LLNL-TR-404701

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Certification of Completion of ASC FY08 Level-2 Milestone ID #2933

D. A. Lipari

June 16, 2008

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Certification of Completion of ASC FY08

Level-2 Milestone ID #2933

Don Lipari
June 11, 2008

This report documents the satisfaction of the completion criteria associated with ASC FY08 Milestone ID #2933: *Deploy Moab resource management services on BlueGene/L*. Specifically:

> This milestone represents LLNL efforts to enhance both SLURM and Moab to extend Moab's capabilities to schedule and manage BlueGene/L, and increases portability of user scripts between ASC systems.

The completion criteria for the milestone are the following:

1. Batch jobs can be specified, submitted to Moab, scheduled and run on the BlueGene/L system.
2. Moab will be able to support the markedly increased scale in node count as well as the wiring geometry that is unique to BlueGene/L.
3. Moab will also prepare and report statistics of job CPU usage just as it does for the current systems it supports.

This document presents the completion evidence for both of the stated milestone certification methods:

> Completion evidence for this milestone will be in the form of (1) documentation - a report that certifies that the completion criteria have been met; and (2) user hand-off.

The sections below present the evidence of satisfaction of the completion criteria.

General Description of the Effort

As the selected Tri-Lab workload manager, Moab was chosen to replace LCRM as the enterprise-wide scheduler across Livermore Computing (LC) systems. While LCRM / SLURM successfully scheduled jobs on BG/L, the effort to replace LCRM with Moab on BG/L represented a significant challenge.

Moab is a commercial product developed and sold by Cluster Resources, Inc. (CRI). Moab receives the users' batch job requests and dispatches these jobs to run on a specific cluster. SLURM is an open-source resource manager whose development is managed by members of the Integrated Computational Resource Management Group (ICRMG) within the Services and Development Division at LLNL. SLURM is responsible for launching and running jobs on an individual cluster.

Replacing LCRM with Moab on BG/L required substantial changes to both Moab and SLURM. While the ICRMG could directly manage the SLURM development effort, the work to enhance Moab had to be done by Moab's vendor. Members of the ICRMG held many meetings with CRI developers to develop the design and specify the requirements for what Moab needed to do.

Extensions to SLURM are used to run jobs on the BlueGene/L architecture. These extensions support the three dimensional network topology unique to BG/L. While BG/L geometry support was already in SLURM, enhancements were needed to provide backfill capability and answer "will-run" queries from Moab.

For its part, the Moab architecture needed to be modified to interact with SLURM in a more coordinated way. It needed enhancements to support SLURM's shorthand notation for representing thousands of compute nodes and report this information using Moab's existing status commands. The LCRM wrapper scripts that emulated LCRM commands also needed to be enhanced to support BG/L usage.

The effort was successful as Moab 5.2.2 and SLURM 1.3 was installed on the 106496 node BG/L machine on May 21, 2008, and turned over to the users to run production.

Completion Criteria

1. Specifying, submitting, and scheduling batch jobs on BlueGene/L.

Moab was modified to accommodate a BG/L job specification. Moab submits these jobs to SLURM which launches and runs the job.

2a. Moab scales to the much larger node count

Moab's msub submission command accepts a large node count. For convenience, it accepts the "k" shorthand to represent 1024 nodes. Native Moab status commands now present node counts in "k" values.

2b. Moab / SLURM support for BG/L's unique, 3D network wiring geometry

Support for the complex network wiring geometry was relegated to SLURM. This is a change for Moab in that is does not schedule individual nodes. Instead, it wraps job requests into "will-run" queries that SLURM evaluates and answers.

Moab follows its scheduling policy and submits jobs to SLURM in priority order and lets SLURM set up the required network geometry.

3. Moab reports the same job statistics on CPU usage that LCRM did.

The script used to retrieve job statistics from Moab on non-BG/L machines was used to retrieve job statistics on BG/L. This script is run periodically on BG/L and feeds job reports to the lrmusage database. Reports of BG/L CPU usage can be retrieved using the standard lrmusage command.

User Notes for Running Jobs on BG/L Under Moab / SLURM

General information on running jobs on LC machines under Moab can be found online at:
https://computing.llnl.gov/jobs/moab/

General information on running jobs on BlueGene/L can be found at:
https://asc.llnl.gov/computing_resources/bluegenel/basics/

SLURM information specific to the BlueGene/L machine can be found at:
https://computing.llnl.gov/linux/slurm/bluegene.html

Moab and SLURM 1.3 have been installed on the BlueGene machines. The six scripts that emulate LCRM commands on Moab scheduled machines in LC are available to BlueGene/L users. There are only a few differences in submitting jobs to Moab to run on the large BG/L machine, bgl, or bgldev compared to the other LC machines.

1. Use the `msub -l nodes=<value>` option to specify the number of BlueGene compute nodes (c-nodes).

   Example: `msub -l nodes=1024` …

2. As with LCRM, “k” is recognized as shorthand for 1024.

   Example: `msub -l nodes=2k`

3. Moab/SLURM 1.3 impose a tighter restriction than LCRM/SLURM 1.2 when scheduling nodes on BGL. The job’s requested nodes must exactly match an available block in the (default or specified) SLURM partition. LCRM/SLURM 1.2 rounded up to the next available block size. To see block and partition information, invoke `smap -Db -c`. Use the `msub -q <partition>` option to target your job to a specific SLURM partition.
4. On bgldev, you can request a portion of the mid-plane. You should specify a node count supported by the hardware (32, 128, 512, or 1024 for bgldev).
5. Use the new `msub --slurm` option to pass BGL-specific options to SLURM. The `--slurm` option **must** be the last msub option on the command line. All options after that are passed directly to SLURM’s `sbatch` command.

Example: `msub <program> -l nodes=512 --slurm --no-rotate`

6. We have configured BGL’s pdebug partition to allow job submissions directly to SLURM using the `sbatch` command. This is an alternative to submitting jobs via msub (and psub) requesting the pdebug class (i.e., SLURM’s debug partition). Jobs submitted directly to SLURM will be invisible to Moab. This practice allows for faster job initiation in the pdebug partition than Moab can provide and conforms to the policy in place on other LC machines with pdebug partitions.

7. The psub wrapper exists on BGL and bgldev, but its use is deprecated. The following options are not supported:

    a. The `psub -ln` k shorthand notation (to represent 1024 compute nodes) is not supported. Either specify the node count in numeric form only or use the msub command described in item 2 above.

    b. The `psub -bgl <attributes>` option is not supported. See item 5 above for the recommended alternative.

User Validations

The following letters are from key scientists who have run their jobs through Moab on the production BlueGene/L machine and attest to Moab / SLURM’s readiness.

TO: LLNL ASC Office

FROM: David Richards

SUBJECT: User Validation of LLNL ASC Level 2 Milestone 2933

DATE: June 9, 2007

As an LLNL scientist asked to participate, I certify that the Moab / SLURM installation on BlueGene/L on May 21, 2008 provided the job specification, job scheduling and job launch capability required to run my jobs at the scale required.

Since that date, I have successfully run jobs on the bgl machine using the Moab Workload Manager. I was able to submit jobs to the job queue, view job status, and run my jobs successfully to completion. I found that Moab scheduled my jobs to run in a reasonable amount of time, comparable to the LCRM scheduler.

In addition, I exercised the LCRM emulation commands (psub and pstat) to submit my jobs to Moab and receive status of my jobs from Moab. I was also able to get usage information on my jobs via the lrmusage command.

Signed: David F. Richards on date: 9-Jun-08

TO: LLNL ASC Office

FROM: Ron Soltz, N-Div

SUBJECT: User Validation of LLNL ASC Level 2 Milestone 2933

DATE: June 9, 2007

As an LLNL scientist asked to participate, I certify that the Moab / SLURM installation on BlueGene/L on May 21, 2008 provided the job specification, job scheduling and job launch capability required to run my jobs at the scale required.

Since that date, I have successfully run jobs on the bgl machine using the Moab Workload Manager. I was able to submit jobs to the job queue, view job status, and run my jobs successfully to completion. I found that Moab scheduled my jobs to run in a reasonable amount of time, comparable to the LCRM scheduler.

In addition, I exercised the LCRM emulation commands (psub and pstat) to submit my jobs to Moab and receive status of my jobs from Moab. I was also able to get usage information on my jobs via the lrmusage command.

Signed: Ron Soltz on date: 6/11/08

Milestone (ID#2933): Deploy Moab resource management services on BlueGene/L

Level: 2

Fiscal Year: FY08

DOE Area/Campaign: ASC

Completion Date: Jun-08

ASC nWBS Subprogram: Computational Systems & Software Environment, Facility Operations & User Support

Participating Sites: LLNL

Participating Programs/Campaigns: ASC

Description: In September 2006, the Moab Workload Manager was selected to become the standard batch scheduling system for exclusive use across the tri-lab HPC facilities. Moab is a commercial product that is developed and sold by Cluster Resources, Inc. LLNL's existing batch system, LCRM, will gradually be replaced by Moab on all platforms. Moab workload management services were installed on several LLNL platforms in early FY07. The BlueGene/L system currently runs the SLURM resource manager and is scheduled by LCRM. This milestone represents LLNL efforts to enhance both SLURM and Moab to extend Moab's capabilities to schedule and manage BlueGene/L, and increases portability of user scripts between ASC systems.

Completion Criteria: This milestone is complete when batch jobs can be specified, submitted to Moab, scheduled and run on the BlueGene/L system. Moab will be able to support the markedly increased scale in node count as well as the wiring geometry that is unique to BlueGene/L. Moab will also prepare and report statistics of job CPU usage just as it does for the current systems it supports.

Customer: NNSA/ASC Headquarters

Milestone Certification Method: Completion evidence for this milestone will be in the form of (1) documentation - a report that certifies that the completion criteria have been met; and (2) user hand-off.

Supporting Resources: Tri-lab CSSE and FOUS products and personnel with the support from CRI to add the necessary enhancements to Moab.

Codes/Simulation Tools Employed: TBD

Contribution to the ASC Program: Enhances effectiveness of the infrastructure to run ASC codes on the BlueGene/L platform(s).

Contribution to Stockpile Stewardship: Supports the overall SSP goal that rely on codes that run on BlueGene/L, including UQ analyses, advanced weapons science studies, and enhanced predictive capabilities.

No.	Risk Description	Risk Assessment (low, medium, high)		
		Consequence	Likelihood	Exposure
1.	This milestone has dependencies on an external vendor (Cluster Resources) to provide topology-aware capabilities and scalability for large BlueGene/L systems.	Low	Moderate	Low