

# BNL KNL Cluster Memorandum of Understanding

Revision 0

December 21, 2017

## 1. Purpose

This Memorandum of Understanding (MOU) describes the agreement between the Computational Science Initiative (CSI) and departmental stakeholders of the KNL cluster. Stakeholders are defined as those groups at the Laboratory with direct financial contributions to the capital procurement and/or operations and are the customary users of the KNL cluster. The MOU shall remain in effect for the lifetime of the KNL cluster for all stakeholders, except for the RIKEN BNL Research Center (RBRC) and the LQCD-ext II Computing Project (LQCD). For the RBRC, this agreement shall remain in effect until April 29, 2018, at which point it shall be reviewed for an extension. 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 KNL 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 KNL 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)

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 guideline. 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 cluster 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 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 KNL cluster is lost, unless prior arrangements 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 BNL 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 KNL cluster contributed. The report should also provide science highlights at least twice a year, demonstrating how the use of the KNL cluster advanced their scientific discovery process.

All published papers, presentations and funded proposals which made use of the KNL 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. KNL 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 KNL Cluster. Time allocation on the KNL cluster is based on the guidance of the Allocation Committee. The committee membership will include: a) CSI, b) a representative for each primary stakeholder, c) representatives of other stakeholders, d) RACF, and e) representatives of appropriate BNL science directorates to insure allocation time is consistent with short and long-term goals at the Lab.

## Appendix A. KNL Cluster Description

The KNL Cluster will initially be configured with the following specifications:

1. 144 compute nodes each consisting of the following
  - a. One Intel Xeon Phi 7230 CPU (64 cores), 16 GB RAM on the chip and clock speed of 1.3 GHz
  - b. 2 x 512 GB high-throughput SSD (with 512 MB internal buffer) for local storage
  - c. 192 GB DDR4 dual-rank RAM
  - d. Dual-rail (2x) Intel Omni-Path Host Fabric Interface Adapter 100 series
2. Intel TOR Omni-Path switches capable of supporting up to 144 compute nodes
  - a. dual-rail, non-blocking
  - b. 400 Gbps peak aggregate bi-directional bandwidth

## Appendix B. Stakeholder Allocations

Stakeholders that contribute to the purchase of the computing portion and the operations of the KNL cluster are defined as primary stakeholders, and stakeholders who make an investment in the cluster with a research grant are defined accordingly. Currently 136 out of 144 nodes (see table below) have been assigned to non-institutional stakeholders. The sharing of storage resources with Institutional Cluster (IC) will be determined later.

Table 1. Institutional Cluster Compute Allocations

| Stakeholder | Type           | Compute Allocation (# nodes) | Compute Allocation Period   | Compute Allocation (node-hrs) | Compute Allocation Cost (\$K) |
|-------------|----------------|------------------------------|-----------------------------|-------------------------------|-------------------------------|
| Lehner      | Research Grant | 36                           |                             |                               |                               |
| LQCD        | Secondary      | 66                           | Jan 14, 2018 – Sep 30, 2018 | 410,256                       | 209.231                       |
| RIKEN       | Secondary      | 36                           |                             |                               |                               |
| CSI         | Institutional  | 6                            |                             |                               |                               |
|             |                |                              |                             |                               |                               |

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.

Research grant stakeholders are not responsible for any capital or operational costs, apart from the initial investment and storage costs. Primary and institutional stakeholders are responsible for operational costs proportional to their fractional share of the cluster. Other users (defined as “secondary stakeholders”) are charged capital, cyber infrastructure and staff support costs proportional to their fractional share of the cluster.

Since allocation is done on a whole node basis, capital and operational costs are calculated accordingly. The capital cost is \$0.18 per node-hour of computing resources. Physical infrastructure costs are \$0.09 per node-hour. All components are volatile and may fluctuate on a yearly basis. The current cyber infrastructure costs are \$0.04 per node-hour. The cost of staff is \$0.20 per node-hour-FTE. For storage, the capital cost (including required licenses) is \$9.50 per TB-month. All costs include BNL overhead.

The following table summarizes the cost model.

Table 3. Institutional Cluster Cost Model

| Stakeholder | Computing (per node-hr) | Physical Infrastructure (per node-hr) | Cyber (per node-hr) | Staff (per node-hr) | Total Cost (per node-hr) | Storage (per TB-month) |
|-------------|-------------------------|---------------------------------------|---------------------|---------------------|--------------------------|------------------------|
| Research    | -----                   | -----                                 | -----               | -----               | -----                    | \$9.50                 |
| Primary     | -----                   | \$0.09                                | \$0.04              | \$0.20              | \$0.33                   | \$9.50                 |
| Secondary   | \$0.18                  | \$0.09                                | \$0.04              | \$0.20              | \$0.51                   | \$9.50                 |

The capital and operational costs above were calculated using current (as of October 2016) expenses and are subject to change. Operational costs are expected to drop if the KNL cluster expands, because they do not generally grow linearly with the number of nodes in a cluster.

Costs will be reviewed (and adjusted accordingly) on a periodic basis. Invoices will be generated on a quarterly basis and sent to all stakeholders.

Storage costs are for enterprise-level, high-performance GPFS-based storage. Support (staff and operations) for non-SDCC-managed storage systems is not included in this MOU.

Below are some hypothetical use cases:

Example 1: Research stakeholder A is assigned 36 nodes and 100 TB of storage for 12 months. Stakeholder only incurs storage cost of \$11,400 ( $\$9.50/\text{TB-month} \times 12 \text{ months} \times 100 \text{ TB}$ ).

Example 2: Primary stakeholder B is assigned 36 nodes and 50 TB of storage for 12 months. Storage cost is \$5,700 ( $\$9.50/\text{TB-month} \times 12 \text{ months} \times 50 \text{ TB}$ ), physical and cyber infrastructure costs are \$40,997 ( $\$0.13/\text{node-hr} \times 36 \text{ nodes} \times 8,760 \text{ hr}$ ) and staff cost is \$63,072 ( $\$0.20/\text{node-hr} \times 36 \text{ nodes} \times 8,760 \text{ hr}$ ). The total cost is \$109,769.

Example 3: Secondary stakeholder C would like to invest \$10k on computing resources with an estimated 100 TB of storage for 2 months. The cost of computing and operations is \$0.51/node-hr ( $\$0.18 + \$0.09 + \$0.04 + \$0.20$ ). The cost of storage is \$9.50 per TB-month, so the \$10k investment translates to 15,882 node-hr of computing, after subtracting \$1,900 ( $\$9.50/\text{TB-month} \times 100 \text{ TB} \times 2 \text{ months}$ ) for storage costs.

## **Appendix D. Other 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. Custom requirements (I/O throughput, storage needs, etc) are addressed by a correspondingly structured cost model.

### Access to Wide-Area Network (WAN):

High-bandwidth access to the WAN must first be negotiated with ESNET and the RACF on a case-by-case basis. Depending on requirements and ESNET approval, purchase and maintenance of additional equipment may be necessary and will be addressed accordingly in the cost model.

---

Sam Aronson  
Director, RIKEN BNL Research Center

---

Date

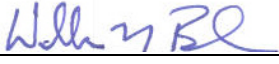
---

Taku Izubuchi  
Computing Group Leader, RIKEN BNL  
Research Center

---

Date

---

  
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

---

Christoph Lehner  
Associate Scientist, Physics Department

---

Date

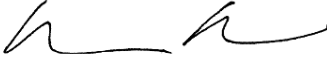
---

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