Update on Cray Earth Sciences Segment Activities and Roadmap

31 Oct 2006

12th ECMWF Workshop on Use of HPC in Meteorology

Per Nyberg

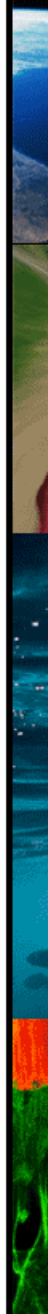
Director, Marketing and Business Development

Earth Sciences Segment

nyberg@cray.com

Topics

- Cray Update
  - Highlights of the Past Year
- Earth Sciences Segment Activities
- Increasing System Scale
- Roadmap Update



Highlights of the Past Year

- Spain's National Institute of Meteorology (INM) Begins Production Weather Prediction on Cray X1E - June 23, 2005.

- Pittsburgh Unveils Big Ben the Supercomputer - July 20, 2005:
  - 10 Tflops XT3
  - Protein simulations, storm forecasting, global climate modeling, earthquake simulations.

- ERDC Supercomputer Most Powerful In DoD - Dec 8, 2005:
  - 24 Tflops XT3
  - Sea-ice and storm surge modeling

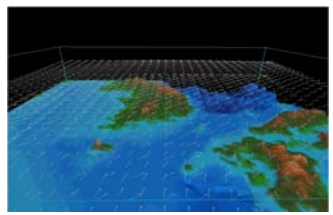
ERDC
US Army Engineer Research and Development Center

Highlights of the Past Year

- Cray Selected by UK'S AWE - Jan 24, 2006:
  - 41 Tflops Cray XT3 selected by United Kingdom's Atomic Weapons Establishment.
  - Sustained performance 30 Times the existing Blue Oak System.

- Korea Meteorological Administration's New Cray X1E Supercomputer Is World's Fastest Weather Prediction System - Feb 06, 2006
  - 18.5 Tflops X1E.

기 상 청
Korea Meteorological Administration



- Cray X1E Supercomputer at India's NCMRWF Will Be the Most Powerful Weather Prediction System in the Region - Mar 13, 2006.

राष्ट्रीय मध्यम अवधि मौसम पूर्वानुमान केन्द्र
National Centre for Medium Range Weather Forecasting (NCMRWF)
Department of Science & Technology

Highlights of the Past Year

- Swiss National Supercomputing Center Will Expand Cray XT3 System - May 17, 2006:
  - Upgrade Will Boost Computing Power to 8.6 Teraflops.
  - Climate and weather is 2nd largest user community: ECHAM5, CCSM, LM.

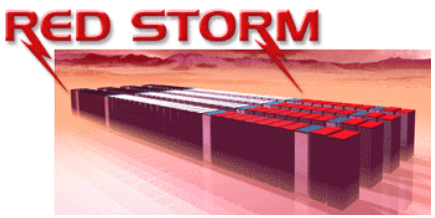
- Cray Signs $200 Million Contract to Deliver World's Largest Supercomputer to Oak Ridge - June 15, 2006:
  - Multi-year contract with the U.S. Department of Energy's (DOE) Oak Ridge National Laboratory (ORNL) to provide the world's first petaflops-speed supercomputer.
  - DOE-NSF Climate-Science Computational End Station established at ORNL.

Highlights of the Past Year

- UK'S EPSRC Selects Cray Inc. to Negotiate Multi-Year Contract for HECToR Procurement - June 27, 2006:
  - EPSRC is the main funding agency in the UK for research in engineering and the physical sciences.
  - Next generation national high performance computing service for the UK academic community with an initial theoretical peak capability of over 50 Tflops.

EPSRC Engineering and Physical Sciences Research Council

- Sandia Red Storm Moves to 120 Tflops.



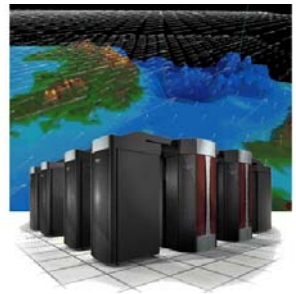
Highlights of the Past Year

- Cray Wins $52M Contract with NERSC - Aug 10, 2006:
  - Contract to install next-generation supercomputer at the DOE's National Energy Research Scientific Computing Center (NERSC).
  - The Hood system installed at NERSC will deliver sustained performance of at least 16 Tflops with a theoretical peak speed of 100 Tflops.

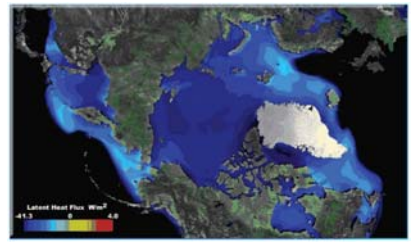
- Finland's CSC Supercomputer Center Selects Cray Hood System - Oct 9, 2006
  - CSC Finland, the Finnish IT center for science, will acquire a Cray Hood system delivering over 70 Tflops.

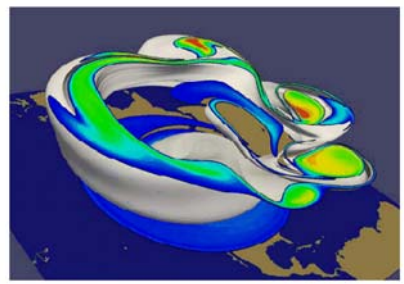
Significant Progress in the ES Market

- **Key wins and upgrades: KMA, NCMRWF, INM, KMA, AWI, NMOO, CERFACS...**
- **CWO modeling represents significant usage of large scale wins: ORNL, PSC, CSCS, Sandia, ERDC, HECToR, CSC,…**
- **Establishment of the KMA-Cray Earth System Research Center.**
- **Leading weather and climate science on Cray systems:**
  - **IPCC assessments on ORNL X1E.**
  - **DOE/NSF Climate-Science Computational End Station at ORNL using X1E and XT3.**
  - **Partner in Germany-Korea cooperation – "Earth System modelling and data analyses focusing on the East Asian region".**
- **Active participation in the community.**
- **Performance records on a number of leading weather and climate models.**
  - **Unprecedented scalability.**



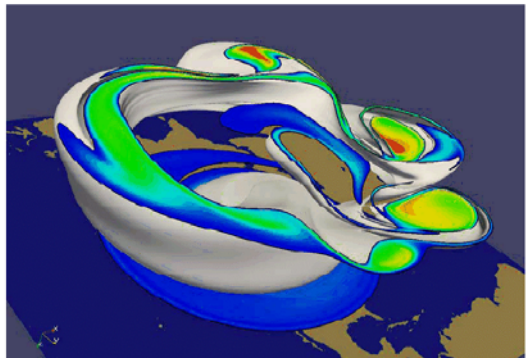
esrc
earth systems research center



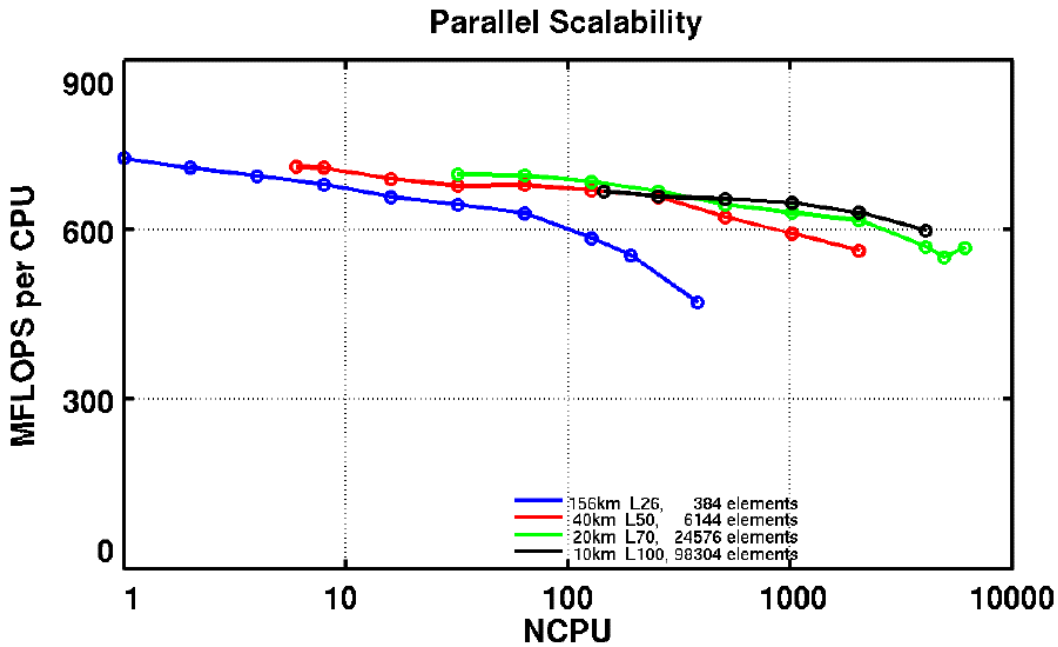


SEAM Scalability

- Spectral Element Atmospheric Model
- Spectral elements replace spherical harmonics in horizontal directions.
- Coupled to the Community Atmospheric Model (CAM)
- Run to 10km global resolution



Billion grid point simulation of a polar vortex that has trapped air at the pole.



Performance of 4 fixed problem sizes, on up to 6K CPUs. The annotation gives the mean grid spacing at the equator (in km) and the number of vertical levels used for each problem.

Sandia National Laboratories

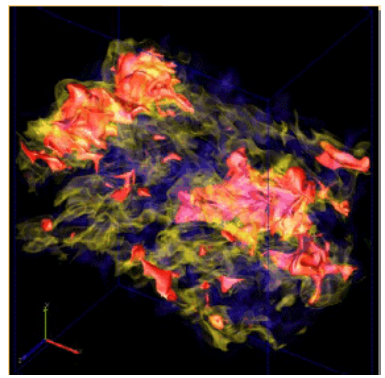
Increasing System Size

- Sandia National Laboratories Red Storm recently upgraded to 120 Tflops.
- NERSC 100 TFLOPS Hood in 2007.
- ORNL 50 TF, 100TF, 250TF, …1 Pflop by end of 2008.
- Preferred bidder at EPSRC/HECToR
  - Initial peak capability of over 50 TFLOPs
- AWE: 41 Tflops XT3
- CSC: >70 Tflops Hood
- Inertia of experience…

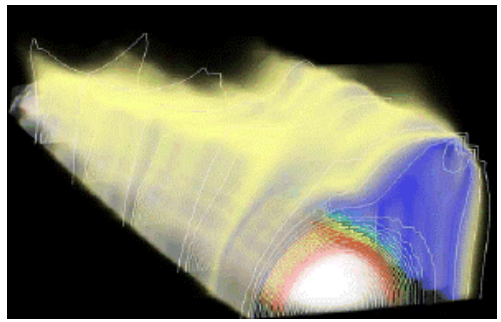
National Leadership Computing Facility

- Cray-ORNL Selected by DOE for National Leadership Computing Facility (NLCF).
- Goal: Deliver National Leadership Computing Facility for science:
  - Focused on grand challenge science and engineering applications
- Principal resource for SciDAC and other programs:
  - Specialized services to the scientific community: biology, climate, nanoscale science, fusion
- Today: 18 TF X1E and 50 TF XT3
- 250+ TF capability by 2007 and PF in late 2008.

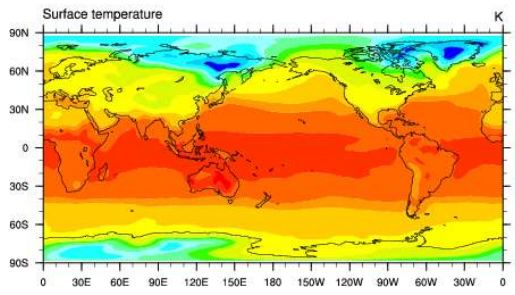
OAK RIDGE NATIONAL LABORATORY
CCS The Center for Computational Sciences
DOE High Performance Computing Research Center



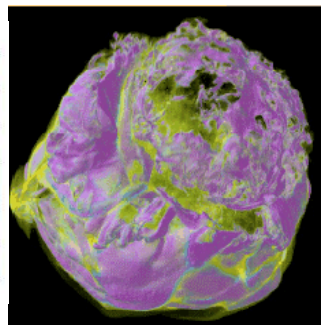
Combustion



Plasma Behavior Simulation



Climate



Astrophysics

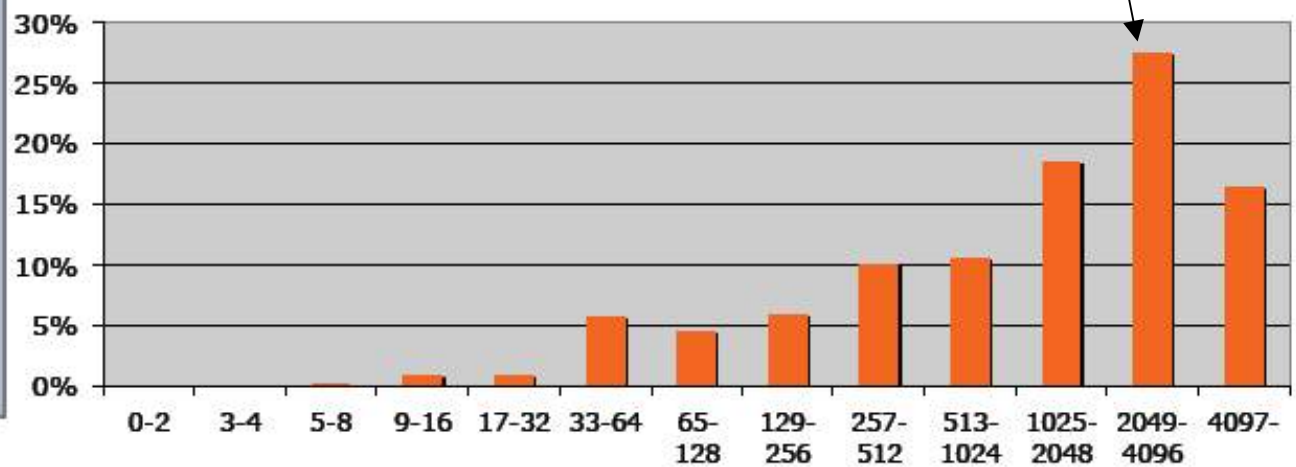
Jaguar
5,294 processors and 11 TB of memory

Since upgraded to 10,000+ processors

Accepted in 2005 and routinely running applications requiring 4,000 to 5,000 processors

A "good" MPP will generate usage patterns like this: Lots of jobs using 1000s of processors

- 43% of time used by jobs using 40% of system or more
- 61% of time used by jobs requiring 1000+ processors



Machine Usage by Number of Processors

Jaguar's Path to 250 TF

Jaguar 2006 upgrade

- Upgrade single-core to dual-core (2.6 GHz)
- Upgrade memory to maintain 2 GB per core
- Add 68 cabinets (total of 124)
- 11,508 dual-core compute sockets
- 119 TF peak
- 46 TB memory
- 900+ TB disk storage
- 55 GB/s disk bandwidth

Jaguar 2007 upgrade

- Upgrade 68 cabinets to multi-core
- O(5000) dual-core compute sockets
- O(6000) multi-core compute sockets
- Double memory in upgraded nodes
- 250+ TF peak compute partition
- 50+ TF data analysis partition
- *70 TB memory*

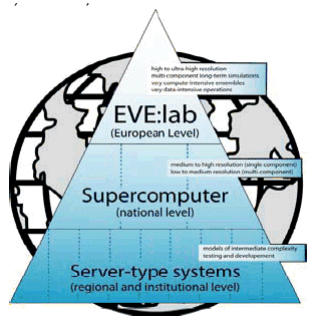
Petascale Ambitions

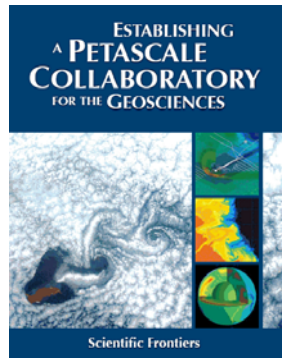
- **Japan, Europe and US climate communities have all announced ambitions to reach Petaflop level by early in the next decade:**
  - **Japan Earth Simulator II (MEXT)**
  - **Europe ENES EVE:lab**
  - **US NSF Petascale Collaboratory**

- **US and Japanese government funding into R&D petascale computing:**
  - **MEXT Petaflop Project**
  - **DARPA HPCS**

- **Ongoing NSF "Track 1" solicitation for Petascale facility by 2011.**

- **First actual contract is between ORNL and Cray to deliver a Pflop by 2008.**





京 速 計算機

DARPA

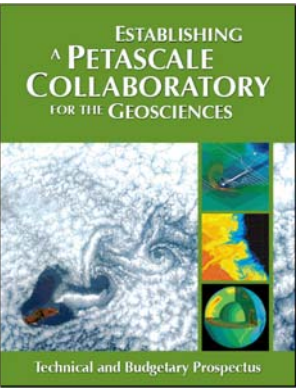
MEXT
Ministry of Education, Culture, Sports, Science and Technology

National Science Foundation
WHERE DISCOVERIES BEGIN

ornl
OAK RIDGE NATIONAL LABORATORY

CRAY

Diverse Application Requirements



Couple of comments:
- all apps will be "massively" parallel.
- Definition of performance metrics vary greatly.

Table 2. Application computational, memory, data storage, and disk bandwidth requirements for various geoscience applications.

Application Name/ Discipline	Problem	Max Required Sustained TFLOPS	System Memory (TBytes)	Mass Storage Archive Rate (PBytes/ year)	Disk Bandwidth for 5% overhead (GBytes/sec)
flow_solve/ oceanography	3-D turbulence	2.5	6.5	0.14	1.1
POP/ oceanography	10 km global mesoscale eddy	6	0.15	0.32-3.2	0.2-2.0
POP/ oceanography	5 km global mesoscale	120	1.5	3.2-32	2-20
MITgcm/ocean data assimilation	15 km global ocean	7.3	0.82	0.66	0.4
WRF/meteorology	10 m tornado simulation	150	20	2-24	25-300
	5 years of 3 km global nonhydro-static simulation	66	1.75	1.0	8
CAM/climate modeling	Five instances of T341L52	13	0.5	4.6	1.1
CRCP/climate modeling	2 km global sub-grid scale model	22	-	-	-
ABINIT/minerology	DFT calculation	1.6	-	-	-
inverse problem/regional seismology	100M point inverse problem	17	0.01	0.12	0.07
forward problem/global seismology	36.6 billion degrees of freedom	10.4	7.3	0.01	0.00002
LADHS/regional hydrology	100 m Columbia river basin	10	0.3	20.8	0.66

Achieving Effective System Scalability

- Computation is increasingly cheap compared to data movement.
- Most hardware cost is in interconnect:
  - Packages, circuit boards, connectors, wires, routers, electro-optics, fibers, etc.
- Rate of increase in system scale is increasing:
  - Reliability, fault detection/containment/recovery requirements.
  - Network bandwidth becomes even more critical (cost and performance).
  - Need to get the system software right.
  - Scalability is primary determinant.
  - A holistic approach is mandatory: computing, system software, I/O, data management,….

Achieving Effective System Scalability

- Moore's Law has affected all systems:
  - Classic vector systems relied upon uniform connectivity to global shared memory (no memory hierarchy).
  - For vector systems to advance at Moore's Law rate, they have been forced to adopt hierarchical memory.
  - This is a *system* architecture issue; orthogonal to processor architecture.

- All systems are massively parallel…
- …. Only a question of how effectively they can scale and be utilized.
- Must invest in a wide range of technologies to achieve effective system scalability.

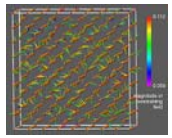
Cray MPP Experience and Lessons

- **Began with MPP Advisory Group**
  - **Formed in 1991, a mechanism to engage user community to understand MPP issues.**
- **Cray XT3 is 3rd generation MPP product.**
- **2 key areas:**

- **Lightweight Kernel:**
  - **Jitter free OS permits finest grain synchronization for MPP performance on broadest range of problems**
  - **No time sharing**
  - **No large SMPs**
  - **No paging**
  - **No spurious daemons**
  - **Minimal OS interrupts**
- **Interconnect:**
  - **Scalability in all aspects**
  - **Interconnect architecture**
  - **High bandwidth global I/O where files are equally accessible from any part of the system**
  - **Administrative ease with single boot image and fast boot time**
  - **Fast application launch**
  - **Resiliency**

Capability MPP Computing at Cray

Cray T3E1200:
- Sustained TF achieved on 1480 processors
- Gordon Bell Prize Winner



Cray X1:
- First system delivered
- Vector MPP

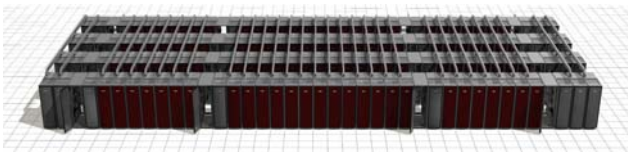
KMA Cray X1E:
- World's Fastest NWP system
- ~6 Tflops sustained

"BlackWidow"

Common Infrastructure
- ***Transparent***
- ***Scalable***
- ***Optimized Scalar + Vector***

1998 2000 2002 2004 2005 2006 2007

Sandia Red Storm Contract:
- 10,000+ processor machine
- Delivery in 2004
- Balanced, 40Tflops System



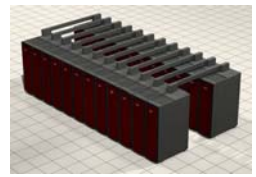
Cray XT3:
- 3rd Generation Microprocessor MPP
- First Cray XT3 Order and Deliveries

Cray XT3 Results:
- Scaling to 1000's of processors.

Key Large Scale Wins:
- AWE, NERSC, HECToR, CSC, …

Three 100+ TF Systems Installed

Cray "Hood"



The Cray Roadmap

Following the Adaptive Supercomputing Vision

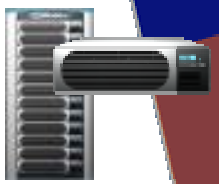
Cray X1E

BlackWidow
2007

Eldorado
2007

Rainier

Cascade



Cray XD1

2006
Cray XT3
Dual-Core

2007
Hood
Dual-Core

2008
Hood
Multi-Core

2009
Baker

Cray XT3

Phase I: Rainier
Multiple Processor Types with
Integrated User Environment

Phase II: Cascade
Fully Integrated
Multi-Architecture System

Summary

- Continued product innovation focused on high performance computing:
  - Leverage and integrate key technologies.
  - Invest in emerging technologies to meet future requirements.
  - Invest in customer relationships to further develop technologies.
- The successful realization of production petascale facilities will require both the earth system modeling and HPC vendor communities to maximize every level of parallelism available.
- Cray's MPP experience and technologies will play an important role in the realizing future needs of earth system modelers.


Thank you for your attention
CRAY
THE SUPERCOMPUTER COMPANY