

# *Run-time operating system environment*

DOE HPC Operations Review  
San Francisco, November 5-6, 2013



**Run-time operating system environment,  
including logging, monitoring, schedulers, and  
allocations.**

## Breakout participants

- Adam Bertsch, LLNL \*
- Steve Monk, SNL \*
- Reese Baird, LANL
- Kevin Harms, ANL
- Doug Jacobsen, LBL
- Jason Hill, ORNL
- Sue Kelly, SNL
- Larry Pezzaglia, LBL
- Jay Srinivasan, LBL

\* Denotes breakout session lead

# Processes (scope of activity)

What needs to be done?

- Monitoring
  - monitoring for verification of bring up..talk to everything
  - nodediag configuration is setup to match the machine hardware configs
  - Ability to gather diag. info is important and should be specified in the RFP
- Resource manager and job scheduler configured
  - Multiple phases of allocation structure
    - early users have resource restrictions (time, node count, etc)
    - friendly users at first
- OS image build:
  - Use tools such as Cfengine, Xcat. GMI etc.
  - version control is very important!
  - nice if tools are portable to black box vendor gear example RHEL with TOSS tools running on Sequoia

## Processes (scope of activity), cont.

What begins first: timeline for activities (before or after hardware)?

- Before:
  - Model with virtual machines to help build configuration
  - obtain information required for integration (e.g. MAC addresses)
  - work with vendor to ensure support for HW/SW environment
  - process planning
- After:
  - Most of this process happens here

## Processes (scope of activity), cont.

What is the role of early hardware access (either locally or remotely) and prototype systems?

- Not just compute nodes, but other infrastructure as well
  - Switches, PDUs, etc
- Remote:
  - difficult for sys-admin activity
  - can be helpful to find issues early
- Local:
  - connectivity, physical access
  - validate hardware with OS image
    - BIOS settings, EDAC

## Processes (scope of activity), cont.

What is the role of vendor partnerships?

- Early: verify the run time OS works
- Early: verify the monitoring tools function correctly
- On-going: consistent and updatable firmware is important
- On-going: support of new versions of operating system
  - vendor supplied OS can cause some risks to moving forward
    - security risks from vendor being behind with supported version
    - moving forward takes you off of vendor supported configuration

## Processes (scope of activity), cont.

What are the roles of research and design and engineering ?

- Build your own image/modified site configurations require an on-going local design and research effort (e.g. TOSS rolling upgrades)
- Proactive monitoring and fault model identification
- Node provisioning (e.g. Open Stack)
- Tuning (hardware and software)
- working with component manufacturers

## Processes (scope of activity), cont.

What are the roles of research and design and engineering (NRE)?

- NRE:
  - Vendors don't develop these features without us funding it
    - Blue Gene Q and dynamically linked libraries
    - Lustre contracts to enable new features etc. (Lustre Center of Excellence)
    - Power monitoring and control, Burst Buffer etc.
    - OpenSFS, IB trade association can be NRE like

# Processes (scope of activity), cont.

What resiliency activities are executed (for example, redundancy)

- Repeatabile OS image and cluster configuration allows for quick cluster re-creation
  - image tracking is important!
- Monitoring: event based needs validation
  - poll model or verify that your push happens
- Hardware failure scenarios and their impact on the OS are good to know early on
  - does your redundancy work? e.g. dual fed UPS/House racks are only as resilient as a single power supply
  - keep the tests as simple as possible
  - can the monitoring system detect a loss in redundancy
- HA on resource managers and schedulers
  - single point of failure on some vendor systems

## Organization and management

What is the structure of the integration and preparation teams?

- Project Manager for large installations
  - cross functional teams underneath

## Organization and management, cont.

What are the necessary skills for the activity team?

- Skills of core admin/integration teams are typically sufficient for the scope of this effort
  - sys-admin, development environment and user services
- testing/QA skills are very important and may not be a “normal” skill of a typical sys-admin

## Experiences and lessons learned

- What were the good and bad experiences and lessons learned?
- Good:
  - Common cluster OS images (e.g. TOSS)
- Bad:
  - Cluster images provided by Hardware vendor can be problematic and not adaptable to site needs
- Lessons learned:
  - Limit changes to allocation policies
  - Sameness is a huge win!
  - Verification of monitoring tools
  - Saving monitoring data

## Experiences and lessons learned, cont.

What were the most productive activities?

- Investing in Tools development can save a bunch of future time
- Develop monitoring for previously observed conditions
  - it's an iterative process
- Monitor early, Monitor often!

## Experiences and lessons learned, cont.

What were the resiliency experiences?

- Scheduler failures fixed (ANL)
  - continued development improve resiliency over time
- “Yank” testing can expose problems in your redundant plan

## Experiences and lessons learned, cont.

What was the highest risk? Was it a surprise or expected?

- Liquid cooled machines have a whole new set of variables...surprise
- Tuning of shared library codes related to node counts...somewhat of a surprise
- Expect to execute risk mitigation on high risk areas related to scaling, such as file systems, interconnects, etc.

## Most significant observation

Provide a summary statement for the most significant observation

- configuration management with vigorous validation

## Effort estimate

How big of an effort was this?

- At least 5 staff years for a big machine
  - higher or lower