Post-Mortem of the NERSC Franklin XT Upgrade to CLE 2.1

James M. Craw, **Nicholas P. Cardo, Yun (Helen) He**
Lawrence Berkeley National Laboratory
Berkeley, CA
craw@nersc.gov, cardo@nersc.gov, yhe@lbl.gov

And

Janet M. Lebens
Cray, Inc.
jml@cray.com

May 4, 2009
Atlanta CUG

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Introduction

This presentation will discuss the lessons learned of the events leading up to the production deployment of CLE 2.1 and the post install issues experienced in upgrading NERSC's XT4™ system called Franklin

NERSC

- NERSC is a Production Computing Facility for DOE Office of Science
- NERSC serves a large scientific population
  - Approximately 3,000 users,
  - 400 projects,
  - 500 code instances
- Focus is high end computing services

NERSC-5 Systems

Franklin (NERSC-5): Cray XT4 installed in 2007

- 9,680 compute nodes; 19,360 cores
- ~ (100 Tflops/s peak)
- 16 Login, 28 I/O Server Nodes (4 MDS Nodes)
- 2 Boot, 2 syslog, 4 network

Silence upgraded to Quad-Core in summer 2008

- 68 compute nodes; 272 cores
- 2 login, 4 I/O, 4 DVS
- 1 Boot, 1 syslog, 2 network

Gulfstream (partition of Franklin) to "burn-in" upgraded Quad-Core H/W

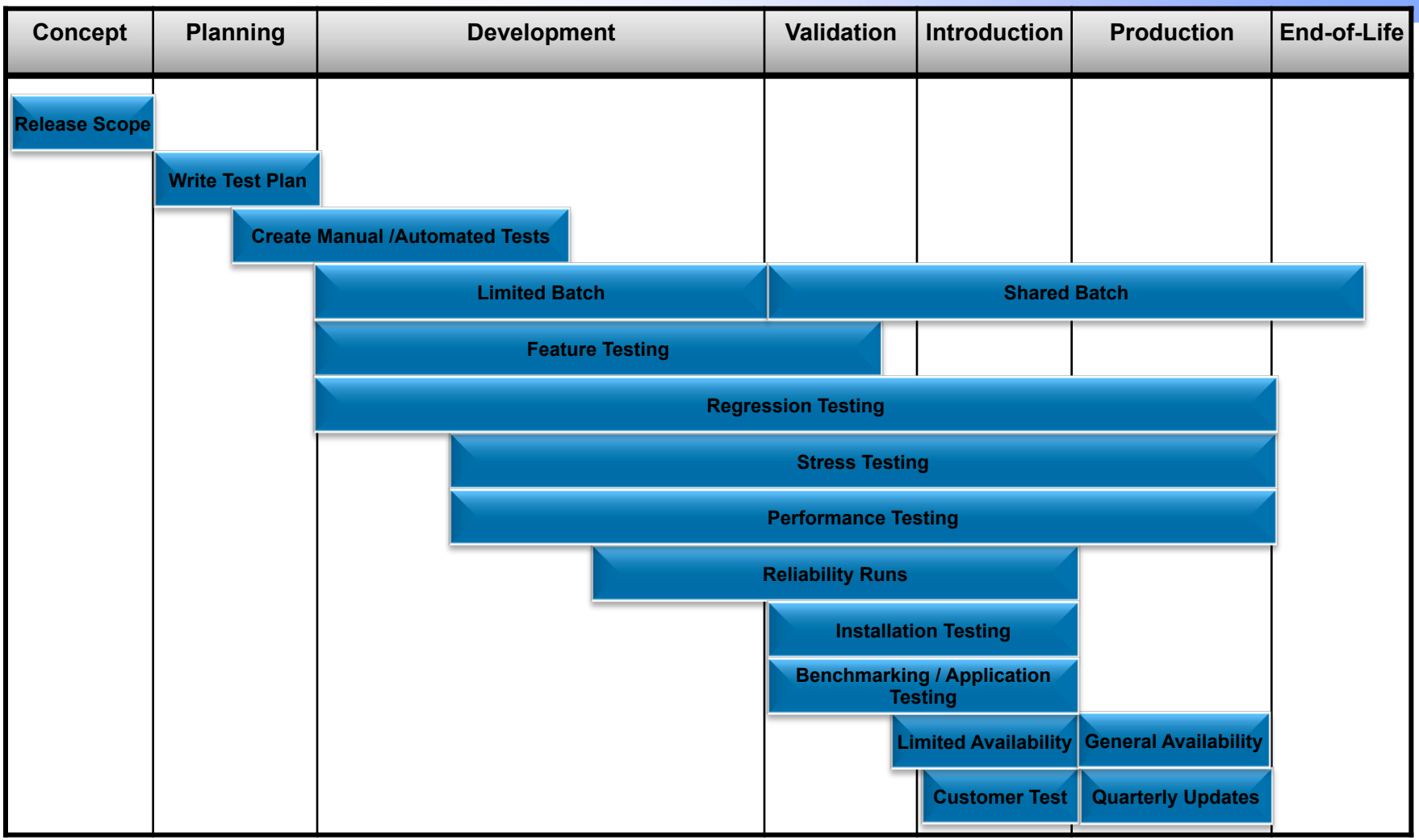
- maximum size of 48 cabinets, at largest stage, max 18,432 cores
- 2 login, 4 I/O, 4 DVS
- 1 Boot, 1 syslog, 2 network

Franklin Quad-Core upgrade completed in October 2008

- 9,592 nodes; 38,368 cores
- ~ (355 Tflops/s peak)
- 16 Login, 56 I/O Server Nodes (4 MDS Nodes)
- 20 DVS, 2 Boot, 2 syslog, 4 network

Cray's Test Strategy

Cray Product Life Cycle and Test Participation



Cray System Test Components (Suites)

- **OS:** system calls, commands, OS features
- **Interconnect:** portals, Seastar, inter-node communication
- **MPI:** MPI based applications/test codes
- **SHMEM:** shmem based applications/test codes
- **UPC:** UPC based applications/test codes
- **CUST:** 22 current customer application codes (6-18 months)
- **Application:** over 500 older applications which have found problems
- **PERF:** specific performance measures for system
- **IO:** exercise IO/networking capabilities and the file system
- **ALPS**

Cray Use of Test Suites

- **Regression tests:**
  - **All automated suites run weekly; manual tests also run**
  - **Results are checked for Pass/Fail**
- **Stress tests:**
  - **All suites run concurrently to put a heavy load on the system for four to six hours**
  - **Focus is on how the system holds up instead of individual Pass/Fail**
- **Reliability runs:**
  - **Weekly, run system for 72 hours straight under heavy load**
  - **Goal of no overall system failures, no nodes lost**

Note: all testing performed with released versions of 3rd party software (e.g. MOAB/TORQUE, PBS Pro) supported by Cray and documented in the Release Overview.

Other Cray Important Testing

- **Installation Testing – upgrade and initial install testing**
  - **Software group testing**
  - **Service group testing**
  - **Use draft installation documentation and provide feedback**
- **Benchmarks/Applications**
  - **Run customer applications for correctness and performance**
  - **Use Cray Programming Environment and provide feedback**
- **Performance Testing**
  - **Specific automated performance tests are run to measure: node-to-node throughput, ping-pong, multi-pong, all-to-all, HPCC latency, 8 node barrier times**
  - **Suites: HPCC 1.2.0, IMB, Pallas, Comtest (Sandia), memory usage-service and compute nodes, Lustre read/write**

Cray Customer Test Program Goals

Partner with 1-2 customers to obtain additional exposure and testing for upcoming feature releases

Benefits:

- Customers will be able to find problems that Cray would not experience otherwise: scaling, production workload, specific customer testing of some features
- Prove the release is stable at scale by testing in three stages:
  - Dedicated time Cray testing (features at scale, overall system at large scale)
  - Dedicated time "friendly user" application testing
  - Run solidly in production at customer site
- Gives Cray the opportunity to fix these problems before most customers upgrade to GA
- Several weeks in duration; problem reporting via Crayport/ Bugzilla

Gulfstream Test Schedule

Central Time	Pacific Time	Friday 9/26/08	Saturday 9/27/08	Sunday 9/28/08	Monday 9/29/08	Tuesday 9/30/08	Wednesday 10/1/08	Thursday 10/2/08	Friday 10/3/08	Saturday 10/4/08	Sunday 10/5/08	Monday 10/6/08	Tuesday 10/7/08	Wednesday 10/8/08
2am-6am	12am-4am					Apps Apps Apps Apps	Friendly Users & Workload	Perf Perf Perf Memtest	Friendly Users & Workload	Friendly Users & Workload	Friendly Users & Workload	Friendly Users & Workload	Friendly Users & Workload	Friendly Users & Workload
6am-10am	4am-8am		HW Update	HW Update	HW Update	Apps Apps I/O Functional DVS & NFS	Friendly Users & Workload	Memtest Memtest HW Update HW Update	Friendly Users & Workload	Friendly Users & Workload Hw/Sw Update	Friendly Users & Workload	Friendly Users & Workload HW Update	Friendly Users & Workload	Friendly Users & Workload HW Update
10am-2pm	8am-12pm	Install 2.1.36 Install PE Setup DVS, CSA, & CPR	24hr RunLong	Last chance for 24hr RunLong HW stability test	HW Status NERSC Security Scan	HW/SW Status NERSC Workload Checkout	HW/SW Status Friendly Users & Workload	HW/SW Status Friendly Users & Workload	Friendly Users & Workload & Cray Apps	DVS & NFS DVS DVS DVS	Friendly Users & Workload	Stress HSN Stress Stress Stress	CPR Functional CPR CPR CPR Demo	24hr RunLong
2pm-6pm	12pm-4pm	Install PE Quick checkout Security Scan		HW Status	Queue issue MemTest MemTest Platinum	Friendly Users & Workload	Friendly Users & Workload	Friendly Users & Workload	Friendly Users & Workload & Cray Apps	Stress Scale Stress Stress Stress	Friendly Users & Workload	DVS & GPFS DVS DVS DVS	CPR CPR & NSFv4 CPR CPR	
6pm-10pm	4pm-8pm	24hr runlong			Platinum Platinum Apps Scale Apps	CPR Functional CPR CPR CPR	DVS & GPFS DVS DVS DVS	Cray Apps & Friendly Users	Cray Apps & Friendly Users	Stress Stress Stress Stress	Friendly Users & Workload	Friendly Users & Workload	Friendly Users & Workload	
10pm-2am	8pm-12am				Apps Apps Apps Apps	CPR CPR CPR CPR	Perf Scale Perf Perf Perf	Cray Apps & Friendly Users	Cray Apps & Friendly Users	Apps Run Apps Apps Apps	Friendly Users & Workload	Friendly Users & Workload	Friendly Users & Workload	

Legend
Cray Hardware
Cray Software
NERSC
Shared
NERSC offsite

NERSC Test Strategy

Silence Test Strategy

- Before any software is installed on Franklin, it is installed and checked out on a single cabinet - independent test system - called Silence
- CLE 2.1 was first installed on Silence back in June 2008
- The primary testing goals for Silence was to:
  - Identify procedural issues
  - Become familiar with the upgrade process
  - Validate the new functionality achieved by the upgrade
  - Gain insight into the stability of the upgrade
  - Perform basic functionality tests
  - Perform limited performance tests

Gulfstream Test Strategy/Results

- Gulfstream, was a temporary partition of Franklin and was being used as a rolling quad-core hardware upgrade vehicle
- CLE 2.1 was first installed on Gulfstream back in July 2008
- The primary testing goals for Gulfstream was to:
  - Build on Silence testing goals particularly issues of scale
  - Gain insight into the stability of the upgrade at scale
  - Perform scale performance tests
- Test results positive; no major issues that didn't have a workaround

Franklin Post 2.1 Install

- Joint NERSC/Cray decision to proceed with Franklin 2.1 upgrade made; upgrade was preformed December 3, 2009
- Issues encountered:
  - Bad SeaStar netmask caused networking issue
  - Access control problem with pam_access.so
  - Franklin stability worsens
  - Virtual Channel 2 impact unknown and NERSC turns off
  - HSN congestion appears related to many system crashes
  - MPT 2.0 applications and libraries crashing system
- Many new patches get installed (December – March)

Light At The End of Tunnel

- In mid March, numerous patches installed to resolve SeaStar related issues and the NERSC wrapper for aprun (that blocked MPT2 compiled applications) appeared to be working
- Franklin still had a large number of individual patches installed and getting new fixes was becoming increasingly more difficult
- So the mother of all Patches Sets (UP01) is under consideration to install – NERSC takes the plunge and installs Patch Sets: PS01, PS01a, & PS02

Summary

- After nearly five months, the end result has been a significant improvement in the software stability of the system
- Even with all of the shared pain, amongst Cray and NERSC staff, and even NERSC users, regarding the 2.1 upgrade of Franklin; the eventual benefits (2.1 stability and functionality) out weighed the pain
- Many lessons were learned along the way also…

Lessons Learned Highlights

- Even when testing is going well; don't schedule a major upgrade right before a major holiday
- Because of the large number of changes incorporated in CLE 2.1, including upgrades to SuSE SLES and Sun Lustre, the release would have been better named "CLE 3.0"
- Open, two-way communications are key to the project success
- The assumption that a successful test on Gulfstream meant that CLE 2.1 was ready for NERSC production.
- Need to really run on a large "production" system (not just a set of test systems) at a customer site before officially GA'ing
- Utility was needed to identify non-compatible software (MPT)
- Customer needs ability to review all outstanding bugs before deciding to go production (GA) – first large site

Recommendations

- Add additional tests to the Cray test suite include:
  - Injection of additional HSN traffic to simulate congestion
  - 3D Torus test
  - I/O stress test, e.g. IOR test
- Increase the size of Cray's test system to better validate scaling issues., beyond the current 16 cabinet test system
- Continue joint Cray and customer Post-Mortems with future test partners
- NERSC and Cray should formally and jointly write a "Post-Mortem" document
- Cray and NERSC should have reviewed all (internal) problems previously found in testing
- Finally, Cray should allow NERSC to share all of its CLE 2.1 bugs with other interest sites

Acknowledgements

- The authors would like to thank the many Cray staff that helped with the Franklin upgrade, from pre-planning to post-mortem. Particularly the Cray On-site: Verrill Rinehart, Terence Brewer, Randall Palmer, Bill Anderson, and Steve Luzmoor; Jim Grindle, Brent Shields and the rest of the OSIO Test Group. Kevin Peterson, for excellent overall planning and as the Cray focal point
- The authors would also like to thank the NERSC staff that helped and worked long hours to make 2.1 a success on Franklin
- The NERSC authors are supported by the Director, Office of Science, Advanced Scientific Computing Research, U.S. Department of Energy under Contract No. DE-AC02-05CH11231