BNL Institutional 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) as Operator of the BNL Institutional Computing Cluster resource and the projects, groups and departmental stakeholders of the system. Primary 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 Institutional Cluster (IC) 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 institutional 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. The RACF will also support operations of Blue Gene Q until September 2017 and the CFN (Center for Functional Nanomaterials) gen3 and gen4 clusters in the context of CSI until end of operational usefulness.

SDCC (Scientific Data & Computing Center)

SDCC will operate and maintain the Institutional 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 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 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 Guidelines document. Stakeholders acknowledge that computing resources are time sensitive: unused computing time on the institutional 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 Operations Guidelines document.

Stakeholders will be empowered to direct their allocated resources to specific researchers associated with their program, subject to guidance of the IC 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 institutional cluster contributed. The report should also provide science highlights to CSI demonstrating how use of the institutional cluster advanced their scientific discovery process.

All published papers, presentations and funded proposals which made use of the institutional 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. Institutional 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 IC Allocation Committee is to solicit, review (or modify as necessary) and approve proposals to use the Institutional Cluster. Time allocation on the institutional cluster 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. Institutional Cluster Description

The Institutional Cluster is configured with the following specifications:

1. 108 compute nodes, each consisting of the following
    a. Two (2) Intel Xeon E5-2695V4 Broadwell based CPU's with a total physical core count of 36 and clock speed of at least 2.1 GHz.
    b. Two (2) Nvidia K80 GPUs
    c. 180GB SSD for OS/Swap
    d. 1.8TB SAS drive for data and swap.
    e. 256 GB ECC RAM
2. 54 nodes – similar to above but with two Nvidia P-100 GPU's per node, instead of K80s.
3. Two master nodes, each with 400GB of disk storage for login access.
4. Non-blocking EDR fabric with complete complement of spine blades/switches capable of supporting at least 240 compute nodes
5. 1GbE network fabric for cluster management
6. 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

Primary stakeholders are stakeholders which contributed to the purchase of the computing portion of the institutional cluster. Currently 94 of the 108 nodes (see table below) and 200 TB (out of 1 PB of usable storage) have been assigned to primary and secondary stakeholders.

Table 1. Institutional Cluster Compute Allocations

Stakeholder	Type	Compute Allocation (# nodes)	Compute Allocation Period	Compute Allocation (node-hrs)	Compute Allocation Cost ($K)
CFN	Primary	35			
Material Science	Primary	19			
LQCD	Secondary	40	Oct 1, 2017 – Sep 30, 2018	349,440	0
LQCD	Secondary	25	Jan 14, 2018 – Sep 30, 2018	155,400	153.846
LQCD	Secondary	15	Feb 1, 2018 – Sep 30, 2018	86,760	85.892

Table 2. Institutional Cluster Storage Allocations

Stakeholder	Type	Storage Allocation (TB)	Storage Allocation Period	Storage Allocation Cost ($K)
CFN	Primary			
Material Science	Primary			
LQCD	Secondary	200	Jan-Sep 2018	17.1

Appendix C. Capital and Operational Cost Basis

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.

BNL will pay for physical infrastructure costs for the lifetime of the IC. BNL will support for 0.5 FTE, and the stakeholders will cover 1.5 FTE. Individual stakeholder share will be determined by level of effort provided to each stakeholder. The chosen metric unit for level of effort is node-hr).

Primary stakeholders are responsible for certain capital costs (storage and associated software licenses), cyber infrastructure and staff support costs. Other users (defined as "secondary stakeholders") are charged capital, cyber infrastructure and staff support costs.

Since allocation is done on a whole node basis, capital and operational costs are calculated accordingly. 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 costs 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	$0.58	$0.06	$0.35	$0.99	$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 institutional 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 and sent to all stakeholders on a quarterly basis.

Below are some hypothetical use cases:

Example 1: Primary stakeholder A is assigned 40 nodes for 10 days and 200 TB of storage for 2 years. Storage cost is $45,600 ($9.50/TB-month × 24 months × 200 TB) and cyber infrastructure cost is $576 ($0.06/node-hr × 40 nodes × 240 hr). Total cost is $46,176. Staff cost is structured as ($0.35/node-hr-FTE) × (1.5 FTE) × (# of nodes) × (# of hr).

Example 2: Secondary stakeholder B requests 30 nodes for 20 days and 100 TB of storage for 18 months. Computing cost is $8,352 ($0.58/node-hr × 30 nodes × 480 hr), storage cost is $17,100 ($9.50/TB-month × 18 months × 100 TB), cyber infrastructure cost is $864 ($0.06/node-hr × 30 nodes × 480 hr). The total cost is $26,316. Staff cost is charged as ($0.35/node-hr-FTE) × (1.5 FTE) × (# of nodes) × (# of hr).

Example 3: Secondary stakeholder C would like to invest $10k on computing resources with an estimated 100 TB of storage for 2 months. Using a cost of $0.64/node-hr ($0.58 + $0.06) for computing and $0.35 per node-hr-FTE for staff costs and after subtracting $1,900 ($9.50/TB-month × 100 TB × 2 months) for storage costs, the initial $10k investment buys $8,100/(0.64/node-hr + (0.35/node-hr-FTE) × (1.5 FTE)) = 6,952 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 representatives to insure storage usage is consistent with agreed-upon allocations. 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.

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.

Mark Hybertsen
CFN Group Leader, Theory and Computation

Date

Chuck Black
Director, CFN

Date

Samuel Aronson
Director, RIKEN BNL Research Center

Date

Taku Izubuchi
Computing Group Leader, RIKEN BNL Research Center

Date

12/21/2017

William Boroski
Project Manager, LQCD-Ext II

Date

Norman Christ
Acting Chair, USQCD Executive Committee

Date

Robert Konik
Chair, Condensed Matter Physics & Materials Science Department

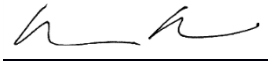
Date

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

Date

Hong Ma
Chair, Physics Department

Date



12/21/2017

Kerstin Kleese van Dam
Director, Computational Science Initiative

Date

Robert Tribble
Deputy Director for Science and Technology

Date