Oak Ridge National Laboratory
Cray X1 and Black Widow
Evaluation and Plans

Thomas Zacharia
Associate Laboratory Director
Oak Ridge National Laboratory

Presented to the DOE ASCAC
March 13-14, 2003
Center for Computational Sciences

- Evaluate new hardware for science
  - Development and evaluation of emerging and unproven systems and experimental computers
- Deliver leadership-class computing for DOE science
  - Offer specialized services to the scientific community
  - Principal resource for SciDAC
  - By 2005: 50x performance on major scientific simulations
  - By 2008: 1000x performance
- Educate and train next generation computational scientists
- Designated User Facility in 1994

Intel Paragon  IBM Power3  IBM Power4  Cray X1
54% of CCS resources dedicated to SciDAC
FY03 CCS SciDAC usage

SciDAC project usage as total of 54%

- Fusion: 21%
- Climate: 26%
- Chemistry: 19%
- Astrophysics: 9%
- QCD: 18%
- ISICs: 7%
- Accelerator: 0%
All orders spectral code in 3D for fusion simulation completed

- AORSA3D is MPI code that uses SCALAPACK to solve linear systems arising from spectral discretization
- Preliminary calculation for Fast Wave minority heating on LHD stellarator
- $16 \times 50 \times 50$ modes in $\phi$, $x$, $y$ (10 independent solutions - one per field period)

- Each solution requires 576 processors on ORNL Eagle system for 6.5 h. Total of 3744 CPU hr. Sequence of 10 runs $\rightarrow$ 37,440 CPU hr $\approx$ 100,000 MPP hours
- Convergence study calculation with 32 toroidal modes $\times 40 \times 40$ modes required almost $10^6$ MPP hours
FY03 Oct-Feb CCS usage by discipline

- Physics: 17%
- Fusion: 13%
- Materials: 16%
- Biology: 11%
- Chemistry: 20%
- Climate: 17%
- Other: 6%
CCS is a National User Facility

- Available to the national user community
- Four types of user agreements
  - UC-Nonproprietary no cost agreement for commercial users
  - UA-Nonproprietary no cost agreement for educational users
  - UR-Nonproprietary cost required agreement for all users
  - UF-Proprietary cost required agreement for all users
- One of twenty user facilities managed by Oak Ridge National Laboratory
  - >500 user agreements in place
- Cray is most recent
  - User Agreement UF-03-277


World Class CCS Facilities
Office of Science computing capability is significantly behind leadership class

The last time Office of Science researchers had access to a leadership class computer was in 1995.
ASCAC statement

Without robust response to the Earth Simulator, U.S. is open to losing its leadership in defining and advancing frontiers of computational science as new approach to science. This area is critical to both our national security and economic vitality. (Advanced Scientific Computing Advisory Committee, May 21, 2002)

ORNL-CCS response
- CCS held series of workshops and meetings with users and vendors
  - Cray, HP, IBM, SGI, Others
Even though clusters of general purpose SMPs dominate U.S. HPC....

- Largest DOE systems
  - NNSA: LANL (HP), LLNL (IBM, Intel)
  - SC: LBNL (IBM), ORNL (IBM), ANL (Intel), PNL (Intel)
- Largest NSF systems
  - PSC (HP), NCAR (IBM), SDSC (IBM), NCSA (Intel)
- Largest (known) DOD systems
  - NAVO (IBM), ARL (IBM)
- Largest of other U.S. agencies
  - NOAA (Intel), NASA (SGI)
- Largest state systems
  - LSU (Intel), SUNY (Intel), FSU (IBM), NCSC (IBM)
... the science community we serve and our users found:

**Increasing**
- Processor speed
- Parallelism
- Algorithm efficiency
- Computational requirements for scientific simulation
- Relative memory and interconnect latencies
- Power consumption
- Heat generation
- System complexity
- Software complexity

**Decreasing**
- Relative memory bandwidth
- Relative interconnect bandwidth
- Relative I/O speed
- % of peak performance

Our users requested a balanced, leadership-class system based on science needs
Cray X1 provides balanced system for science applications

Cray X1

- Commercial name for “SV2” project that Cray has been building for NSA for more than 4 years
- Combines multi-streaming vector processors with a globally addressable memory similar to T3E
- Offers best opportunity for leadership class system for delivered performance in scientific applications such as climate, materials, astrophysics, fusion
- Proposal entitled “Reasserting U.S. Leadership in Scientific Computation” was submitted to evaluate and deploy Cray X1 on July 4, 2002
Cray X1 scalable vector architecture

- Powerful vector processors with integer and bit operations
- Very high memory bandwidth, but with cache
- Works well for short vectors
- Ultra-high bandwidth interconnect
- 2-D torus topology

Diagram showing 12.8 GF (64bit) CPUs connected to 8 to 64 GBytes Memory with 200 GB/sec bandwidth. I/O connections are also shown with 100 GB/s capacity.
Comparisons of peak 10TF systems based on current OASCR computers

**CCS-3**
- IBM Power4
- 10 teraflops peak
- 1920 processors
- 5.2 gigaflops processor
- 32 processors per cabinet
- 60 cabinets
- Well understood, stable system; *Federation interconnect should make this machine more attractive*
- Commodity manufacturing
- Cost: $35M (includes 5-year maintenance)

**NERSC-3E**
- IBM Power3
- 10 teraflops peak
- 6,656 processors
- 1.5 gigaflops processor
- 64 processors per cabinet
- 104 cabinets
- Well understood, stable system
- Commodity manufacturing
- Cost: $75M* ($45M+$30M) (includes 5-year maintenance)

*estimate

**CCS-4**
- Cray X1
- 10 teraflops peak
- 768 processors
- 12.8 gigaflops processor
- 64 processors per cabinet
- 12 cabinets
- Most balanced system
- Designed for scientific computing
- Commodity manufacturing
- Cost $71M (includes 5-year maintenance)
CCS-Cray plans: near term
MOU with Cray on 8-14-2002

2003

X1
12.8 GF
35 GB/s/p mem BW
76 GB/s/p cache BW

End of ‘04

X1e
Technology upgrade
Faster clock
Denser Package
Mix&Match with X1

2005+

X2
(Black Widow)
Bigger
Faster
Cheaper

Synergy?
I/O?

2003

Red Storm
(Development)

End of ‘04

Red Storm
(Install)
Catamount LWK
Linux service
AMD 2 GHz X86-64

2005

Red Storm
Product (?)
Linux Service
Compute OS?
CCS-Cray plans: not so near term
Cascade technical affiliate program

- DARPA HPCS program
- Shared memory locales
  - UMA, NUMA
- Heavy-weight processors
  - Multithreading, vectors, streams
- PIM (LWP)
Cray X1/X2
4-phase evaluation and deployment

- X2 architecture per DOE apps
  Design changes for X2 based on ORNL-Cray partnership
  Phase 1: 3.2 TF, 1 TB, 20 TB
  Phase 2: 8.192 TF, 2.621 TB, 24 TB
  Phase 3: 40.96 TF, 13.107 TB, 102 TB
  Phase 4: 120 TF, 40 TB, 400 TB

- Earth Simulator
  Phase 1: 3 TF (256 CPU)
  Phase 2: 8 TF (640 CPU)
  Phase 3: 40 TF (3200 CPU)
  Phase 4: 120 TF

Timeline:
Statement of work

- Seller: Cray Inc., Small Business
- Items: Cray X1 Systems, Cray Black Widow Systems, Source Code Licenses, and Maintenance
- Delivery Schedule
  - Base System/Phase 1: March 2003 – Summer 2003
    - Eight half populated cabinets
    - Contract signed and approved
  - System Option II/Phase 2: October 2003 – December 2003
    - Upgrade phase 1 system to fully populated 8-12 cabinet systems (10TFlops peak) based on results of initial evaluation
  - System Option III/Phase 3: January 2004 – June 2004
    - Add sufficient additional cabinets based on science and community needs for leadership class machine
    - Expandable to 64 cabinets
  - System Option IV/Phase 4: Commencing late 2005
    - Upgrade to 120 TF Cray Black Widow system
Center for Computational Sciences
Cray X1 System

- Picture of Cray X1 at factory awaiting shipment to ORNL
- Delivery scheduled for March 18th
- 32-processor, liquid-cooled cabinet
- 128 GB memory
- 8 TB disk
Summer 2003

- 3.2 TFlops
- 256 processors
- 1 TB shared memory
- 32 TB of disk space
- 8 cabinets
Detailed evaluation plan developed in concert with user community (http://www.csm.ornl.gov/meetings/)

- Applications Workshops, November 5-6, 2002
  - Climate, Materials, Biology, Fusion, Chemistry, Astrophysics
  - Cray, ORNL, LANL, PNNL, NCI

- Cray X1 Tutorial, November 7, 2002
  - >100 attendees from 20+ sites

- Cray-Fusion Workshop, February 3-4, 2003
  - Cray, ORNL, PPPL, U. Wisconsin, U. Iowa, General Atomics

- SciDAC CCSM Workshop: Porting CCSM to the Cray X1
  - NCAR, February 6, 2003 (followed CCSM Software Engineering Working Group meeting)
  - Cray, NEC, ORNL, NCAR, LANL

- Computational Materials Science: Early Evaluation of the Cray X1
  - Austin, March 2, 2003 (in conjunction with APS Meeting)
  - Invitees from 15 sites

- Cray Biology Workshop, March 18, 2003
Science applications driven evaluation and benchmarking

- Potential application
  - important to DOE Office of Science
  - scientific goals require multi-terascale resources

- Potential user
  - knows the application
  - willing and able to learn the X1
  - motivated to tune application, not just recompile

- Set priorities
  - potential performance payoff
  - potential science payoff

- Schedule the pipeline
  - porting/development
  - processor tuning
  - scalability tuning

  - science!
Climate (CCSM) simulation resource projections

At current scientific complexity, one century simulation requires 12.5 days
Single researcher transfers 80Gb/day and generates 30TB storage each year

Science drivers: regional detail / comprehensive model

Machine and Data Requirements

- Blue line represents total national resource dedicated to CCSM simulations and expected future growth to meet demands of increased model complexity
- Red line shows data volume generated for each century simulated
Climate science

- Collaboration with CCSM project
  - NCAR will provide two development branches for NEC and Cray
  - ORNL and NCAR gate-keepers will propose merges to CCSM
- Community Atmospheric Model (CAM)
  - “Physics” stresses single-processor vectorization and multistreaming
  - “Dynamics” stresses interconnect bandwidth
- Community Land Model (CLM)
  - Stresses vectorization and memory bandwidth
  - Atmospheric coupling stresses interconnect
- Parallel Ocean Program (POP)
  - “Baroclinic” should scale well - does it?
  - “Barotropic” stresses interconnect latency
  - Latency bound on SX-6
Porting strategies for the CCSM to Cray X1 and NEC SX

**Cray X1**
- Atmosphere (ORNL-CRAY)
  - Activate Eulerian vector dynamics
  - Radiation dropped 25% to 7%
- Ocean (LANL-CRAY)
  - Replace utilities 75% to 0%
  - 2-D Barotropic with CoArray Fortran
- Land (ORNL-NCAR)
  - Move column loop
- Sea-Ice (LANL)
- Coupler (ANL)

**Nec SX**
- Atmosphere (NCAR-NEC)
  - FV dynamics
  - Radiation same
- Ocean (LANL-CRIEPI)
  - Vectorized out of the box by replacing one module
- Sea-Ice (CRIEPI)
  - Vectorization improved x 50
<table>
<thead>
<tr>
<th>Vectorization task descriptions</th>
<th>Lead</th>
<th>Others</th>
<th>FTEs</th>
<th>months</th>
<th>FTE-yrs</th>
<th>Status</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Software Management</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repository branch for Cray</td>
<td>Craig</td>
<td></td>
<td>0.33</td>
<td>25%</td>
<td>1Q03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repository branch for NEC</td>
<td>Craig</td>
<td></td>
<td></td>
<td>25%</td>
<td>1Q03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merge to dev branch</td>
<td>Boville</td>
<td>Drake</td>
<td>2</td>
<td>2</td>
<td>0.33</td>
<td>ongoing</td>
<td>4Q03</td>
</tr>
<tr>
<td><strong>Atmospheric Model (CAM2)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUL dynamical core</td>
<td>Courdery</td>
<td>Worley</td>
<td>1</td>
<td>2</td>
<td>0.17</td>
<td>25%</td>
<td>2Q03</td>
</tr>
<tr>
<td>FV dynamical core</td>
<td>Parks</td>
<td>Boville</td>
<td>1</td>
<td>2</td>
<td>0.17</td>
<td>25%</td>
<td>2Q03</td>
</tr>
<tr>
<td>Radiation physics</td>
<td>Courdery</td>
<td>White</td>
<td>3</td>
<td>3</td>
<td>0.75</td>
<td>25%</td>
<td>2Q03</td>
</tr>
<tr>
<td>Other physics</td>
<td>White</td>
<td></td>
<td>1</td>
<td>3</td>
<td>0.25</td>
<td></td>
<td>1Q03</td>
</tr>
<tr>
<td>Message passing optimization</td>
<td>Worley</td>
<td>Putman</td>
<td>1</td>
<td>1</td>
<td>0.08</td>
<td></td>
<td>3Q03</td>
</tr>
<tr>
<td><strong>Ocean Model (POP2)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baroclinic</td>
<td>Jones</td>
<td>Levesque</td>
<td>2</td>
<td>2</td>
<td>0.33</td>
<td>75%</td>
<td>1Q03</td>
</tr>
<tr>
<td>Barotropic solve</td>
<td>Jones</td>
<td>Levesque</td>
<td>2</td>
<td>4</td>
<td>0.67</td>
<td>25%</td>
<td>2Q03</td>
</tr>
<tr>
<td>Message passing optimization</td>
<td>Jones</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0.08</td>
<td></td>
<td>3Q03</td>
</tr>
<tr>
<td><strong>Land Model (CLM2.1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prototype structure</td>
<td>Hoffman</td>
<td>Vernstien</td>
<td>1</td>
<td>2</td>
<td>0.17</td>
<td>25%</td>
<td>1Q03</td>
</tr>
<tr>
<td>Modify pft process routines</td>
<td>Hoffman</td>
<td>Vernstien</td>
<td>2</td>
<td>3</td>
<td>0.50</td>
<td></td>
<td>2Q03</td>
</tr>
<tr>
<td>River routing scheme</td>
<td>White</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0.08</td>
<td></td>
<td>3Q03</td>
</tr>
<tr>
<td><strong>Sea Ice Model (CICE2)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incremental remapping</td>
<td>Jones</td>
<td></td>
<td>1</td>
<td>2</td>
<td>0.17</td>
<td>25%</td>
<td>2Q02</td>
</tr>
<tr>
<td>Elastic-Viscous Plastic dynamics</td>
<td>Jones</td>
<td>Hunke</td>
<td>1</td>
<td>2</td>
<td>0.17</td>
<td></td>
<td>1Q02</td>
</tr>
<tr>
<td><strong>Coupler (CPL6)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Coupling Toolkit</td>
<td>Larson</td>
<td></td>
<td>1</td>
<td>3</td>
<td>0.25</td>
<td></td>
<td>2Q03</td>
</tr>
<tr>
<td>Unit test</td>
<td>Kaufman</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0.08</td>
<td></td>
<td>3Q03</td>
</tr>
<tr>
<td>Live model test</td>
<td>Craig</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0.08</td>
<td></td>
<td>4Q03</td>
</tr>
<tr>
<td><strong>Total Effort</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Status and Target**
- **25%**
- **50%**
- **75%**
- **done**
16P X1 provides best POP performance

<table>
<thead>
<tr>
<th>System</th>
<th>Resource (Res)</th>
<th>Y/day</th>
<th>Nproc</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES40</td>
<td>1</td>
<td>60.0</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>3.6</td>
<td>960</td>
</tr>
<tr>
<td>IBMp4</td>
<td>1</td>
<td>24.8</td>
<td>256</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.118</td>
<td>480</td>
</tr>
<tr>
<td>CrayX1</td>
<td>1</td>
<td>35.3</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.25</td>
<td>16</td>
</tr>
</tbody>
</table>
Extremely optimistic results on X1 however, much work remains

Co-Array Fortran

SHMEM

32 processors, one p690
Science projects with CCSM

Cray
- ORNL – Evaluation of X1 architecture for high resolution climate change simulations

CRIEPI
- NCAR – Whole Atmosphere Simulations
- LANL – Eddy resolving ocean simulations and hi-res IPCC runs
An opportunity for extraordinary discovery through effective integration

Spallation Neutron Source (SNS)
Center for Nanophase Materials Science (CNMS)

$1.3B
$150M/yr

Synthesis & Characterization
~ 12,000 atoms
~ 4,000 surface + sub

First principles simulation size
Current largest 2176-atoms (3TF)

Real device size
Nano dot: 5x5x5 (6TF)
Nano wire: 10x10x60 (250TF)
New algorithms

Theory and Simulation

J_Bulk ≠ J_Surface

Fe

5 nm

250 TF Cray X2

OAK RIDGE NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
DCA-QMC calculations of strongly correlated electronic materials

http://www.physics.uc.edu/~jarrell/Research/myresearch.html
Performance of Boltztran on X1
Calculates Boltzmann neutrino transport

46% efficiency without any tuning for astrophysics application

http://www.phy.ornl.gov/tsi/
Computational fluid dynamics with chemistry shows good performance and scalability on X1
Summary

- Cray X1 offers balanced architecture for science
  - Eight half populated cabinets in FY03
  - Upgrade to 10TFlops in FY04
  - X1 and follow on systems scalable to 100+TF based on Office of Science needs
  - New private sector-funded CCS facility will be ready to house Cray

- Unprecedented performance on Office of Science applications
  - Factor of 3-50 better sustained performance
  - Critical to DOE mission goals

- Access to system source code to tune system for DOE applications

- Strong collaborative partnership and opportunity to guide development of next generation “Black Widow” system based on DOE applications
  - Design changes for X2 based on ORNL-Cray partnership