Compute Node Linux: Overview, Progress to Date & Roadmap

David Wallace

Cray Inc.

CRAY
THE SUPERCOMPUTER COMPANY

Abstract

- This presentation will provide an overview of Compute Node Linux (CNL) for the CRAY XT machine series. Compute Node Linux is the next generation of compute node operating systems for the scalar line of Cray systems. This presentation will discuss the current status of Compute Node Linux development including results of scaling and performance testing. At the time of CUG, Cray will have shipped limited access versions of CNL to customers. Early customer experiences will be discussed, as well as a vision of the long-term objectives for CNL.

Agenda Topics

- Overview of Objectives and Requirements
- Progress against Objectives
- Migration Planning

Overview of Objectives & Requirements

Compute Node Linux: Objectives

- Stable
  - Must be robust at scale
  - Low maintenance “out-of-the-box”
- Performance & Scaling
  - Meet or exceed Catamount functionality and performance
  - Must scale to 100,000 cores
- Functionality
  - Better Application portability
  - Support for sockets
  - OpenMP on a node, other programming models
- Flexible configuration of OS services
  - Support for mixture of Capability and Capacity environments on a single system

Compute Node Linux Requirements

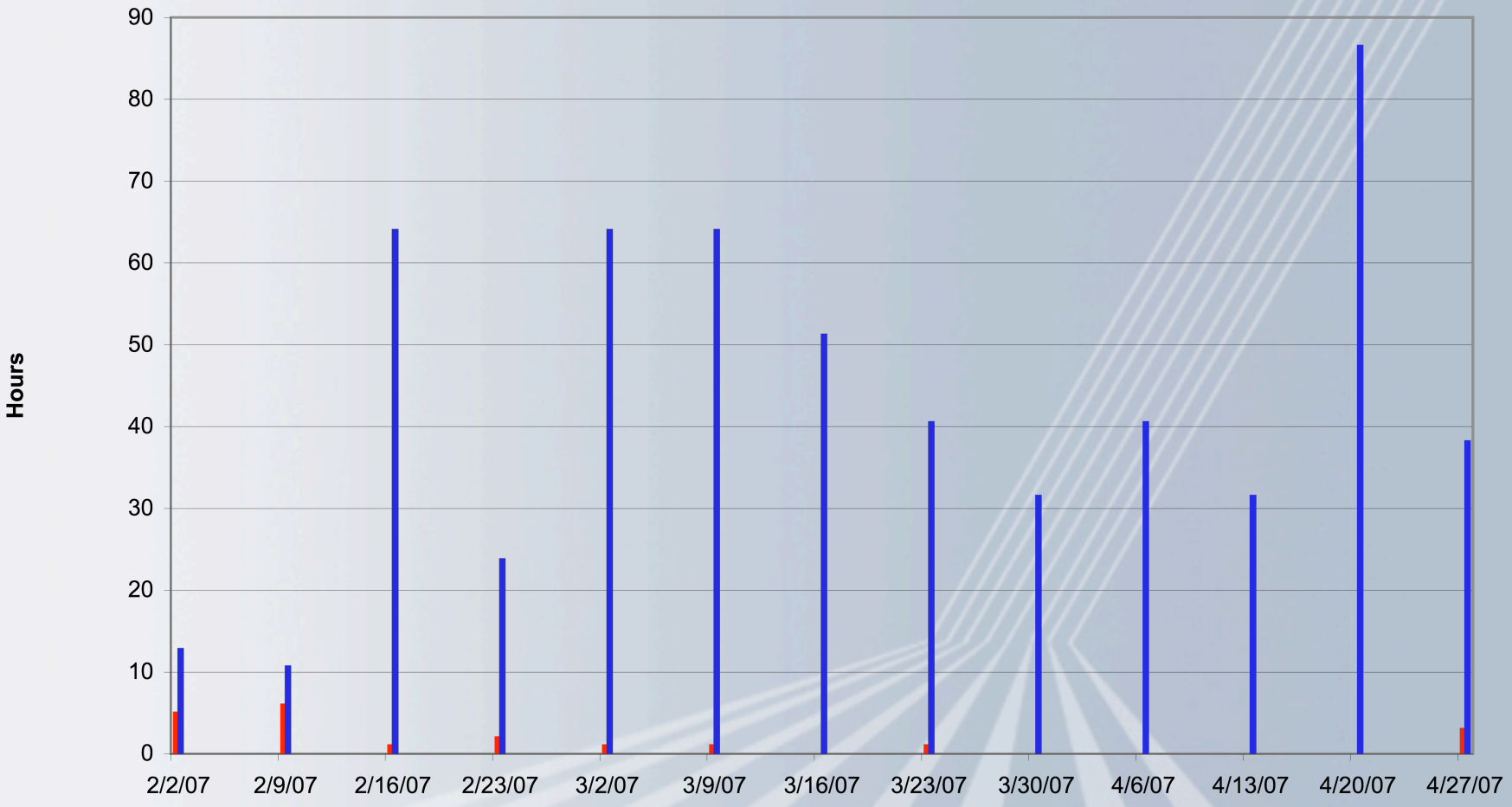
- The requirements for a compute node are based on Catamount functionality and the need to scale
  - Scaling to 20K compute sockets
  - Application I/O equivalent to Catamount
  - Start applications as fast as Catamount
  - Boot compute nodes almost as fast as Catamount
  - Small memory footprint
- Support N-way cores
- Improved application portability
- Support for multiple programming models including:
  - MPI
  - SHMEM
  - OpenMP
  - Global Arrays/ ARMCI
  - PGAS Language Support (CAF and UPC) for Baker

CNL Stability

CNL Stability

- Internal systems
  - Stability ok - frequency of system reboots is low
  - Increasing CNL time to get more exposure
  - In general, CNL can run most applications
- ORNL test time has been very helpful!
  - Have exposed scaling issues with system admin and applications
  - OS Test can crash machine under stress testing
  - Subsequent test shots will continue to focus on stability, performance and scaling, getting “friendly” user exposure

Stability Metrics



	2/2/07	2/9/07	2/16/07	2/23/07	3/2/07	3/9/07	3/16/07	3/23/07	3/30/07	4/6/07	4/13/07	4/20/07	4/27/07
Interrupts	5	6	1	2	1	1	-	1	-	-	-	-	3
MTTI	12.8	10.7	64.0	23.8	64.0	64.0	51.2	40.5	31.5	40.5	31.5	86.5	38.2

Field Exposure: AHPCRC/NCS

- AHPCRC has run key applications on Perch and Salmon
  - Performance has been in the neighborhood of Catamount
- Stability was good during Army's functional tests
- CNL test install at AHPCRC
  - Early version of XT v2.0 installed in April
  - User experience generally 'good'

Field Exposure: AHPCRC/NCS

- Applications
  - WRF: report that CNL is about 5-8% slower.
  - Gaussian doesn't run on Catamount, so CNL will look good here. :-) (about halfway through the port to CNL)
  - CPMD: is running and is starting to optimize
  - Adsorb is about 5% slower under CNL.
  - BenchC seems about 10% faster under CNL.
  - Presto: is still waiting on RSIP to get to Perch or Salmon.
- Observation
  - “the issues I had listed in our previous acceptance report appear to be resolved by going to CNL”

Performance & Scaling

CNL Scaling/Performance

- Initial Goal: Performance within 10% of Catamount (on selected applications)
  - CCSM
  - GTC
  - HPCC
  - LSMS
  - MILC
  - NAMD
  - Paratec
  - POP
  - VH1
- I/O
  - Lustre expected to perform as well under CNL as QK

Performance & Scaling Scorecard

	Relative to Catamount	
	SC	DC
CCSM		
GTC		
HPCC		
LSMS		
MILC		
NAMD		
Paratec		
POP		
VH1		
Other Apps		
LAMMPS		
GYRO		
VASP		
S3D		

Performance Tests: POP

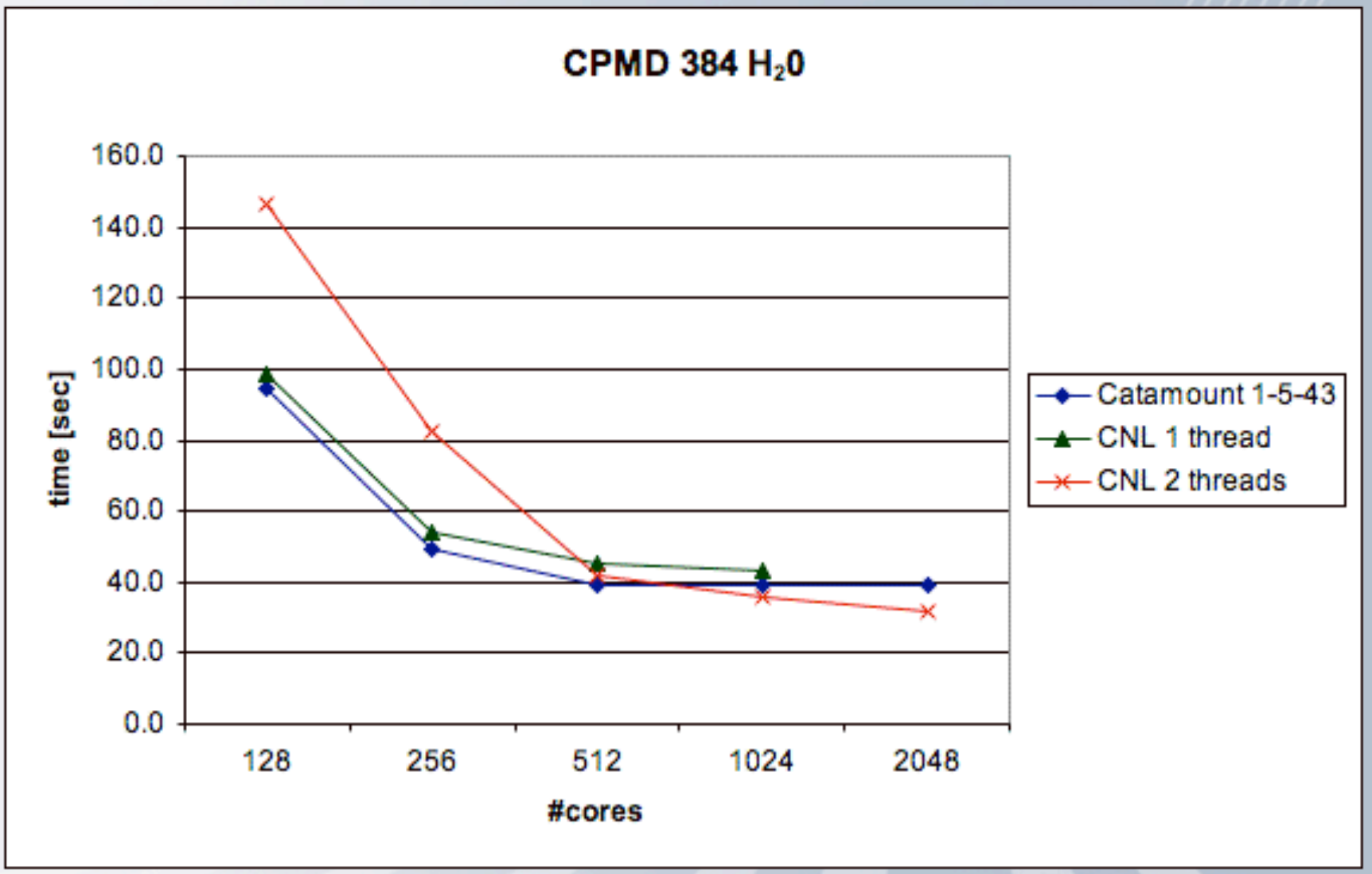
		April 17/April22 (- Slower, + Faster)		13-Apr		7-Jan	
# Cores	Routine	SC	DC	SC	DC	SC	DC
1000	Step/Total	-1%	2%	-15%	-9%	-23%	-46%
1000	Baroclinic	4%	-1%	2%	-2%	-2%	-15%
1000	Barotropic	-7%	6%	-40%	-22%	-40%	-95%
2000	Step/Total		0%	-6%	-5%		
2000	Baroclinic		-3%	0%	-1%		
2000	Barotropic		2%	-26%	-14%		
4000	Step/Total	3%					
4000	Baroclinic	1%					
4000	Barotropic	4%					
8000	Step/Total		-5%				-150%
8000	Baroclinic		-4%				-22%
8000	Barotropic		-5%				-179%

Himeno-BMT, a 3D CFD Poisson kernel benchmark code

Compute Node OS	CPU Speed	# of Cores	Performance
Catamount	2.6 GHz	256	232 MFLOPS
CNL	2.6 GHz	256	292 MFLOPS

- Program runs 1.26X faster on CNL.

OpenMP on CNL



8 processor IOR single file, stripe 1 MB, count 4, start 0

Test	Size	Operation	QK	CNL	Measure	Diff %
IOR	65536	Write	48.734	678.953	MB/sec	1293%
IOR	65536	Read	123.220	569.569	MB/sec	362%
IOR	1048576	Write	524.558	688.652	MB/sec	31%
IOR	1048576	Read	404.950	517.401	MB/sec	28%
IOR	4194304	Write	725.626	693.292	MB/sec	-4%
IOR	4194304	Read	603.823	537.106	MB/sec	-11%
mdtest	1	Dir_create	2402.801	3484.987	ops/sec	45%
mdtest	1	Dir_stat	4854.606	4887.840	ops/sec	1%
mdtest	1	Dir_rm	2000.467	3161.859	ops/sec	58%
mdtest	1	File_create	1514.006	2716.896	ops/sec	79%
mdtest	1	File_stat	4686.043	4740.014	ops/sec	1%
mdtest	1	File_rm	1658.783	2685.556	ops/sec	62%

CNL Functionality

CNL Functionality

- Meeting requirements
  - OpenMP
  - Dynamic libraries*
  - POSIX API*
- Booted and run on 11508 nodes (23016 cores)!
- Better support for ISV applications over time
  - Multi-phase project
  - This is a configuration/image management issue
  - Requirements for ISV applications vary:
    - Different libraries, different OS services, etc
    - Developing infrastructure to allow custom Application Partitions
    - Working on Prototype feature

Migrating to CNL

CNL Migration Planning

- Requires lots of planning
- Current strategy based on using System Partitioning to allow concurrent Production and CNL migration work
  - Partitioning has specific requirements
    - Separate SIO nodes
    - Separate file systems
  - Not as flexible as we would like
- Catamount/CNL differences document for Programming Environment
- When to migrate?
  - UNICOS/lc 2.0 LA release for Test systems
    - 2Q07
  - UNICOS/lc 2.0 GA release for Production
    - 1Q08

Questions?

David Wallace
dbw@cray.com

CRAY
THE SUPERCOMPUTER COMPANY