application using CPE.
- User submits their application job in the WLM queue using WLM client commands.
- After the job completes, the results need to be in the shared storage of the HPE Cray Supercomputer, ClusterStor, for example.
- User analyzes their results.
- User logs out of the UAN.

B. User Access Instance (UAI)

Because UAIs are single user containerized entities, they require a workflow which may be unfamiliar to previous Cray users. Before a UAI can be used, it must be created. The Cray CLI tool provides the command line interface to create and launch UAIs on the HPE Cray Supercomputer. Once launched, the user uses SSH to log into their UAI. This process is similar to using the Cloud-Native APIs from Amazon and Google to launch instances of their services. The Cray Switchboard utility is provided to facilitate this new workflow and make it more like that of the UAN. The Switchboard utility can be configured to allow users to SSH to the host running it, and be automatically routed to their existing UAI, if just one, or be presented with a list of their existing UAIs from which to choose. If the user has no existing UAIs, it creates one for them and then passes the user to it.

Note that the only difference between UAN and UAI workflows are the first and last steps.

- User creates a UAI
  - Cray CLI (*cray uas create* command)
  - Or via Switchboard
- User logs into their UAI via SSH
- User builds their application using CPE.
- User submits their application job in the WLM queue.
- After the job completes, the results need to be in the shared storage of the HPE Cray Supercomputer, ClusterStor, for example.
- User analyzes their results.
- User logs out of their UAI.

- User deletes their UAI or leaves it running for later use.

VIII. User Control and Security

Controlling access to the system is important. Fitting into the customer's existing user control system is important. The following subsections describe user control and security for UAN and UAI.

A. User Access Node (UAN)

The UAN is a Linux server. The customer can install whatever Linux supported user control system they require. Support for LDAP with SSSD is included by default.

B. User Access Instance (UAI)

The UAI is implemented as a container managed by Kubernetes. The UAI container runs as the user and is only accessible by the user or root user using SSH keys. SSHD is running as the user and the UAI container is unprivileged. It does not have direct access to host networking. Creation of the UAI is restricted to authorized users using LDAP or local Keycloak accounts.

IX. Sizing

Computing how many UANs or UAI hosts are needed depends on the type of workload which is expected. The following formula is used to get an idea of how may UANs or UAI hosts are recommended.

UANs or UAI hosts for interactive use	
Memory size per user	8GiB
Fractional (Skylake) CPU per user	1/32
# UANs or UAI hosts = max(#of users/(UAN or UAI host memory size/memory size per user), # of users/UAN or UAI #CPUs/fractional CPU per user))	

UANs or UAI hosts for compiler use	
Memory size per compile	16GiB
Fractional (Skylake) CPU per compile	1/8
# UANs/UAI hosts = max(#of users/(UAN or UAI host memory size/memory size per compile), # of users/UAN or UAI #CPUs/fractional CPU per compile))	

NOTE: UAIs do not support swap space whereas UANs do have local disk for swap.

The following table show the results for the given parameters:

Example Parameters	
Cores per UAN or UAI host node	128
Memory per node	512 GiB
Number of interactive sessions or UAIs	100
Number of compilations or UAIs	50
Number of analysis jobs	20

Interactive Use	
Memory size per user	8GiB
Cores per user	1
Required Memory (calculated)	800GiB
Required Cores (calculated)	100
Compilation Use	
Memory size per user	16GiB
Cores per user	4
Required Memory (calculated)	800GiB
Required Cores (calculated)	200
Analysis Use	
Memory size per user	32GiB
Cores per user	8
Required Memory (calculated)	640GiB
Required Cores (calculated)	160
UAN or UAI host counts needed	
Required for Memory (calculated)	5
Required for Cores (calculated)	4

X. Limitations

A. User Access Node (UAN)

Currently, the UANs do not boot from local disk. This makes rebooting dependent on the Cray Management System (CMS) being available. The Cray Programming Environment (CPE) is provided by the Content Projection Service (CPS) which is part of the CMS.

B. User Access Instance (UAI)

The UAI is a Kubernetes managed pod and therefore relies on the Kubernetes service being active. UAIs also do not have local disk for swap so it is possible to run a UAI out of memory. UAN nodes have swap space available to help prevent this from happening. The network bandwidth for customer access network (CAN) is limited to 1GbE vs the 2x 40GbE bonded CAN network of the UAN. The 1GbE connection is per UAI host node and is shared among all UAIs being hosted on that node. The 2 bonded 40GbE interfaces on UANs are shared by all users on that particular UAN.

XI. Future Work

There will be work in the future to improve the functionality and usability of both UAN and UAI based on customer feedback as we improve the user access experience and the management of the user access experience for HPE Cray Supercomputers.