Los Alamos
NATIONAL LABORATORY
EST.1943

Delivering science and technology
to protect our nation
and promote world stability

HPC Systems Acceptance: Controlled Chaos

SC'16 - Inaugural HPC Systems Professionals Workshop

Salt Lake City, UT

Paul Peltz Jr, Parks Fields
Scalable Systems Engineer
HPC Design

11/14/2016

NNSA National Nuclear Security Administration
Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

Presentation Overview

- **The Importance of Acceptance**
- Procurement Process
- Performance and Reliability Testing
- Acceptance Phases
- System Integration
- Bug and Issue Tracking
- Conclusions and Lessons Learned

The Importance of Acceptance

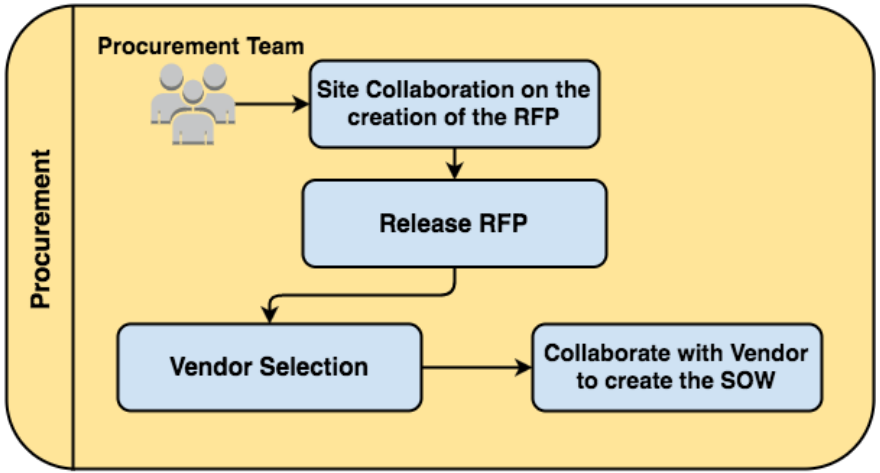
- **Acceptance is about more than the Applications**
  - Hardware
  - Software
  - Facilities
  - Monitoring
- **Testing Each of these Areas is Critical**
- **Develop an Acceptance Plan**

Presentation Overview

- The Importance of Acceptance
- Procurement Process
- Performance and Reliability Testing
- Acceptance Phases
- System Integration
- Bug and Issue Tracking
- Conclusions and Lessons Learned

Procurement Process

- **Request for Proposal (RFP)**
  - Site's solicitation for a proposal for the problem they are trying to solve
- **Vendor Selection**
  - Review Proposals
- **Creation of Statement of Work (SOW)**
  - Contract between site and vendor to obligate the vendor to provide the solution that was proposed in the RFP



Procurement Process

Statement of Work (SOW)

- **Complexity/Length of the SOW depends upon the system**
- **What we as Administrators should have in the SOW**
  - Homogeneity of HW components
    - DIMMs, Power Supplies, etc.
      - DIMM - Variable Performance, Failure Rates, Parity Failure Rates
      - PS - Inconsistent power output, Failure Rates
  - Identical part supplies for the lifetime of the system's warranty
  - Performance/Capability of components
    - DDR speed, Interconnect speed, bisection bandwidth
  - Software Provided with the system
    - Work Load Manager, compilers, debuggers
    - Vendor software complies with site security requirements

Procurement Process

Statement of Work (SOW) cont.

- Failure Rates
  - Mean Time Between Failure (MTBF)
    - Defines how long between component failures
    - Spare parts cache is sized accordingly
  - Job Mean Time to Interrupt (JMTTI)
    - Minimum time allowed between job failures
    - HW or SW event that takes down a node
  - System Mean Time Between Interrupt (SMTBI)
    - Availability of the System
    - Network Failure, PFS failure
    - SW or HW event that brings down the machine

Presentation Overview

- The Importance of Acceptance
- Procurement Process
- Performance and Reliability Testing
- Acceptance Phases
- System Integration
- Bug and Issue Tracking
- Conclusions and Lessons Learned

Performance and Reliability Testing

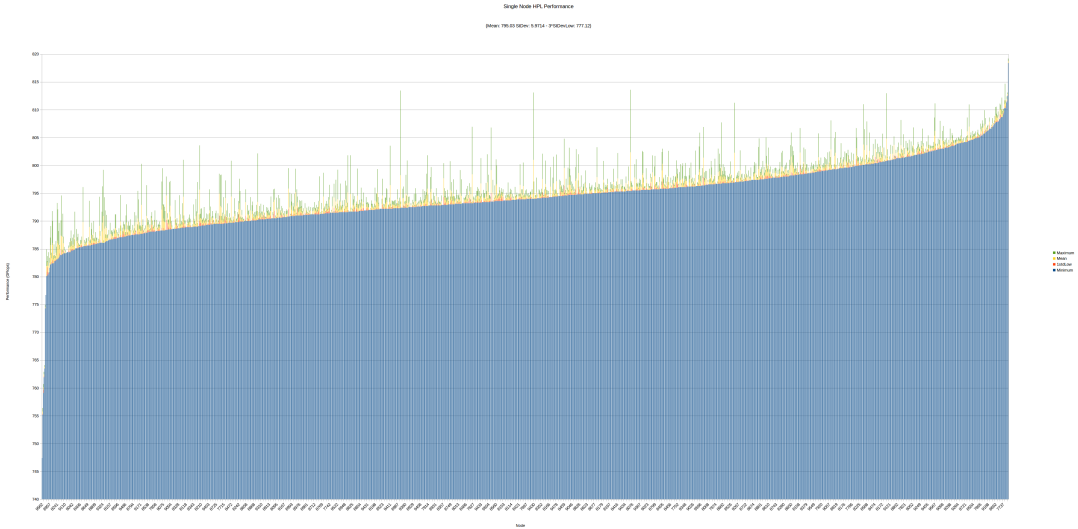
Performance

- **Synthetic Benchmarks**
  - Do not typically reflect the systems workload
  - HPL
    - FLOP/s
  - HPCG
    - Bookend for HPL
  - STREAM/STRIDE
    - Memory tester
  - Network Benchmarks
    - OSU, IMB, System Confidence

Performance and Reliability Testing

Performance (cont.)

- **HPL – More than a benchmark**
  - HW Infant Mortality
  - CPU Testing
    - Performance Variations
      - CPUs can exhibit much higher performance variations now (Anecdotal)
      - Find “under performers”
    - Correctness
      - High residual value causes the HPL Result to be invalid



Performance and Reliability Testing

Performance (cont.)

- Thermal Testing
  - Validate that system components do not exceed their thermal threshold
  - Find hot spots in the system
    - Thermal paste issues
    - Fans set in the wrong direction
- Facility Testing
  - Test to make sure the system does not exceed the high end power draw
  - Facility can adequately cool the machine

Performance and Reliability Testing

Performance (cont.)

- **Representative Applications**
  - Suite of Applications that represent the typical workload
  - Stress various aspects of the system
    - I/O intensive
    - Memory Intensive
    - CPU Intensive
    - Cache Thrashing

Performance and Reliability Testing

Reliability

- **Test System Stability**
  - Fault Injection
    - Test failures of different components of the system
    - Test HA functionality
  - Tracking Failures
    - Track job failures to verify JMTTI
    - Track system failures to verify SMTBI
    - Component Failure
      - Are components failures meeting the expected MTBF
      - If not, this could lead to lower JMTTI and/or SMTBI values
    - Ask Vendor to root cause each failure

Presentation Overview

- The Importance of Acceptance
- Procurement Process
- Performance and Reliability Testing
- Acceptance Phases
- System Integration
- Bug and Issue Tracking
- Conclusions and Lessons Learned

Acceptance Phases

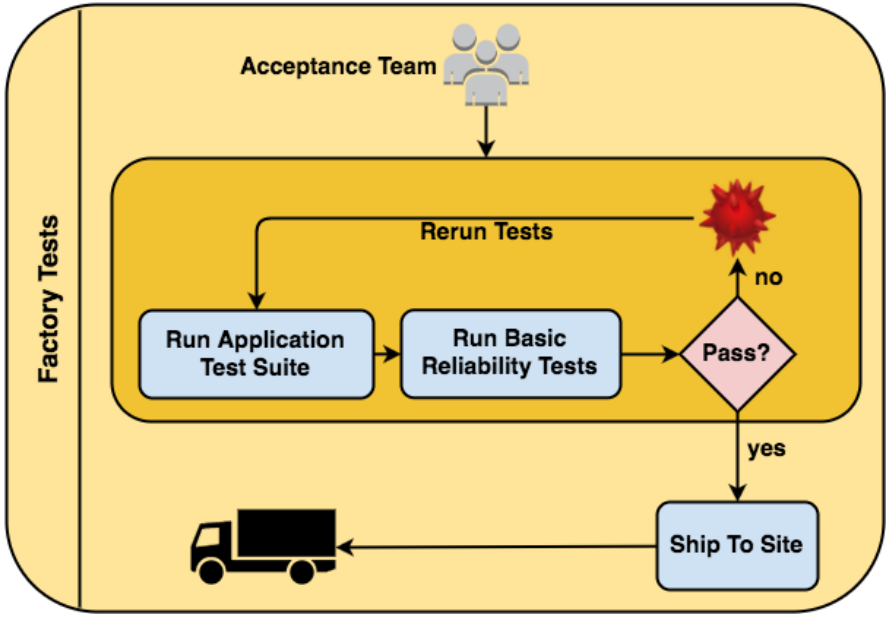
Test Harness

- **LANL uses pavilion**
  - Framework for launching tests and getting results
  - Allows site to define tests
  - Define multiple applications to run simultaneously
  - Utilizes batch scheduler to launch jobs to run continuously
  - Ability to define a Pass/Fail for the applications
  - Launch jobs and triage failures

Acceptance Phases

Factory Trial

- **Purpose**
  - Testing at vendor facility before shipment
  - Test for Systemic Hardware Issues
  - Do not test performance during this time
  - Verify hardware is fully functional
  - Usually synthetic benchmarks only
  - Verify no “forklift” replacements will have to be done on site



Acceptance Phases

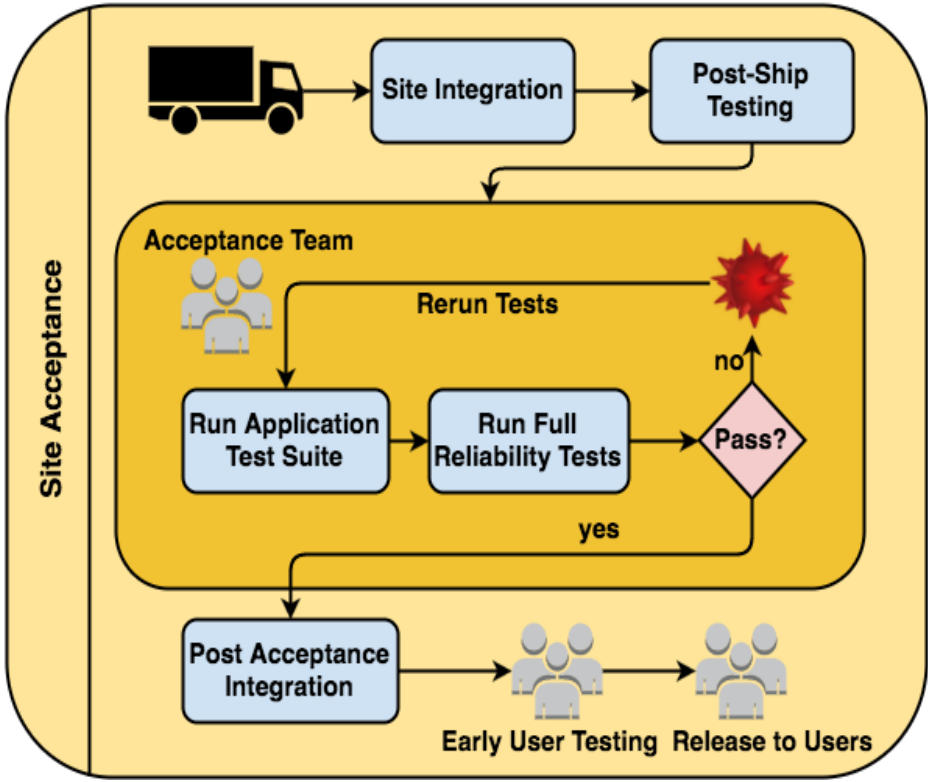
Post Shipment Tests

- **Purpose**
  - Verify there was no damage during shipment
  - Verify no problems during installation at the site
  - Rerun of the factory trial tests
  - Test if the Facility integration was successful
    - Power, Water, and Cooling

Acceptance Phases

Acceptance Testing

- **Verification that the System fulfills the SOW**
- **Application Testing**
  - Capability Improvement (CI)
    - problem-size-increase x run-time-speedup
    - Usually only for the advanced technology system (ATS)
  - Application Scaling Tests
- **Full Scale System Reliability**
  - Tracking failures to calculate JMTTI and SMTBI
  - System runs full set of applications for ~2 weeks



Acceptance Phases

Regression Testing

- **Pavilion acceptance results are saved**
  - system is tested to verify there is no degradation in performance
    - Kernel upgrades
    - Driver Upgrades
    - OS Upgrades
  - Track system degradation/improvement over time
  - Usually only on the large systems

Presentation Overview

- The Importance of Acceptance
- Procurement Process
- Performance and Reliability Testing
- Acceptance Phases
- System Integration
- Bug and Issue Tracking
- Conclusions and Lessons Learned

System Integration

- **The System is the vendors until it is accepted**
  - Especially a problem if using vendor software
  - Tracking changes and configuration settings the vendor makes to the system
    - Typically the system is tuned/configured to pass acceptance
    - Not always ideal for production
  - LANL uses a combination of a version control system and configuration management to track changes on the system

System Integration

Vendor Software

- **Test Vendor provided software**
  - Security
  - Functionality
  - Integrates into sites infrastructure
  - Fixes to bugs come in the form of an RPM
  - Monitoring and Logging

System Integration

Site Software

- **Commodity Clusters**
  - Site usually has a system provisioning solution
    - Warewulf, xcat, nfsroot
  - Testing is mostly focused on hardware testing
    - Performance
    - Reliability

Presentation Overview

- The Importance of Acceptance
- Procurement Process
- Performance and Reliability Testing
- Acceptance Phases
- System Integration
- Bug and Issue Tracking
- Conclusions and Lessons Learned

Bug and Issue Tracking

- **Large complex systems can have hundreds of bugs generated on the system during acceptance**
- **Weekly meetings with vendor to discuss bugs**
- **Vendor will never resolve all of the bugs before acceptance**
  - Milestone bugs
  - Hold vendor accountable
- **Spreadsheet to manage these bugs**

Trinity Issue Tracker

Weighting Factors	
minor	1
major	4
urgent	16
critical	64

Trinity Issue Tracker

Backlog Metric = 28849.4

Status Updates ==>

Backlog Factor	Days Opened	Days Since Last Update to LANL	LANL Escalation Priority Factor	WAIT	Issue Summary Description	LANL Issue ID	LANL POC	LANL Assigned Severity	LANL Assigned Status	Cray Issue ID	Cray POC	Cray Assigned Severity	Cray Assigned Status	Opened	Last Update to LANL	Date Closed	Product	Component	11/9/16
1507.7	231	111	0	---	Improve OpenStack / eLogin administrative password security			Critical	Open	837141		Critical	Confirmed	3/22/16	7/20/16		CLE	Install	Alpaca 1.1 with UP00 (4/5). Alpaca 1.2 will have X509 Certs (11/2016). (confirmed 3/24. ETA requested. bte)
1060.9	343	36	0	---	SNX11138 zabbix is logging at a much higher rate than SNX11139			Critical	Open	833985		Critical	Confirmed	12/1/15	10/3/16		Sonexion_HW	MMU	Workaround in place. Long term fix needed.
793.9	145	53	0	---	A sequence of large DataWarp jobs on June 15 experienced poor performance and a timeout.			Critical	Open	840948		Critical	Confirmed	6/16/16	9/16/16		CLE	Datawarp	related to 844147, 844158. bte
685.4	287	8	0	---	passwdless root access to/from all nodes		Paul Peltz	Critical	Open	835608	Jeff Becklehimer	Critical	Resolved	1/26/16	10/31/16		CLE	IMPS	RFE IMPS-4360. Still need the ability to disable key generation.
630.9	60	55	0	---	Fatal error in MPI_Init			Critical	Open	843890		Critical	Resolved-fixed	9/9/16	9/14/16		PE MPT	MPI	Waiting for patchset from Cray.
609.5	140	34	0	---	KNL re-provisioning needs to be reliable / resilient			Critical	Open	841094		Critical	Confirmed	6/21/16	10/5/16		Moab	Moab	Bug waiting for ACES to retest on tr2. bte
601.6	252	7	0	---	Live updates on elogins			Critical	Open	836758		Critical	Resolved-fixed	3/1/16	11/1/16		elogin	Build/Install/Configure	Workaround in place, need patch to address it permanently.
562.6	88	41	0	---	NHC's datawarp.sh marks computes as admindown			Critical	Open	842922		Critical	Confirmed	8/12/16	9/28/16		Moab	Moab	duplicate of 834785. bte
439.5	56	35	0	---	SNX11138 zabbix db issues			Critical	Open	843956		Critical	Confirmed	9/13/16	10/4/16		Sonexion_SW	Linux	Workaround in place. Long term fix needed.
436.7	119	20	0	---	M/T starts up compute job even if 'dw_wlm_cli -f pre_run' fails			Critical	Open	841850		Critical	Resolved	7/12/16	10/19/16		Moab	Moab	PS54 Received and LANL needs to test.
369.4	96	18	0	---	DataWarp cleanup with 32K files: rm -rf takes 30+ minutes			Critical	Open	842620		Critical	wait-confirmed	8/4/16	10/21/16		CLE	Datawarp	Cray looking into this again.
347.7	133	7	0	---	Large quantity of KNL reboots can cause dwsd to become unresponsive			Critical	Open	841328		Critical	Confirmed	6/28/16	11/1/16		CLE	Datawarp	PS64 Received and LANL needs to test.
315.1	92	13	0	---	ft - data miscompare running libhio / DataWarp tests			Critical	Open	842732		Critical	Confirmed	8/8/16	10/26/16		CLE	Datawarp	Still an issue, and DDR marginality is not to blame.
256.0	30	21	0	---	during re-provisioning - one node fails due to bounce.			Critical	Open	844974		Critical	Confirmed	10/9/16	10/18/16		SMW	HSS Infrastructure	SMW PS27.

Active | DATAWARP | Integration Entry Criteria | Final Acceptance Requirements | Adaptive | Mutrino | Metric Tracker | closed | Outstanding RedProductionIssues | Before Trinity UP01 Upgrade | dropdown

Presentation Overview

- The Importance of Acceptance
- Procurement Process
- Performance and Reliability Testing
- Acceptance Phases
- System Integration
- Bug and Issue Tracking
- Conclusions and Lessons Learned

Conclusions and Lessons Learned

- **Difficult and Stressful Process**
  - Have a plan
  - Use SOW and Issue Tracker to negotiate progression
- **Milestone payments**
  - Not one lump sum
  - Keeps the vendor motivated to progress towards individual goals and not just final acceptance
  - Helps smaller vendors
- **When to Accept**
  - Do not try and accept at the end of a fiscal year
    - Site
    - Vendor
  - When the system is able to fulfill the site's mission

Questions?

We are Hiring!

Los Alamos
NATIONAL LABORATORY
EST.1943

Delivering science and technology
to protect our nation
and promote world stability