CRAY
THE SUPERCOMPUTER COMPANY

Scaling Efforts to Reach a PetaFlop

Kevin Peterson

May 6, 2009

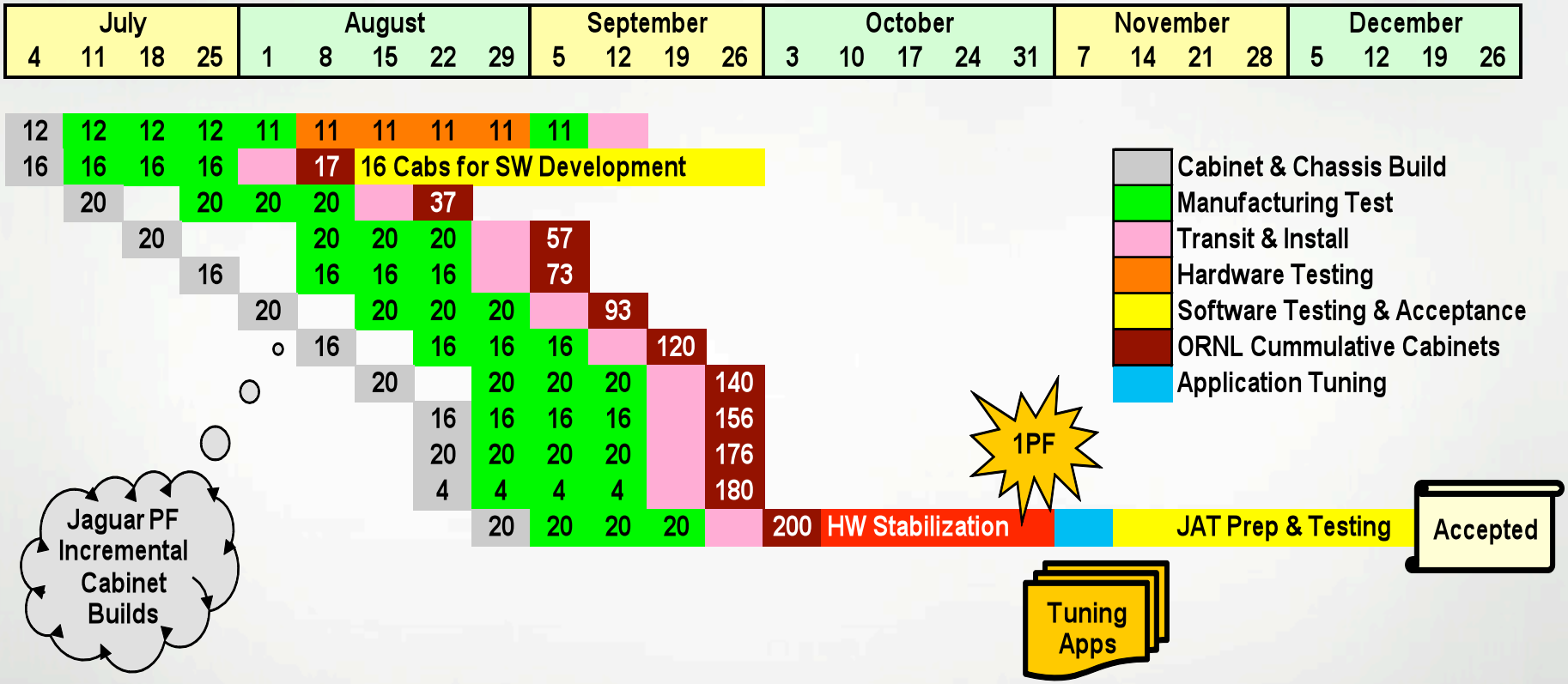
PetaScale Software Scaling Agenda

- Software Scaling Motivation & Goals
- HW Configuration & Scale Out
- Software Scaling Efforts
  - System management
  - Operating system
  - Programming environment
- Pre-Acceptance Work
  - HW stabilization & early scaling
- Acceptance Work
  - Functional, Performance, & Stability Tests
  - Application & I/O results
- Software Scaling Summary

Petascale SW Scaling Motivation & Goals

- Execute benchmarks & kernels successfully at scale on a system with at least 100,000 processor cores
- Validate Cray software stack can scale to > 100K cores
  - Cray Programming Environment scales to >150K cores
  - Cray Linux Environment scales to >18K nodes
  - Cray System Management scales to 200 cabinets
- Prepare for scaling to greater number of cores for Cascade

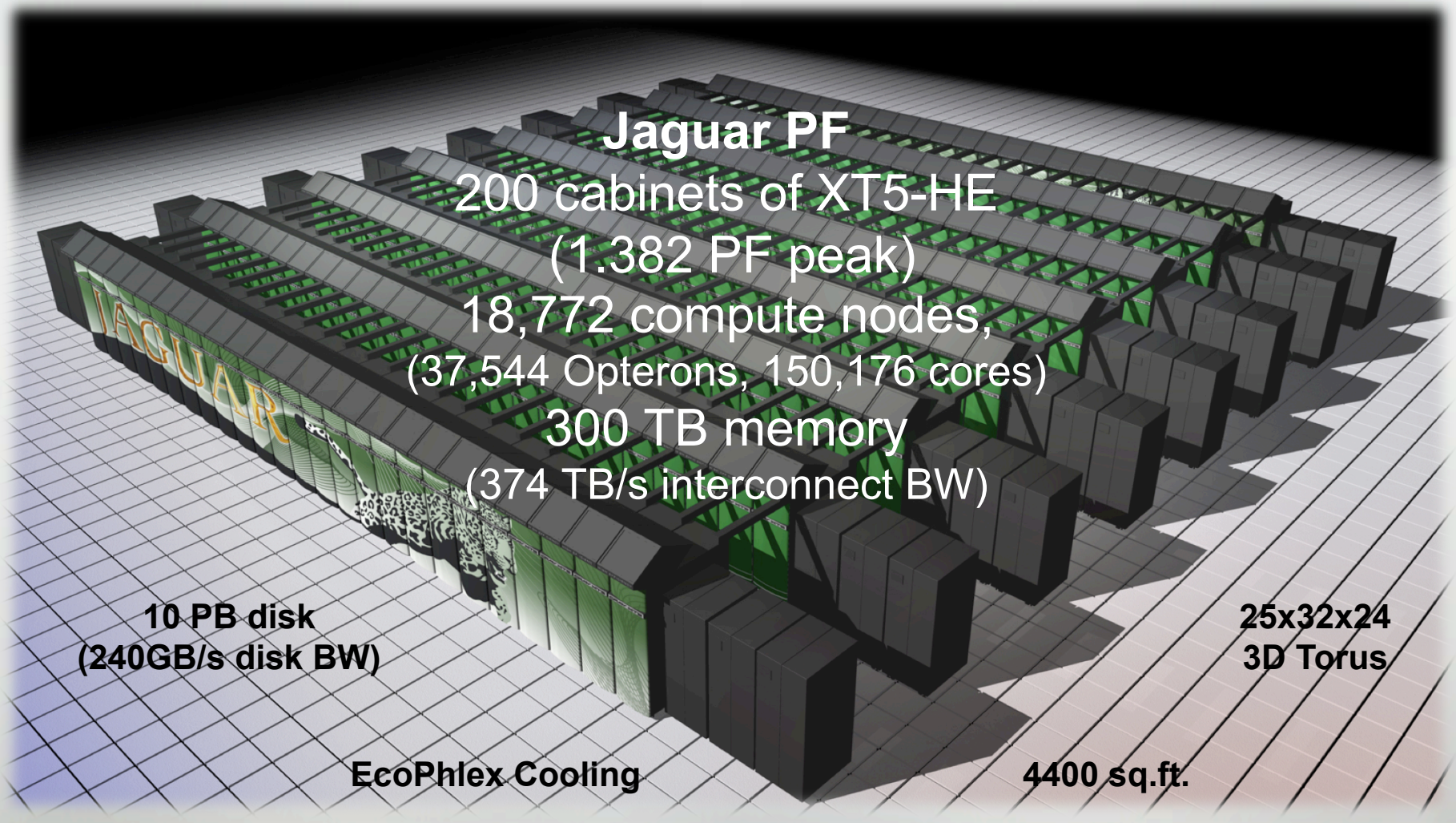
Petascale System Scaling (2008)



Only one quarter to stabilize, scale SW, tune apps, & complete acceptance!

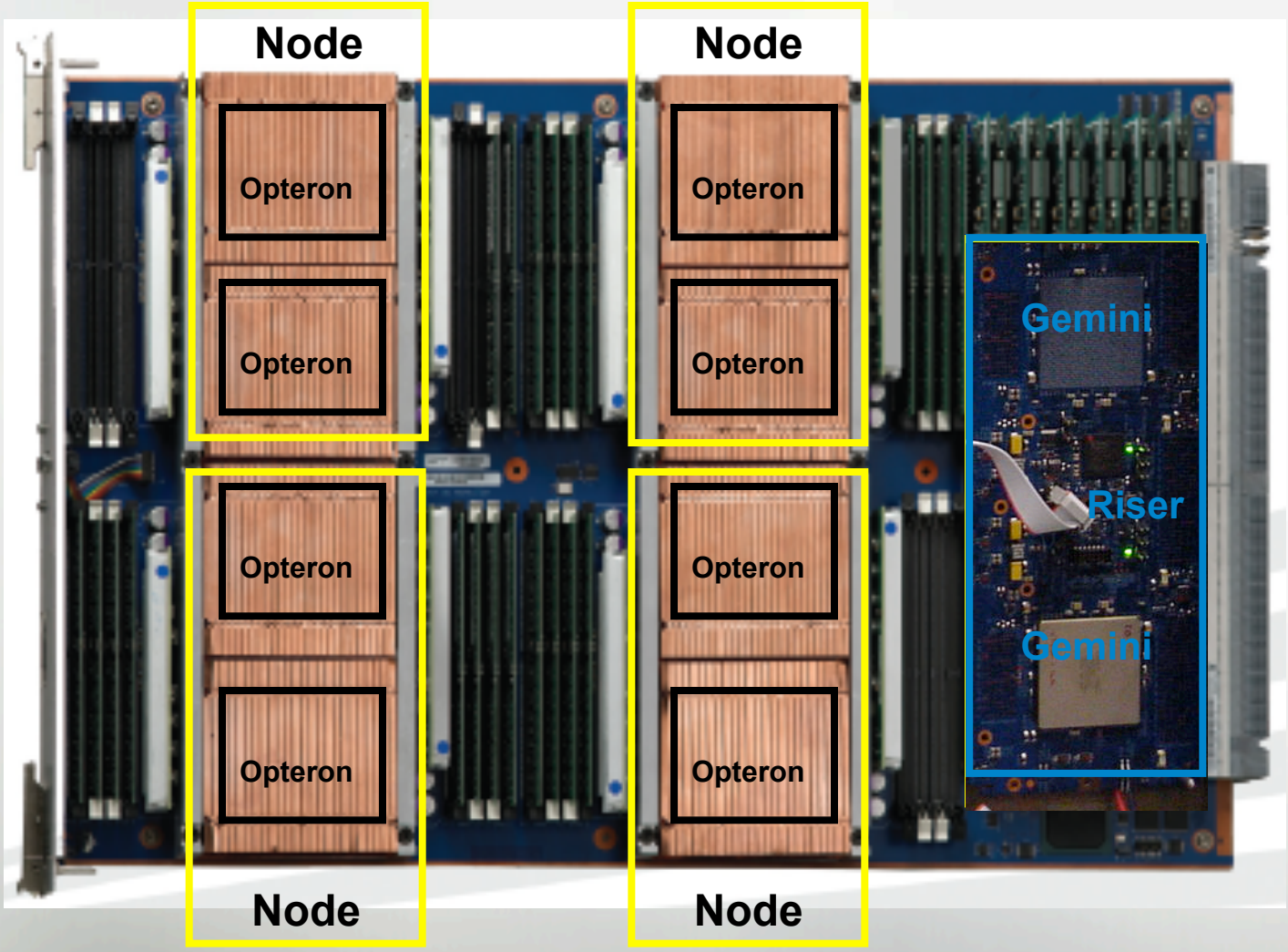
(Due in part to the solid XT foundation)

Petascale System Configuration



XT5 Blade Scaling – 32 cores/blade Today SeaStar,

…tomorrow Gemini



- Each XT5 has 4 nodes
- Each riser has 4 NICs
- Each NIC serves 2 AMD Opterons (4 cores each)

- Gemini risers will replace SeaStar risers
- Each Gemini has 2 NICs

Software Scaling Efforts - SMW

- System Management Workstation
  - Manages the system via the Hardware Supervisory System (HSS)

Hurdles & Strategies

- Single SMW for 200 cabinets
  - Localized some processing on cabinet (L1) controllers
- XT5 double density nodes with quad-core processors
  - Throttled upstream messages at blade (L0) controllers
- HSN 16K node soft limit
  - Increased limit to 32K node (max for SeaStar)

Software Scaling Efforts - CLE

- Cray Linux Environment
  - Operating system for both compute (CNL) and service nodes

Hurdles & Strategies

- Transition from Light-Weight Kernel (Catamount) to CNL
  - Reduced number of services and memory footprint
- Lack of large test system
  - Emulated larger system by under provisioning
  - Ran constraint based testing under stressful loads
- Two socket multi-core support
  - Added ALPS support for 2 socket, 4 core NUMA nodes
  - Modified Portals to handle more cores & distribute interrupts
- Switch from FibreChannel to InfiniBand (IB) for Lustre
  - Tested IB with external Lustre on system in manufacturing
  - Tested IB fabric attached Lustre on site during installation

Software Scaling Efforts - CPE

- Cray Programming Environment
  - Development suite for compilation, debug, tuning, and execution

Hurdles & Strategies

- MPI scaling >100K cores with good performance
  - Increased MPI ranks beyond 64K PE limit
  - Optimized collective operations
  - Employed shared memory ADI (Abstract Device Interface)
- SHMEM scaling >100K cores
  - Increased SHMEM PE max beyond 32K limit
- Global Array scaling >100K cores
  - Removed SHMEM from Global Array stack
  - Ported ARMCI directly to Portals
  - Tuned Portals for better out-of-band communication

Jaguar Pre-acceptance Work

- Hardware Stress & Stability Work
  - Incremental testing as system physically scaled
  - Key diagnostics and stress tests (IMB, HPL, S3D)
- HPL & Autotuning
  - Tiling across system while weeding out weak memory
  - Monitoring performance and power
  - Tuning HPL to run within the MTBF window
- Scientific Application Tuning
  - MPT (Message Passing Toolkit) restructuring for 150K ranks
  - Global Arrays restructuring for 150K PEs

High Performance LINPACK Benchmark$_2$

- **1.059 PetaFlops** (76.7% of peak)
- Ran on **150,152 cores**
- Completed only 41 days after delivery of system

```
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WR03R3C1     4712799   200   274   548          65884.80              1.059e+06
--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV--VVV-
Max aggregated wall time rfact . . . :             13.67
+ Max aggregated wall time pfact . . :             10.99
+ Max aggregated wall time mxswp . . :             10.84
Max aggregated wall time pbcast  . . :           6131.91
Max aggregated wall time update  . . :          63744.72
+ Max aggregated wall time laswp . . :           7431.52
Max aggregated wall time up tr sv  . :             16.98
--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=        0.0006162 ..... PASSED
================================================================================
```

HPC Challenge Benchmarks$_2$

- Four "Class 1" benchmarks after little tuning:
  - HPL 902 TFLOPS #1
  - G-Streams 330 #1
  - G-Random Access 16.6 GUPS #1
  - G-FFTE 2773 #3
- Still headroom for further software optimization

- These HPCC results demonstrate
  balance,
  high-performance,
  & Petascale!

Early Science Application Runs$_2$

Science Area	Code	Contact	Cores	% of Peak	Total Perf	Noteable
Materials	DCA++	Schulthess	150,144	97%	1.3 PF*	Gordon Bell Winner
Materials	LSMS/WL	ORNL	149,580	76.40%	1.05 PF	64 bit
Seismology	SPECFEM3D	UCSD	149,784	12.60%	165 TF	Gordon Bell Finalist
Weather	WRF	Michalakes	150,000	5.60%	50 TF	Size of Data
Climate	POP	Jones	18,000		20 sim yrs/ CPU day	Size of Data
Combustion	S3D	Chen	144,000	6.00%	83 TF	
Fusion	GTC	PPPL	102,000		20 billion Particles / sec	Code Limit
Materials	LS3DF	Lin-Wang Wang	147,456	32%	442 TF	Gordon Bell Winner

These applications were ported, tuned, and run successfully, only 1 week after the system was available to users!

System Scaling Criteria – The JAT_1

- Jaguar Acceptance Test (JAT)
  - Defined acceptance criteria for the system
- HW Acceptance Test
  - Diagnostics run in stages as chunks of the system arrived
  - Completed once all 200 cabinets were fully integrated
- Functionality Test
  - 12 hour basic operational tests
  - Reboots, Lustre files system
- Performance Test
  - 12 hour of basic application runs
  - Tested both applications and I/O
- Stability Test
  - 168 hour production-like environment
  - Applications run over variety of data sizes and number of PEs

Performance Test – I/O

Metric	Description	Goal	Actual
InfiniBand Performance	Send BW Test	1.25 GB/sec	1.54 GB/sec
Aggregate Bandwidth	Sequential Write Sequential Read	100 GB/sec	173 GB/sec 112 GB/sec
	Parallel Write Parallel Read	100 GB/sec	165 GB/sec 123 GB/sec
	Flash I/O	8.5 GB/sec	12.71GB/sec

PetaScale SW Scaling Summary

- Execute benchmarks & kernels successfully at scale on a system with at least 100,000 processor cores
  - Cray Linux Environment scaled to >18K nodes
  - Cray Programming Environment scaled to >150K PEs
  - Cray System Management scaled to 200 cabinets
- Demonstrated productivity
  - Performance: greater than 1 PetaFlop
  - Programmability: MPI, Global Arrays, and OpenMP
  - Portability: variety of “real” science apps ported in 1 week
  - Robustness: Completed Jaguar Stability Test

Acknowledgements & References

1. *NLCF Acceptance Test Plans (50T, 100T, 250T, 1000T-CS) and (1000T-G)*
   - DOE Leadership Computing Facility
   - Center for Computational Sciences
   - Computing and Computational Sciences Directorate
   - December 10, 2008
2. *Jaguar & Kraken – The world's most powerful computing complex* (Presentation)
   - Arthur S. (Buddy) Bland
   - Leadership Computing Facility Project Director
     National Center for Computational Sciences
   - November 20, 2008
3. *ORNL 1PF Acceptance Peer Review* (Presentation)
   - ORNL Leadership Computing Facility
   - Center for Computational Sciences
   - December 29, 2008
4. *Acceptance Status* (Presentation)
   - Ricky A. Kendall
   - Scientific Computing
   - National Center for Computational Sciences
   - October 30, 2008
5. SC08 Awards Website
   - http://sc08.supercomputing.org/html/AwardsPresented.html
   - November 21, 2008
6. Cray XT Manufacturing Plan
   - William Childs
   - Cray Inc., Chippewa Falls, Wisconsin
   - October 2008

Thank you!

kpeterso@cray.com


CRAY
THE SUPERCOMPUTER COMPANY