CUG 2007
NEW FRONTIERS

The Naval Research Laboratory Cray XD1

Wendell Anderson
Jeanie Osburn
Robert Rosenberg
Naval Research Laboratory

Marco Lanzagorta
ITT Corporation

Presentation Outline

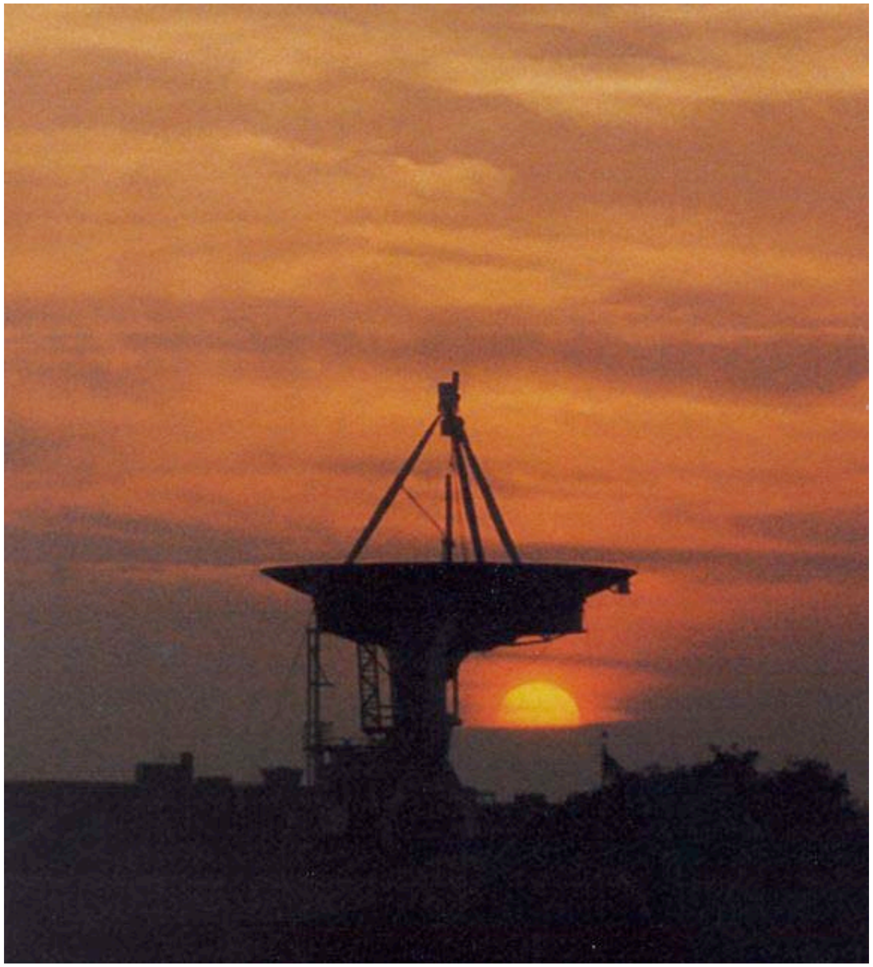
1. The NRL’s Cray XD1System
2. Scientific Supercomputing
3. Reconfigurable Supercomputing
4. FPGA Programming Tools
5. Performance Measurements
6. Problems and Issues
7. Conclusions

Presentation Outline

1. **The NRL’s Cray XD1System**
2. Scientific Supercomputing
3. Reconfigurable Supercomputing
4. FPGA Programming Tools
5. Performance Measurements
6. Problems and Issues
7. Conclusions

US Naval Research Laboratory

NRL is the US Navy's corporate research laboratory under the Office of Naval Research.



CCS

NRL's Center for Computational Science:

- Distributed Center under the HPCMO
- Provides leading edge HPC resources to the Navy
- Conducts evaluation, benchmarking, research, and development in HPC.

NRL's XD1

- 216 nodes with 864 cores and a cumulative speed of 3.5 TF.
  - Each node consists of two Opteron 275 2.2 GHz dual core processors with 8 GB of shared memory, and 73 GB 10K rpm 3.5 in. SATA data.
- 144 Xilinx Virtex-II Pro and 6 Virtex-4 FPGAs.

Software

- Cray modified version of SUSE Linux
- PGI and GNU Fortran and C/C++ compilers.
- MPI support through mpich 2.6
- AMD Core Math Library and Cray Scientific Library.
- Xilinx software and tools, Mitrion-C, Handel-C, and DSPLogic.

Node Usage

The XD1 nodes are used as:

- 4 to support the 30 TB Lustre system
- 1 for monitoring
- 1 for login
- 204 compute nodes scheduled with PBS
- 6 Virtex-4 compute nodes scheduled with PBS

Presentation Outline

1. The NRL's Cray XD1System
2. **Scientific Supercomputing**
3. Reconfigurable Supercomputing
4. FPGA Programming Tools
5. Performance Measurements
6. Problems and Issues
7. Conclusions

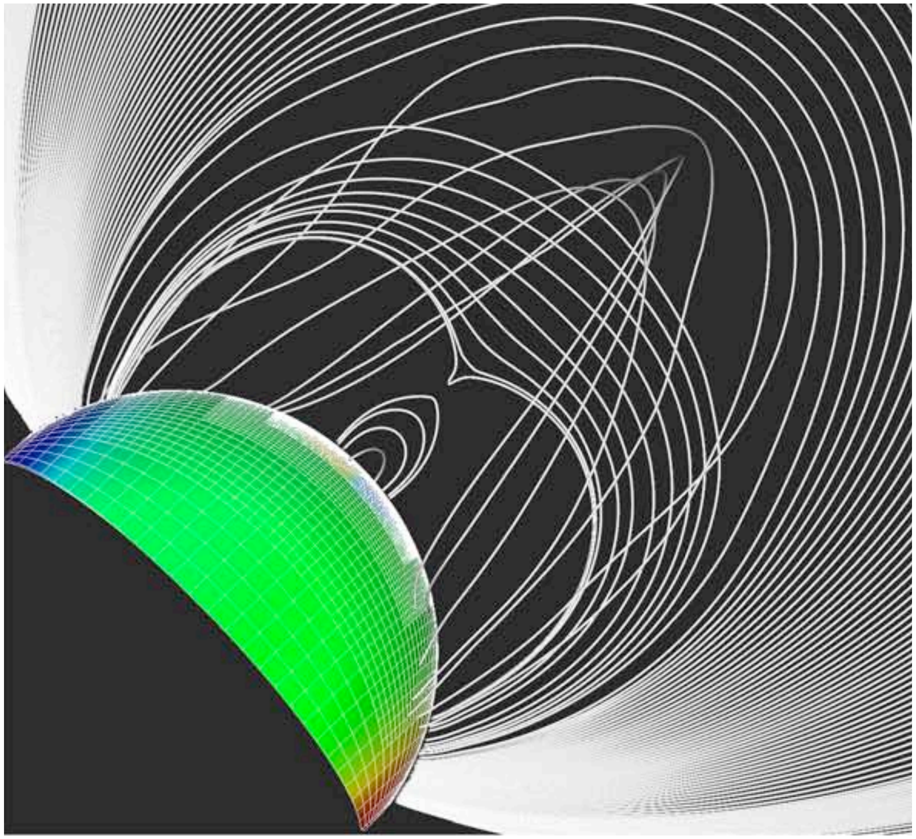
Scientific Applications

- Popular resource for scientific computing
- Provided over 3.3 million core hours.

Code	Avg. # Cores/Run	Core Hrs
ARMS	128-256	1,350,000
NOZZLE	128-256	800,000
NRLMOL	64-96	600,000
ADF	32	120,000
CHARMM	32	90,000
STARS3D	12	80,000

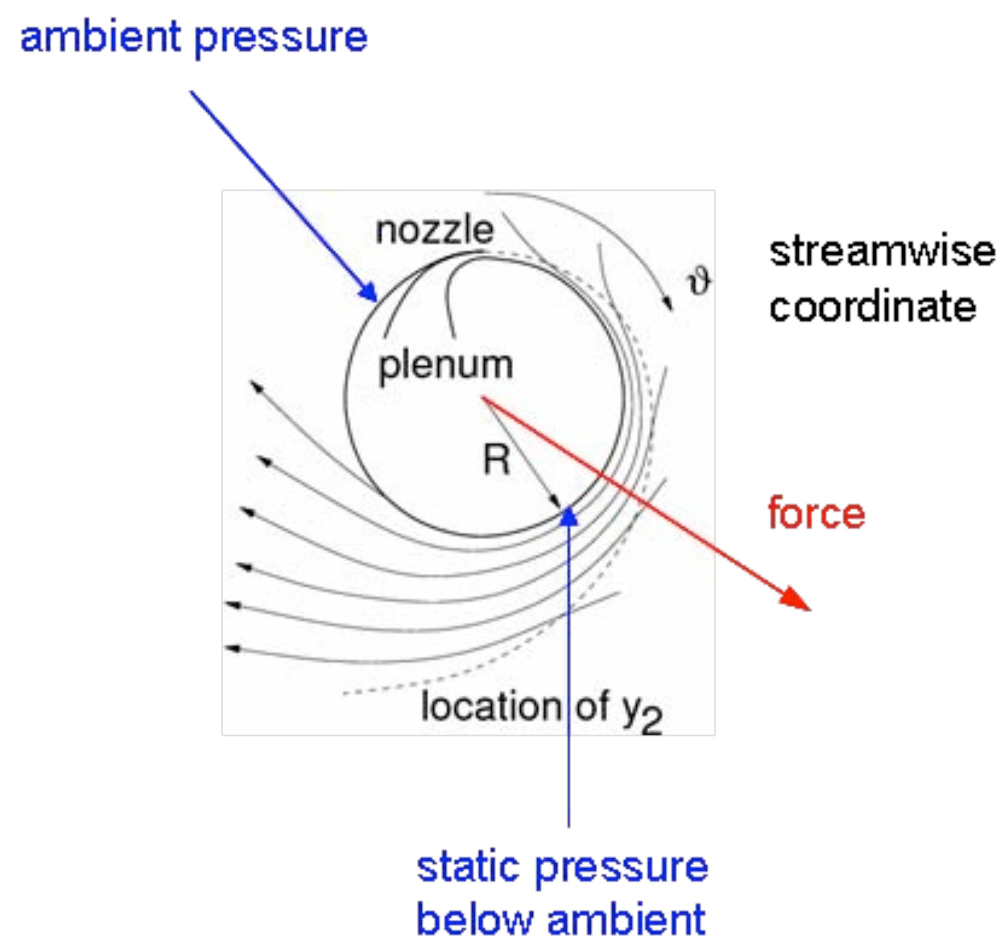
ARMS

Simulation of solar storms by Dr C. R. DeVore and Dr S.K. Antiochos.



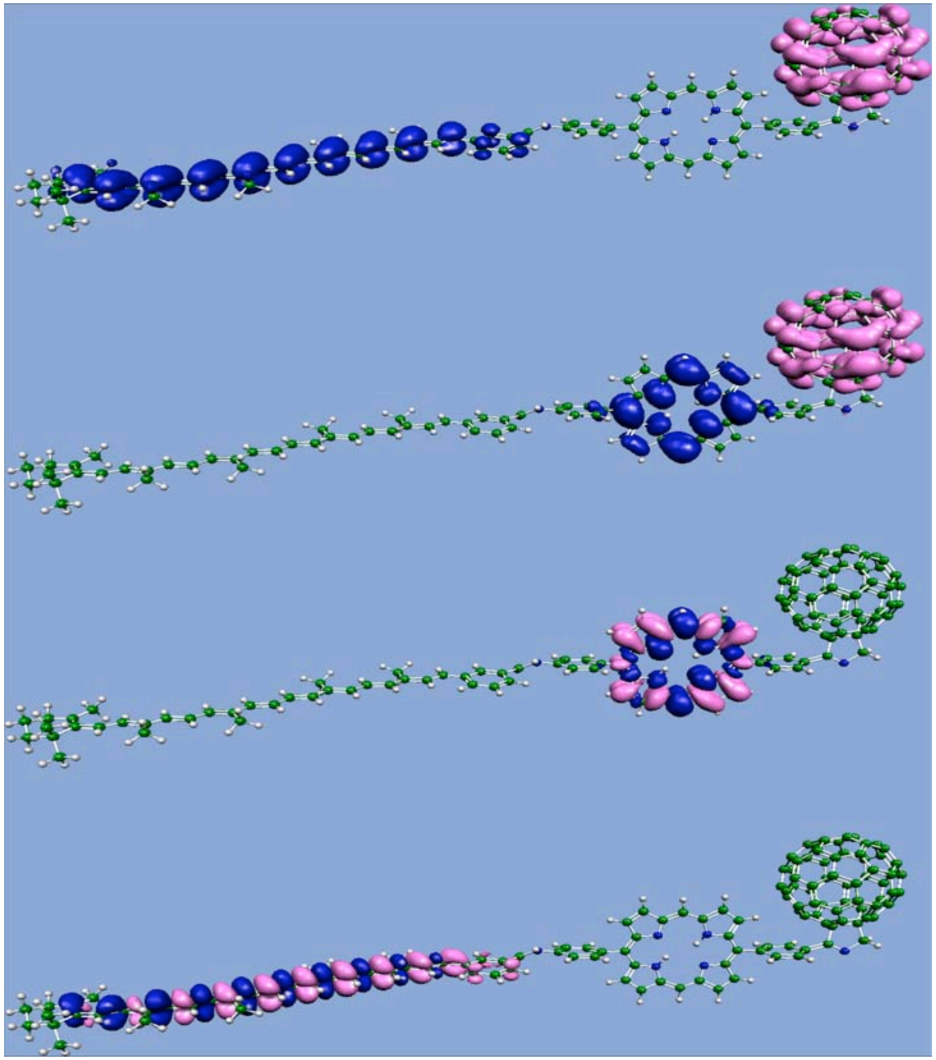
NOZZLE

Simulation of Coanda wall jet experiments by Dr A Gross.



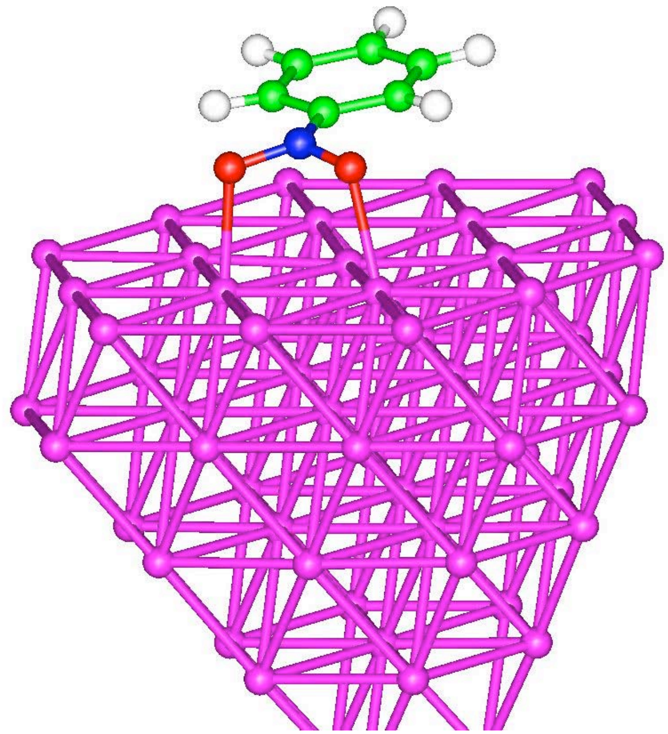
NRLMOL

Dr T Baruah and Dr M Pederson's study of the molecular vibrational effects on the simulation of a light-harvesting molecule.



ADF

Quantum-chemical analysis of the interaction between chemical warfare agents and materials by Dr S Badescu and Dr V Bermudez.

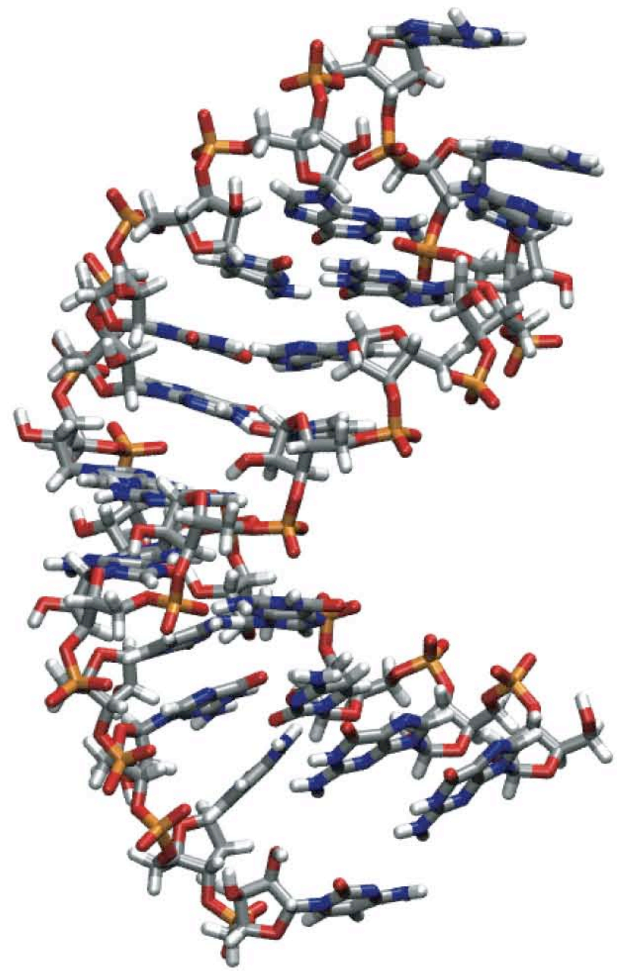


Nitrobenzene

$Ag_{56} + C_6H_5NO_2$

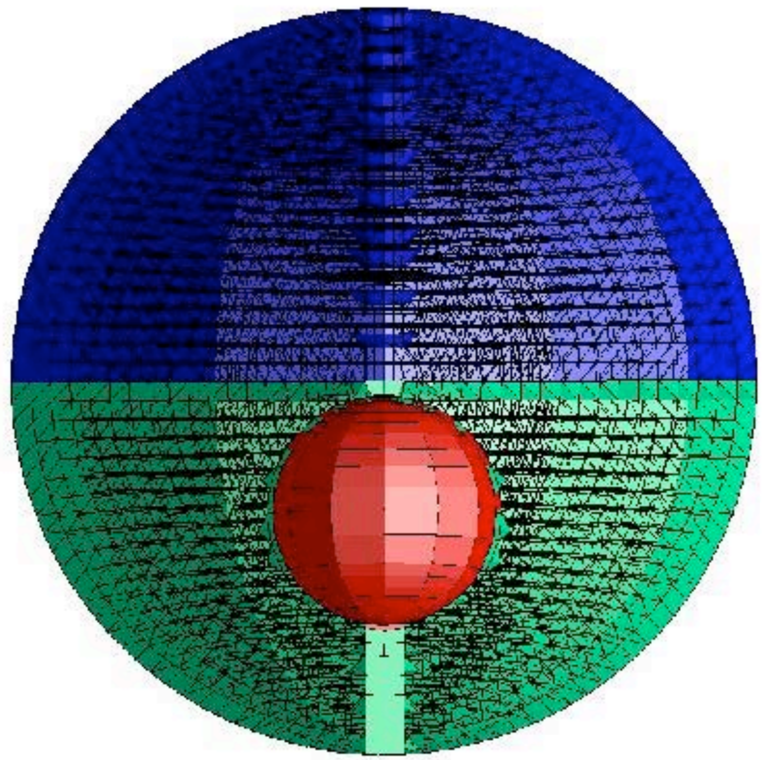
CHARMM

Study of the interaction between urea and P5GA RNA by Dr A MacKerell, D Priyakumar, and Dr Jeff Deschamps



STARS3D

Dr S Dey uses STARS3D to study wideband acoustic radiation and scattering from submerged elastic structures.



Presentation Outline

1. The NRL's Cray XD1System
2. Scientific Supercomputing
3. **Reconfigurable Supercomputing**
4. FPGA Programming Tools
5. Performance Measurements
6. Problems and Issues
7. Conclusions

Reconfigurable Supercomputing

- With 150 FPGAs, NRL's XD1 is the largest reconfigurable Cray supercomputer.

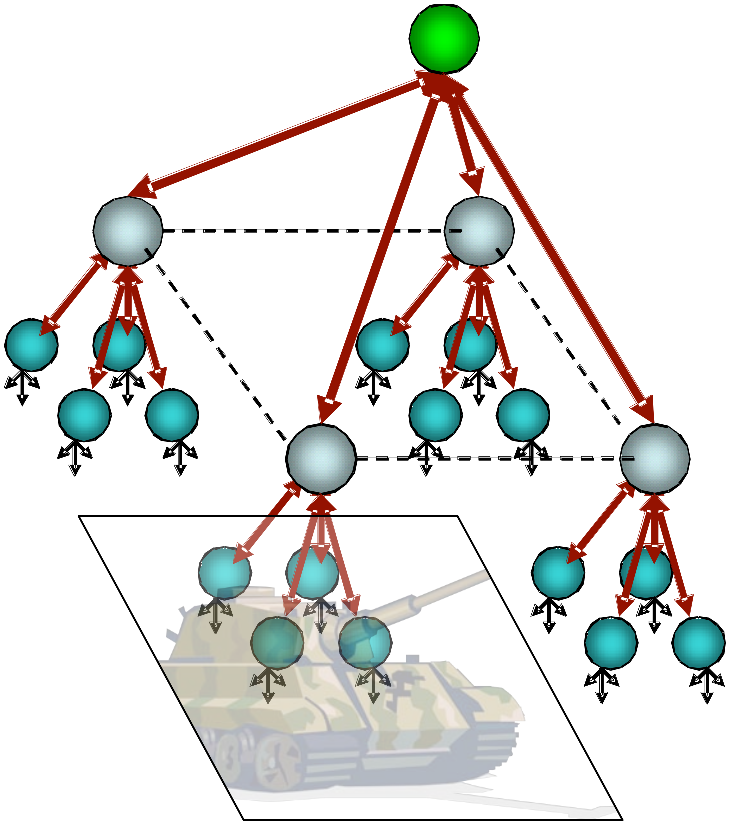
- We have started to explore the application of FPGA to accelerate scientific codes.

Porting Codes

- First applications are from users who already had VHDL codes running on a local system with a single FPGA.
- Main challenge has been the porting of their codes to the XD1.

Neural Networks

- Ken Rice and Tarek M Taha from Clemson University study large-scale models of the neocortex.
- Modeled up to 321 nodes using 64 of the XD1's Virtex-2 FPGAs.



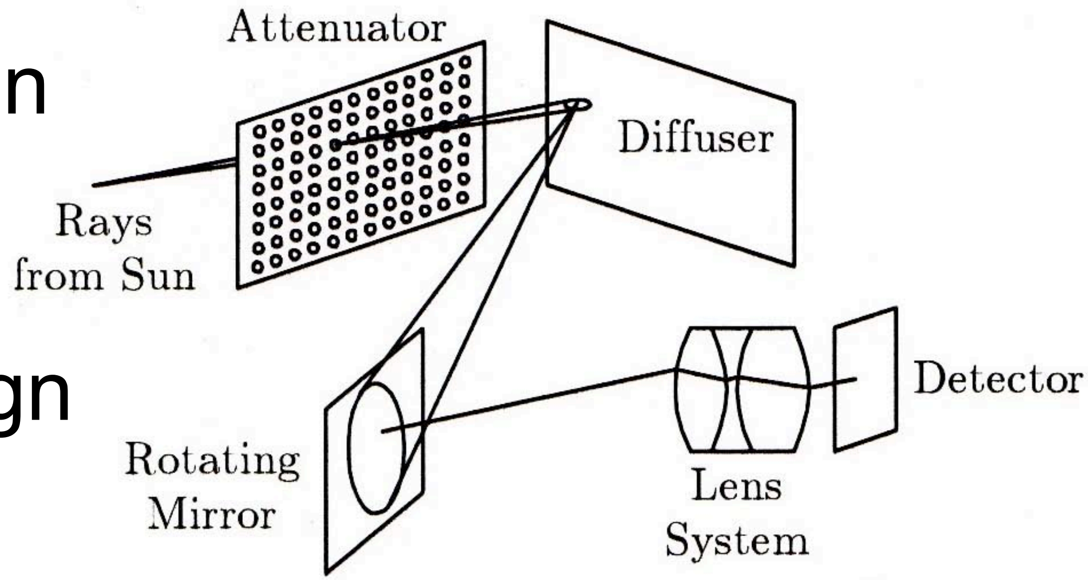
Neural Networks

Preliminary benchmarks suggest the following speedups over a single AMD core:

- Using all 864 cores: 720
- Using all 144 V2P FPGAs, with no SDRAM use: 31,246.
- Using all 144 V2P FPGAs, with SDRAM use: 128,389.

Design of Optical Devices

Commander Charles Cameron has been using ray tracing software to design optical devices.



BLASTN

- NRL is currently working with Mitrionics to port their SGI RASC FPGA BLASTN implementation to the XD1.
- Main problems:
  - SGI uses 128-bit data paths from a pair of QDRAMS. XD1 requires 64-bit data paths from a single QDRAM.
  - Cray has not finished the Virtex-IV interface to the XD1.

Other

Several other scientists are in the initial stages of investigating the potential applications of FPGAs to:

- Cryptography
- Hyper-Spectral Image Processing
- Ray Tracing
- Line of Sight Calculations
- Molecular Dynamics

Principal Challenges

- Identification of the portions of a code that are good candidates for FPGA acceleration.
- Programming of the FPGAs.
- Lack of established FPGA programming strategies for algorithm development.
- Lack of portability across HW platforms and across FPGA programming tools.

Presentation Outline

1. The NRL's Cray XD1System
2. Scientific Supercomputing
3. Reconfigurable Supercomputing
4. **FPGA Programming Tools**
5. Performance Measurements
6. Problems and Issues
7. Conclusions

High Order FPGA Programming Tools

- FPGA programming using VHDL or Verilog is a difficult and time consuming task.
- There are a few software packages that provide simpler methods to program FPGAs.
- We are currently testing and evaluating three of these software packages: Mitrion C, Handel C, and DSPLogic.

Mitrion-C

- Developed by Mitrionics.
- Currently supported on Cray XD1, SGI RASC RC100, and Nallatech BenDATA-DD.

Mitrion-C (+)

Advantages:

- C-like syntax and constructions.
- Straightforward “translation” from ANSI C to Mitrion C.
- Concurrent language with parallel data structures and parallel control flow directives.
- Easier than VHDL or Verilog.
- Good simulation, debugging, and algorithm development tools.

Mitrion-C (-)

Disadvantages:

- Most HPC users are Fortran programmers.
- Concurrent language.
- Mitrion software is closely tied to a specific version of the Xilinx compiler.
- Software maintenance and bug fixes present a big challenge.

Handel-C

- Developed by Celoxica.
- Only runs on Windows based PC.
- The Linux version has just been released, but there are problems with the release.

Handel-C (+)

Advantages:

- C-like syntax and constructions.
- Sequential programming with parallel constructors.
- Straightforward “translation” from ANSI C to Handel C.
- Easier than VHDL or Verilog.

Handel-C (-)

Disadvantages:

- Most HPC users are Fortran programmers.
- The Linux version has just been released, but with many problems.
- Temporary licenses for PCs available, but imply additional work to install and support the software.
- No support for Virtex-4.
- Poor support for the XD1

DSPLogic

- Based on Simulink, a sophisticated graphical interface to Matlab for modelling, simulation, and analysis of dynamical systems.
- Algorithms are implemented by dragging blocks from a library into the workspace, and establishing connections between them.

DSPLogic (+)

Advantages:

- Potential access to low level Xilinx primitives.
- Appears ideal for digital circuit design.
- Block abstraction and code encapsulation may be valuable for very large and complex reconfigurable codes.
- Good simulator and debugging tools inherited from Simulink.

DSPLogic (-)

Disadvantages:

- Only runs on a Windows-based PC.
- User needs to learn/buy Matlab/Simulink.
- Simple algorithms often require dozens of interconnecting blocks.

Presentation Outline

1. The NRL's Cray XD1System
2. Scientific Supercomputing
3. Reconfigurable Supercomputing
4. FPGA Programming Tools
5. **Performance Measurements**
6. Problems and Issues
7. Conclusions

Dual Core Efficiency

- The dual cores in the XD1's Opteron 275 share the same DDR memory controller as the single chip processor version.

- This sharing of memory bandwidth can lead to a degradation of the performance of the codes running on the dual cores chips.

A Measure of Efficiency

- We consider two scenarios:
  - A code running using n nodes and all 4 cores on the node takes T_4 time.
  - A code running using 2n nodes and only one core of each dual core processor takes T_2 time.
- We define the dual core efficiency as:

$$DCE = 100 \times \left(1 - \frac{T_4 - T_2}{T_2} \right)$$

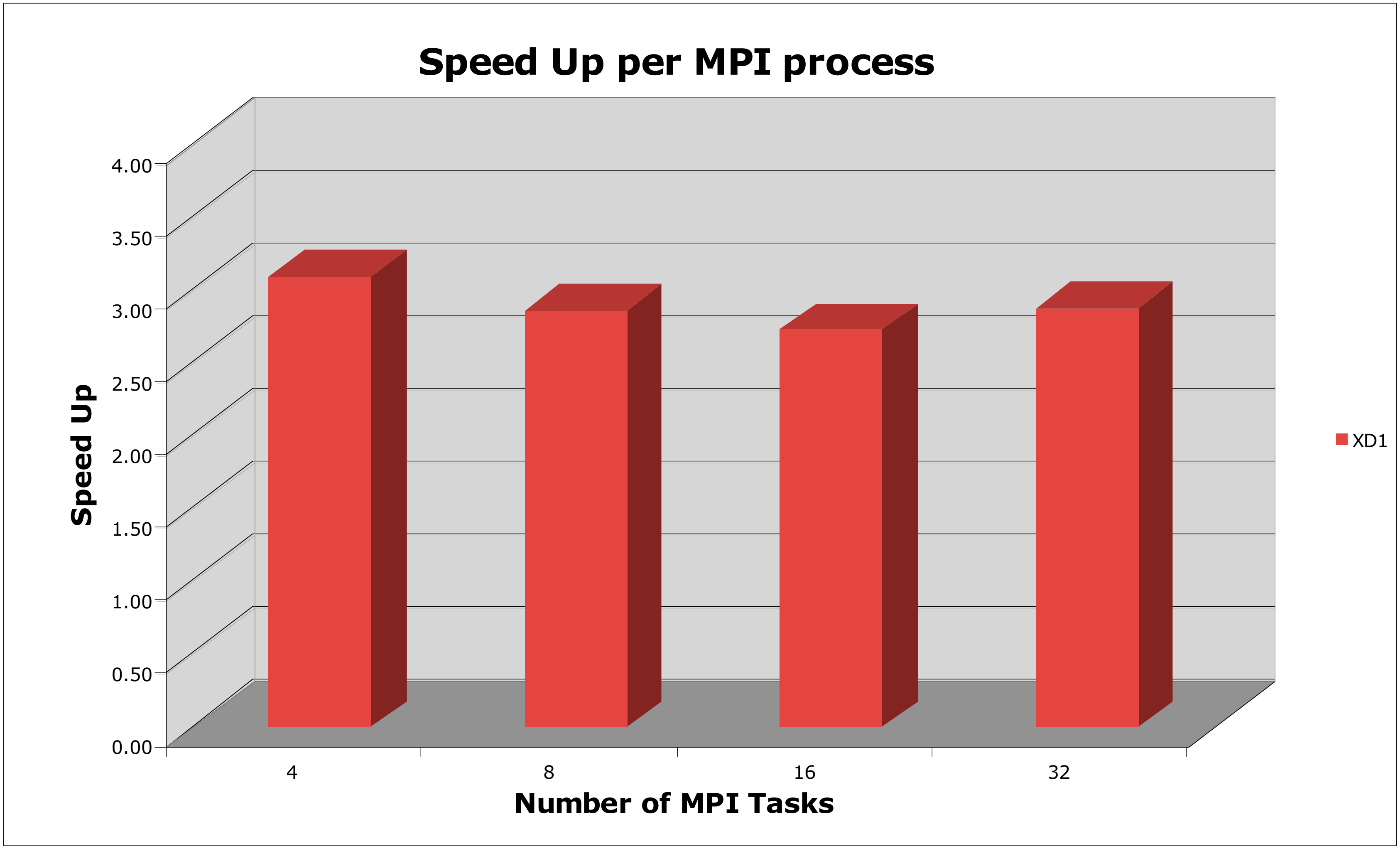
Dual Core Efficiency Results

Application	One Core	Both Cores	Efficiency %
STATIC	313	450	56
CAUSAL	275	293	93
LANCZOS	771	1371	22
NRLMOL	14283	16260	90
ARMS	2090	2524	79
NOZZLE	27498	27286	101
AVUS	1197	963	120
HYCOM	823	849	97
OOCORE	5274	7716	54
RFCTH2	279	448	39

Hybrid Codes

- MPI/OpenMP Hybrid Code
  - Is it more efficient than pure MPI code?
- Developed 3 versions of Causal Code
  - Pure MPI, Pure OpenMP, Hybrid MPI/OpenMP
- Performance
  - Pure MPI and Pure OpenMP had similar performance on 4 cores
  - Pure MPI code still outperformed Hybrid code

Hybrid Code Efficiency



Lustre Systems

- The Lustre system is a high speed parallel file system available to all nodes.
  - Not a mature technology

- We have recently upgraded the Lustre disk system, adding an additional controller and devoting 4 nodes to the running of Lustre (instead of 2).

Lustre I/O Rates

NODES	Read (MB/sec) old/new	Write (MB/sec) old/new
1	206/156	165/417
2	325/326	324/7821
4	629/630	646/1298
8	794/1224	709/1393
16	892/1460	862/1250
32	859/1420	893/1280

Presentation Outline

1. The NRL's Cray XD1System
2. Scientific Supercomputing
3. Reconfigurable Supercomputing
4. FPGA Programming Tools
5. Performance Measurements
6. **Problems and Issues**
7. Conclusions

A Few Problems

We have observed only a few significant issues with the XD1, even though our system is the largest one fielded by Cray.

- XD1 and Lustre file system interaction.
- MPI error messages

Disk Accesses

- Disk accesses were affected when programs were using most of the bandwidth to the Lustre nodes.
- A command to list the files in a directory could take as much as 5 minutes.
- Also the time to rebuild a RAID disk that failed would increase from 3 hrs (stand alone mode) to 3 days (with users).

Large Files I/O

- Some users have reported that their programs crash when writing large files to disk.
- This problem has proved to be very difficult to track down and reproduce, as it may take several days before the failure occurs.
- Tests performed by Cray appear to indicate a problem with GART on a node allocated to the job.

MPI Error Messages

- MPI error messages are misleading and completely useless for debugging purposes:

  `mpiexec:Error: read_rai_startup_ports: Failed to read barrier entry token from rank 3 process on node#"`

- Cray is currently working on **`mpiexec`** to provide more meaningful messages.

Presentation Outline

1. The NRL’s Cray XD1System
2. Scientific Supercomputing
3. Reconfigurable Supercomputing
4. FPGA Programming Tools
5. Performance Measurements
6. Problems and Issues
7. **Conclusions**

Conclusions

- The XD1 has proved to be popular at NRL.
  - Wide variety of scientific codes and applications.
- The development of reconfigurable codes remains a daunting task.
  - Usability of FPGA programming suites is by far the greatest challenge.