UNR High Performance Computing (HPC)
General Overview and Business Model

Executive Summary

High Performance Computing (HPC) at UNR consists of a set of shared hardware and human resources in support of research and instruction requiring large-scale or high capacity computation, data storage, and/or data networking. Here, the term "HPC" refers to computational resources of a scale that would not be found in a typical workplace desktop or laptop computer. Commonly, this would be in the form of cluster computing, "big data" storage and high bandwidth networking. The HPC program is summarized as follows.

Overview of HPC Resources

Three Types of HPC Resources at UNR:

Local Resources: These consist of specialized, application-specific, or individual research group resources. An example would be an individual computing cluster that is not generally available to the campus at-large (for any number of reasons). Local resources are not generally managed or significantly supported by central IT staff as part of their normal duties, but this may be arranged individually on a fee basis.

Centrally-Shared Resources: UNR maintains a set of commonly-available HPC resources for efficient and cost effective shared use by the campus community. These shared resources are significant in scale, and are made available through a combination of funding channels, including both centrally-provided funds and contributions (or "investments") made by researchers through grants, contracts, startup packages, etc. *A significant portion of this document will address these centrally-shared HPC resources in detail.*

Off-Premise Cloud Resources: Users have access to off-campus computing resources (government, academia, and commercial) as arranged by the individual researchers or users. UNR's central IT team is charged with facilitating this and supporting UNR users in the processes of gaining access to and using these off-campus resources.

The Process of Using Shared HPC Resources: Access to UNR's shared resources requires a user account and a minimum level of training and account configuration (software, etc.). The resources are accessed remotely (such as from an office workstation) through a *secure shell* (ssh) login procedure (or similar). Computational tasks will typically be run from the "command line", in a Linux environment. Each user will have a storage quota, with the availability of additional storage space for a fee. Computational jobs will be run through a queue submission process, in which the user's job will wait its turn to begin, and then run to completion. The length of the wait in the start queue will be prioritized through a selection of factors including the user's investment level, the job size, expected job duration, and the extent of recent use of the system by the individual user.

Investing Options

Access to the shared campus HPC resources is organized into three "Tiers" of use. These include up-front leasing for a fixed period, pay-as-you-go use, and a "no fee" use (with restrictions). Due to the overall nature of what is being provided, each of these HPC investment types are considered *services* (not equipment purchases or leases).

Tier 1: Investors

Tier 1 is a "service contract" based lease plan that is available in 4-year increments for a maximum of 8 years (to align with the underlying hardware warranty term) and is paid for up-front (non-refundable). Three types of HPC resources are available to investors:

Tier 1 - Investors	**$/Node-Eq.* (4 years)**	**$/Node-Eq.* (8 years)**	**Cloud** (for comparison)**
CPU-Based Compute Capability	$9,800 ($0.28/hr)	$14,700 ($0.21/hr)	$56k - $93k ($1.60-2.66/hr)
GPU-Based Compute Capability	$35,500 ($1.01/hr)	$53,250 ($0.76/hr)	$250k ($7.20/hr)
Bulk Storage	$330/TB	$495/TB	($1,250/TB)

** Node Equivalents are based here on 2017 generation hardware: A **CPU node** has 32 cpu cores and 256 GB of RAM, with ~1 TF of theoretical capacity). A **GPU node** adds 14,336 GPU cores, for ~22 TF.*

*** Resource costs from a typical cloud service provider (Amazon EC2 and Google Cloud Storage, August 2017 price sheets) are provided for reference only (4-yr estimate).*

These costs include the physical computing and storage hardware, networking equipment, licenses for core software components, facilities (space, power, cooling, security), and technical support (both system and application support). Compute resources may be purchased in units less than a complete node's worth of computing (with OIT consultation), but the lease terms are restricted to multiples of 4 years. These costs are based on the actual hardware and support costs and factor in a subsidy provided by central UNR funds. Hardware purchases last for an estimated 4 year period, during which the equipment is warrantied and supported by the vendor. These costs are based on 2017 estimates and may vary in the future, depending on actual hardware costs and the "financial health" of the HPC program at UNR. Lease amounts beyond the first 4 year period are based on projected decreases in costs for computing power (based on historical trends). At present, the UNR HPC business model assumes that future costs (for a given amount of computing capacity) will be halved every four years. Thus, the cost of an 8 year investment is 150% the cost of a 4 year investment.

A special subcategory of Tier 1 leases is available for users who require resources that cannot be shared with the rest of the campus. Examples may include uniquely configured hardware, or dedicated use systems (such as a web server) that cannot tolerate the queuing procedures that will be in place for the shared compute resources. These "specialized use" leases may be charged an additional fee above Tier 1 pricing (at OIT's discretion) because the resources will

not be available to contribute to the campus's shared access pool, yet will necessitate additional management and support from central OIT resources.

Tier 2: Renters

Tier 2 is a "pay as you go" plan where users are charged on a monthly basis for their actual usage. Two types of HPC resources are available through Tier 2:

Tier 2 - Renters	$/Node-Eq*-hr	Cloud** (for comparison)
CPU-Based Compute Capability	$0.56	$1.60 - 2.66/hr
GPU-Based Compute Capability	$2.02	$7.20/hr
Bulk Storage	$13.75/TB-month	$26/TB-month

** Node Equivalents are based here on 2017 generation hardware: A **CPU node** has 32 cpu cores and 256 GB of RAM, with ~1 TF of theoretical capacity). A **GPU node** adds 14,336 GPU cores, for ~22 TF.*

*** Resource costs from a typical cloud service provider (Amazon EC2 and Google Cloud Storage, August 2017 price sheets) are provided for reference only.*

Tier 2 plan costs will be updated at every hardware purchase interval (for example, approximately every 6 months), based on evaluation of the UNR HPC business model performance and current hardware costs..

Tier 3: No-Fee Use

The primary aim of the shared HPC resources is to support Tier 1 and Tier 2 (paying) users. However, a portion of the resources will be made available to non-paying users who wish to run smaller-scale computational jobs and are willing to work with a lower job priority level (which will result in longer waits for jobs to run, etc.). This category of HPC resource usage is sponsored directly through OIT/VPRI funding.

UNR's shared HPC resources are available to off-campus affiliates as well. External (non-UNR) users will be coordinated through the Nevada Center for Applied Research and charged at higher rates (to be negotiated).

Use of HPC Resources

An investment in UNR's shared HPC resources will provide the investor with access to a specific amount of computational or storage capability of the unit type that was leased (CPU, GPU, Storage). It does NOT provide guaranteed access to any specific physical piece of hardware. It can be thought of as piece of the "local UNR cloud" that can be configured to suit the investor's specific needs or access common resources such as a campus-wide software license.

The mechanism for leasing HPC resources that are shared with the rest of the campus is quite flexible, and is based on the concept of having a certain number of computational clock cycles

that can be used (with some technical limitations) any time over the period of the lease (ex: 4 yrs). If a user leases 10 Node-Equivalents for 4 years, they can effectively use 10 nodes worth of computational power continuously for 4 years straight with no penalty in job priority. Or, alternatively they can use their resources in "burst mode" (within the physical limitations of the system), where they may use (for example) 40 Node-Equivalents, but for only 25% of the time during their 4-year lease (with some more detailed rules in place). If the user's recent usage exceeds their investment level of use, their jobs may still run, but will be progressively dropped in queue priority (i.e., they are not completely "cut off").

Use of shared HPC resources at UNR is generally managed through a job queuing system (typical of HPC systems). The user submits a computational job to the queuing system, and and will (in due course) be executed. How long a job will wait in the queue before starting will depend on a range of factors including 1) The overall activity (load) of the system, 2) the size of the job, 3) expected run time (wall time) of the job, 4) The level of investment that the user is working with, and 5) The recent usage history of the user. The system is designed so that small and short jobs are expected to start within a maximum of 4 hrs after submission. Medium sized conventional jobs are expected to start within 16 hrs of submission. Longer jobs, or jobs that require a large percentage of available campus resources may require much longer waits before starting. Similarly, a user that has recently or consistently been using more than their leased share may also be dropped in queue priority, leading to longer waits for subsequent jobs. These system design goals may vary over time, as the system size changes to accommodate the ebb and flow of demand.

HPC users are provided with various forms of support from OIT staff, including configuration of resources, software installation, training, and individualized assistance with general matters relating to use of on-campus and off-campus HPC resources. These User Support processes are facilitated primarily through the HPC Application Specialist in OIT.

HPC Governance Processes

The overall governance of the centrally-shared HPC resources at UNR is researcher-driven, in order to best represent the diverse high performance computing needs of the campus. HPC governance is overseen by the campus Cyberinfrastructure Committee (CiC), partnered with the Office of Information Technology (OIT), its staff, and the Office of the Vice President of Research and Innovation. The CiC is engaged in developing usage policies, strategic planning, and general oversight of HPC usage and practices. The CiC consists primarily of campus researchers (users), with representation from the technical team in OIT. The OIT collectively acts as the technical planner, operator, and manager (including financial) of the HPC resources and also serves as the "day-to-day" interface to HPC users across campus. HPC users have access to both the CiC and the OIT to seek information, present suggestions or new ideas, or raise concerns about HPC resources and operations on campus.