

BNL Skylake Cluster Memorandum of Understanding

Revision 0

December 19, 2017

1. Purpose

This Memorandum of Understanding (MOU) describes the agreement among the Computational Science Initiative (CSI) as Operator of the BNL Skylake Computing Cluster resource and the projects, groups and departmental stakeholders of the system. Stakeholders are defined as those groups at the Laboratory with direct financial contributions in the procurement of hardware and are projected to be the primary users of the Cluster. The MOU shall remain in effect for the lifetime of the Skylake Cluster (SC) for all stakeholders, except for the LQCD-ext II Computing Project (LQCD). For LQCD, this agreement shall remain in effect from January 1, 2018 through September 30, 2019, subject to funding availability. This MOU may only be modified with the mutual consent of all parties.

2. Institutional Responsibilities

The CSI is the umbrella organization that will deliver, with its partners, the required services for the Skylake cluster operation. Specifically:

RACF (RHIC-ATLAS Computing Facility)

RACF will provide procurement expertise, co-design of the cluster, along with the full range of scientific computing services, including hardware lifecycle management, OS software (operating system, workload scheduler, configuration management, etc.) provisioning, storage services, user account management, network infrastructure support, tape services, etc.

SDCC (Scientific Data & Computing Center)

SDCC will operate and maintain the Skylake Cluster resources (computing and storage). The distinction is made between SDCC and RACF only for the purposes of assigning effort and responsibilities among the various funding sources. The SDCC is responsible for operations, and the RACF provides basic IT services in support of operations.

CSL (Computational Science Laboratory)

The CSL will provide co-design of the cluster. User support will be jointly provided by the CSL and the SDCC. Usage policy, quotas and allocations will be enforced via the governance mechanism outlined in the associated Operations Guidelines. Subject to infrastructure and operational (as defined by the RACF and SDCC) constraints, it will also procure compute resources (in consultation with the RACF and SDCC) purchased by users that will then contribute to their time allocation.

CSI and the Lab

BNL will provide sufficient space, electricity and cooling to house and operate institutional clusters. The clusters will consist of compute nodes, communication fabrics, management networks, and data storage. Account approval and setup will be provided by the GUV

(Guest, User and Visitor) center and RACF, respectively. The Laboratory will provide infrastructure, communications, facility management and network support.

3. Stakeholder Responsibilities

Stakeholders will purchase a guaranteed annual wall-clock time allocation that is determined by the cost of a specified block of compute resources. Allocations must be renewed annually, subject to stakeholder funding and cluster resource availability.

For purposes of resource management, allocations will be assigned on a quarterly basis following processes outlined in the attached operations guideline. Stakeholders acknowledge that computing resources are time sensitive: Unused computing time on the Skylake cluster is lost, unless prior arrangements (subject to cluster resource availability) with the SDCC have been made. One benefit of a shared facility is that the give and take between the needs of multiple stakeholders can smooth this out. However, the operation of the cluster will include mechanisms to decrement unused allocations as a function of time, as documented in the associated operations guidelines.

Stakeholders will be empowered to direct their allocated resources to specific researchers associated to their program, subject to guidance of the allocation committee and the general guidelines for access to BNL resources, e.g. for researchers outside of BNL.

Stakeholders will report semi-annually to CSI. The report must include lists of published papers, presentations given and proposals funded to which the usage of the Skylake Cluster contributed. The report should also provide science highlights to CSI demonstrating how the use of the Skylake cluster advanced their scientific discovery process.

All published papers, presentations and funded proposals which made use of the Skylake cluster must include the following acknowledgement:

“This work was supported by resources provided by the Scientific Data and Computing Center (SDCC), a component of the Computational Science Initiative (CSI) at Brookhaven National Laboratory (BNL).”

4. Skylake Cluster Allocation Committee

While some stakeholders will purchase time for specific projects and services, others such as BNL lab management and CSI, will purchase allocations in support of novel research projects. These allocations are assigned on a competitive basis. The purpose of the Allocation Committee is to solicit, review (or modify as necessary) and approve proposals to use the Skylake Cluster. Time allocation is based on the guidance of the Allocation Committee. The committee membership will include: a) representatives of the stakeholders, b) RACF, c) CSL, and d)

representatives of appropriate BNL science directorates to insure allocation time is consistent with short and long-term goals at the Lab.

Appendix A. Skylake Cluster Description

The Skylake Cluster is configured with the following specifications:

1. 64 compute nodes consisting of the following
 - a. 2 Intel Skylake Gold 6150 CPU's with a total physical core count of 36 and clock speed of 2.7 GHz.
 - b. 16 TB SATA storage for data, swap and OS
 - c. 192 GB ECC RAM
2. Two master nodes with 400GB of disk storage for login access.
3. Non-blocking EDR fabric
4. 1GbE network fabric for cluster management
5. Access to 1 PB of usable RAID 6 storage capacity managed by GPFS with up to 24 GB/s bandwidth accessed via EDR

Appendix B. Stakeholder Allocations

Currently all 64 nodes (see table below) and 200 TB (out of 1 PB of usable storage) have been assigned to LQCD. The 200 TB is shared among Institutional, KNL and Skylake clusters. LQCD storage allocation is also documented in the BNL Institutional Cluster MOU and must not be double-counted. The total storage allocation to LQCD is 200TB.

Table 1. Institutional Cluster Compute Allocations

| Stakeholder | Type | Compute Allocation (# nodes) | Compute Allocation Period | Compute Allocation (node-hrs) | Compute Allocation Cost (\$K) |
|-------------|-----------|------------------------------|-----------------------------|-------------------------------|-------------------------------|
| LQCD | Secondary | 64 | Feb 26, 2018 – Sep 30, 2018 | 331,776 | 301.916 |
| | | | | | |

Table 2. Institutional Cluster Storage Allocations

| Stakeholder | Type | Storage Allocation (TB) | Storage Allocation Period | Storage Allocation Cost (\$K) |
|-------------|------|-------------------------|---------------------------|-------------------------------|
| | | | | |
| | | | | |

Appendix C. Capital and Operational Costs

Costs are divided into capital (computing, storage, all software licenses, etc.) and operational expenses. Operational expenses can be further subdivided into physical (power, cooling and space) infrastructure, cyber (gateway servers, account management, network connectivity, etc.) infrastructure and staff support.

Since allocation is done on a whole node basis, capital and operational costs are calculated in units of node-hr. The capital cost of computing is \$0.58 per node-hour. Cyber infrastructure costs are \$0.06 per node-hour. The cost of staff is \$0.35 per node-hour-FTE. For storage, the capital cost (including required licenses) is \$9.50 per TB-month. All figures include BNL overhead.

The following table summarizes the cost model.

Table 3. Institutional Cluster Cost Model

| Stakeholder | Computing (per node-hr) | Cyber (per node-hr) | Staff (per node-hr-FTE) | Total Cost (per node-hr) | Storage (per TB-month) |
|-------------|----------------------------|------------------------|----------------------------|-----------------------------|---------------------------|
| Primary | ----- | \$0.06 | \$0.35 | \$0.41 | \$9.50 |
| Secondary | \$0.50 | \$0.06 | \$0.35 | \$0.91 | \$9.50 |

The capital and operational costs above were calculated using current (as of June 2016) expenses and are subject to change. Operational costs are expected to drop as the Skylake cluster expands, because they do not grow linearly with the number of nodes in a cluster. Costs will be reviewed (and adjusted accordingly) on an annual basis near the boundary between fiscal years. Invoices will be generated on a quarterly basis and sent to all stakeholders.

Below are some hypothetical use cases:

Example 1: Stakeholder A requests 30 nodes for 20 days and 100 TB of storage for 18 months. Computing cost is \$7,200 ($\$0.50/\text{node-hr} \times 30 \text{ nodes} \times 480 \text{ hr}$), storage cost is \$17,100 ($\$9.50/\text{TB-month} \times 18 \text{ months} \times 100 \text{ TB}$) and cyber infrastructure cost is \$864 ($\$0.06/\text{node-hr} \times 30 \text{ nodes} \times 480 \text{ hr}$). The formula for staff cost is $(\$0.35/\text{node-hr-FTE}) \times (1.0 \text{ FTE}) \times (\# \text{ of nodes}) \times (\# \text{ of hr})$, which is \$5,040 in this example, so the total cost is \$30,204.

Example 2: Stakeholder B would like to invest \$10k on computing resources with an estimated 100 TB of storage for 2 months. Using a cost of \$0.56/node-hr ($\$0.50 + \0.06) for computing and infrastructure, \$0.35 per node-hr-FTE for staff cost and subtracting \$1,900 ($\$9.50/\text{TB-month} \times 100 \text{ TB} \times 2 \text{ months}$) for storage cost, the initial \$10k investment buys $\$8,100 / (\$0.56/\text{node-hr} + (\$0.35/\text{node-hr-FTE}) \times (1.0 \text{ FTE})) = 8,901$ node-hr of computing.

Appendix D. Storage Services

Storage Management

Beyond the volatile local scratch disk of 1.5 TB/machine and user home directory (40 GB/user), all other disk storage space is actively managed to support operational readiness and avoid unexpected loss of

computing time. The SDCC will coordinate with stakeholder representative and insure storage usage is consistent with agreed-upon allocation. Once storage allocation ends, stakeholder will have 30 days to move data somewhere else or back up to tape storage (see optional service below). Data will be deleted, and disk space recovered after 30 days.

Optional Services

Tape back-up:

Access to tape storage is possible if long-term storage of precious data and software is needed. Archival storage (write once and then only accessed to restore lost data on disk) is the most cost-effective solution for back-up support. Estimated cost for archival storage is \$29 per TB per year. This estimate includes the cost of tape and robotic silo slot license. It does not yet include fractional cost of tape drive(s), networking, front-end server(s), software licenses and warranty support.

Usage of tape storage other than archival mode must be discussed with individual stakeholders on a case-by-case basis. Individualized requirements (I/O throughput, storage needs, etc.) require a similarly structured cost model.



William Boroski
Project Manager, LQCD-Ext II

12/21/2017
Date

Norman Christ
Acting Chair, USQCD Executive Committee


Date

Eric Lançon
Chair, Scientific Data & Computing Center
Director, RHIC-ATLAS Computing Facility

Date

Hong Ma
Chair, Physics Department

Date



Kerstin Kleese van Dam
Director, Computational Science Initiative

12/21/2017

Date

Robert Tribble
Deputy Director for Science and Technology

Date