
Performance Evaluation of the Cray X1 Distributed Shared Memory Architecture

Tom Dunigan
Jeffrey Vetter
Pat Worley

Oak Ridge National Laboratory



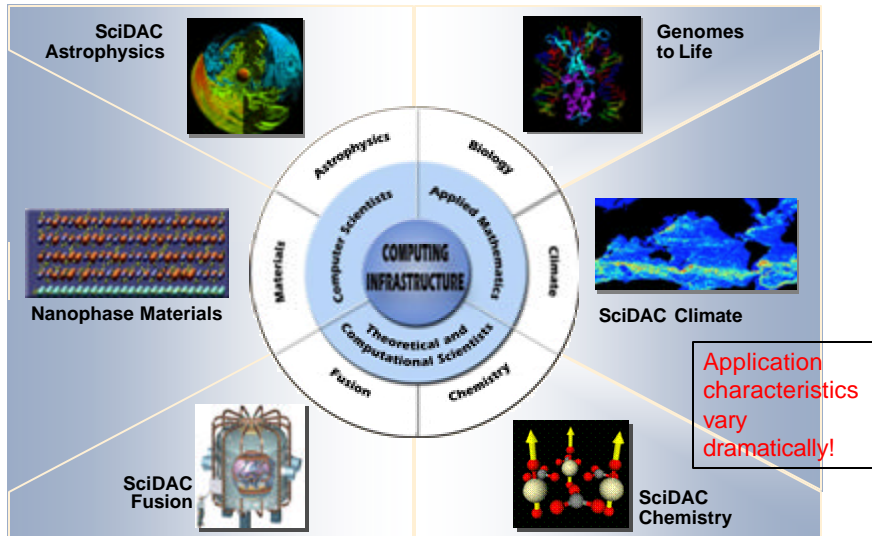
Highlights

- ➔ **Motivation**
 - Current application requirements exceed contemporary computing capabilities
 - Cray X1 offered a 'new' system balance

- ➔ **Cray X1 Architecture Overview**
 - Nodes architecture
 - Distributed shared memory interconnect
 - Programmer's view

- ➔ **Performance Evaluation**
 - Microbenchmarks pinpoint differences across architectures
 - Several applications show marked improvement

ORNL is Focused on Diverse, Grand Challenge Scientific Applications

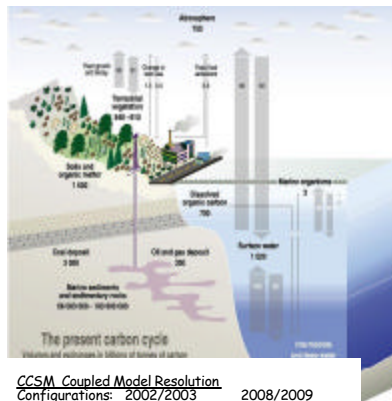


ORNL/JV

3

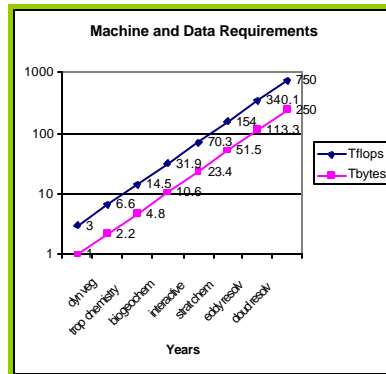
Climate Case Study: CCSM Simulation Resource Projections

Science drivers: regional detail / comprehensive model



CCSM Coupled Model Resolution

Configurations:	2002/2003	2008/2009
Atmosphere	230kmL26	30kmL96
Land	50km	5km
Ocean	100kmL40	10kmL80
Sea Ice	100km	10km
Model years/day	8	8
National Resource (dedicated TF)	3	750
Storage (TB/century)	1	250



- 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

At 2002-3 scientific complexity, a century simulation required 12.5 days.

ORNL/JV

4

Engaged in Technical Assessment of Diverse Architectures for our Applications

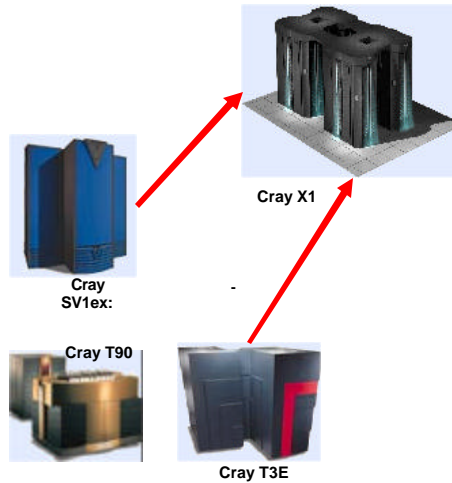
- Cray X1
- IBM SP3, p655, p690
- Intel Itanium, Xeon
- SGI Altix
- IBM POWER5
- FPGAs
- Planned assessments
 - Cray X1e
 - Cray X2
 - Cray Red Storm
 - IBM BlueGene/L
 - Optical processors
 - Processors-in-memory
 - Multithreading
 - Array processors, etc.



Cray X1 System Overview

Cray X1 Combines Features of Previous Cray Platforms

- ➔ T3E
 - MPP Interconnect
 - High bandwidth
 - Low Latency
 - Mesh Topology
 - Scalable
 - Scalable system software
- ➔ T90
 - Vector processors
 - High bandwidth memory subsystem
 - 2L and 1S per cycle/pipe
- ➔ SV1ex
 - Vector cache
 - Multistreaming processors
 - CMOS technology



ORNL/JV

7

Cray X1 @ ORNL

- ➔ Cray X1 @ ORNL
 - 512 processors (12.8 GF per processor)
 - 2 TB memory
 - 32 TB disk
 - 6.4 TeraOps Peak

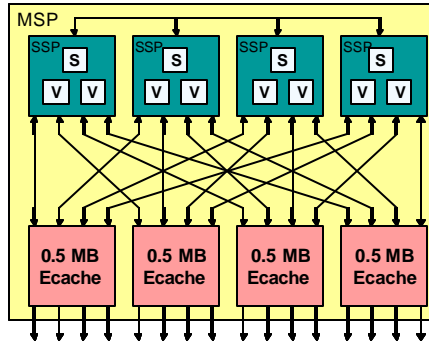


ORNL/JV

8

Multi-Streaming Processor (MSP) is the Basic Building Block for Cray X1

- 12.8 GF/s for 64-bit operations
 - 25.6 GF/s for 32-bit operations
- Comprised of four single-streaming processors (SSPs)
 - Two 32-stage 64-bit floating-point vector units
 - One 2-way super-scalar unit
 - Two clock frequencies
 - 800 MHz for the vector units
 - 400 MHz for the scalar unit
 - Each SSP is capable of 3.2 GF/s for 64-bit operations
 - Four SSPs share a 2 MB "Ecache"
- Vector load buffers permit 2048 outstanding mem ops

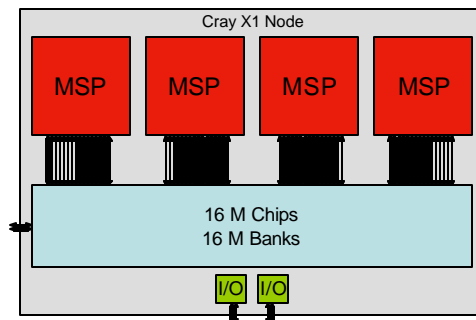


ORNL/JV

9

Four MSPs Connected to Form One Cray X1 Node

- Four MSPs (51 GFLOPS)
- 16 memory controller chips (M-chips)
- 32 memory daughter cards
- 200 GB/s node memory bandwidth
- Single node memory capacity is 8-32 GB
- Cache coherent memory inside node
- Uniform memory latency within node
- M chips service both local and remote memory requests



ORNL/JV

10

Interconnect Topology

- ➔ Small Systems
 - Quad-bristled hypercubes to 512 CPUs
- ➔ Larger systems
 - 2-D Modified Torus
 - 16 planes
 - ORNL configuration
- ➔ Remote loads and stores are handled transparently by the interconnect

Router

- ➔ Derived from SPIDER router
 - 8 Channels, 20b width per channel
 - 1.6 GB/s/channel
 - 12.8 GB/s/chip
 - 25ns latency
- ➔ Each X1 node has 32 of these 1.6 GB/s full duplex channels connecting it to other nodes

X1 Programmer's View

- ➔ Single Node
 - Vector
 - MSP-mode automatically by compiler + user-inserted Cray Streaming Directives (CSDs)
 - SSP-mode
 - OpenMP, pthreads
- ➔ Distributed memory
 - Message Passing Interface
 - Unified Parallel C
 - Co-Array Fortran
 - Cray SHMEM

X1 Naturally Supports Global Address Space Programming Models

- ➔ Globally distributed hardware shared memory
 - References outside of node converted to non-allocate
- ➔ Natural fit for globally addressable languages
 - Unified Parallel C
 - Co-Array FORTRAN
 - Compiler can schedule remote references

Microbenchmark Performance

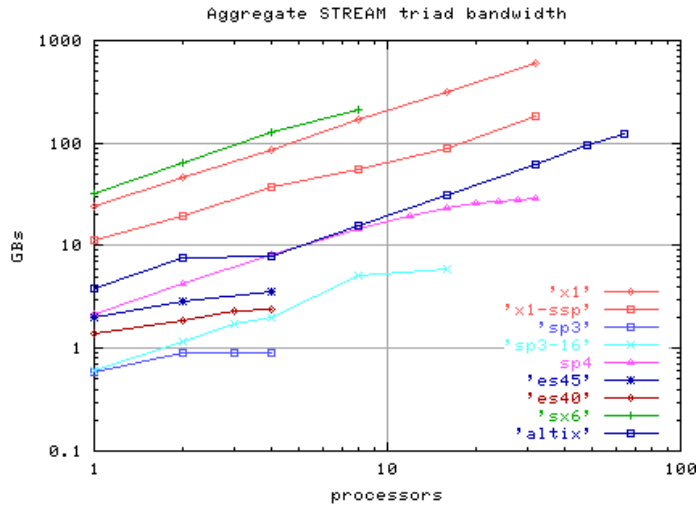
1. P.A. Agarwal, R.A. Alexander *et al.*, "Cray X1 Evaluation Status Report," ORNL, Oak Ridge, TN, Technical Report ORNL/TM-2004/13, 2004, <http://www.csm.ornl.gov/evaluation/PHOENIX/PDF/CRAYEvaluationTM2004-15.pdf>.
2. T.H. Dunigan, Jr., M.R. Fahey *et al.*, "Early Evaluation of the Cray X1," Proc. ACM/IEEE Conference on High Performance Networking and Computing (SC03), 2003.

Evaluation Platforms

	SGI Altix	Alpha SC	IBM SP3	IBM SP4	Cray X1
Proc	Itanium 2	Alpha EV67	POWER3-II	POWER4	Cray X1
Interconnect	Numalink	Quadrics	Colony	Colony	Cray X1
MHz	1500	667	375	1300	800
Mem/Node	512GB	2GB	2GB	32GB	16GB
L1	32K	64K	64K	32K	16K (scalar)
L2	256K	8MB	8MB	1.5MB	2MB (per MSP)
L3	6MB	n/a	n/a	128MB	n/a
Proc Peak Mflops	6000	1334	1500	5200	12800
Peak mem BW	6.4 GBs	5.2GBs	1.6GBs	51 GBs/MCM	26 GBs/MSP

Includes recent results for IBM SP4 w/ Federation Interconnect.

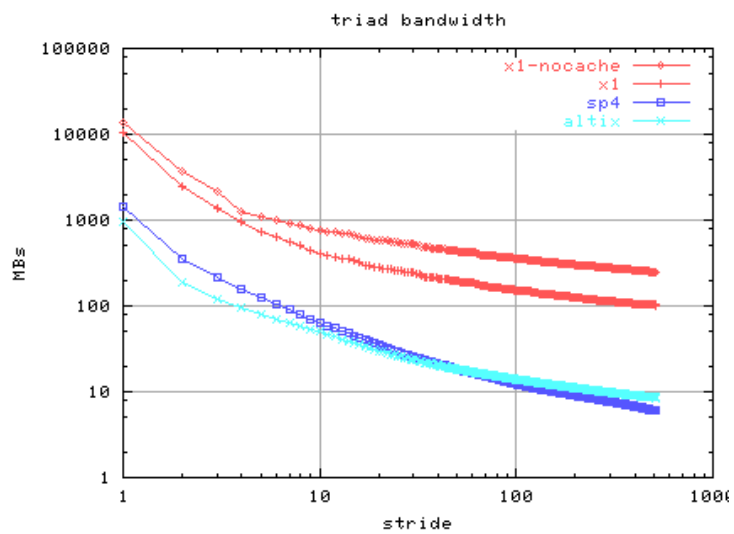
STREAM Triad Aggregate Shared Memory Bandwidth for X1 Node



ORNL/JV

17

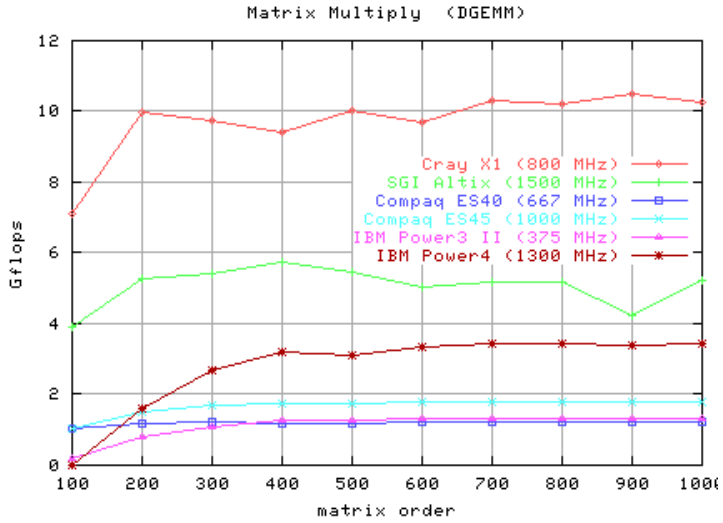
Triad Memory Bandwidth by Stride



ORNL/JV

18

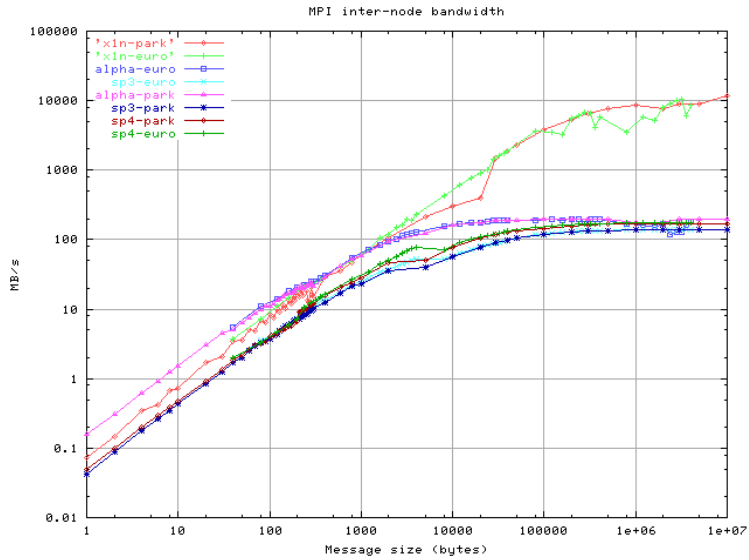
Performance of Matrix Multiply



ORNL/JV

19

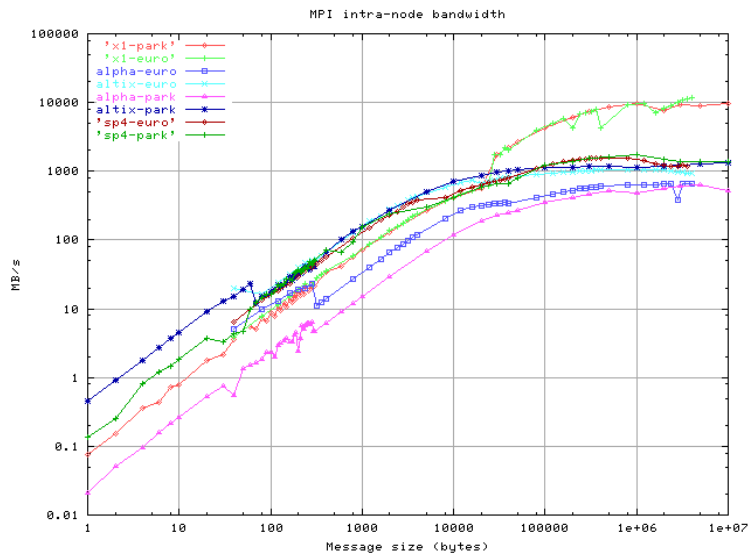
MPI Inter-node Bandwidth



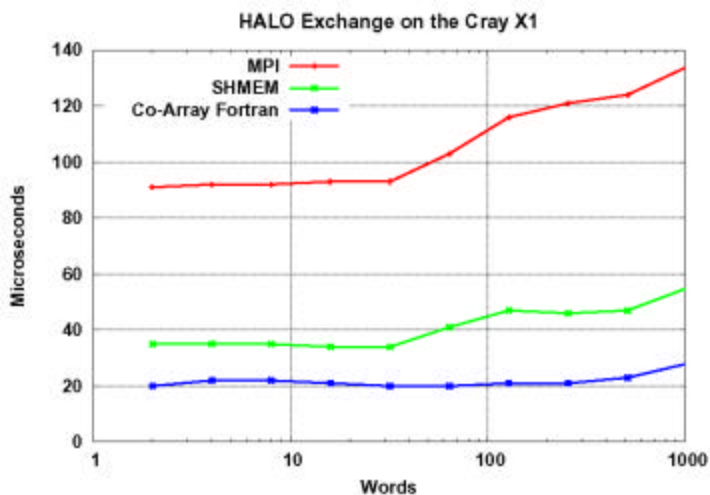
ORNL/JV

20

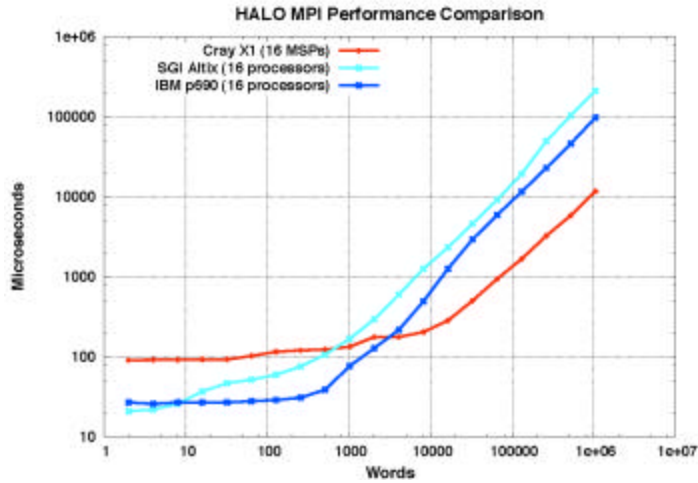
MPI Intra-node Bandwidth



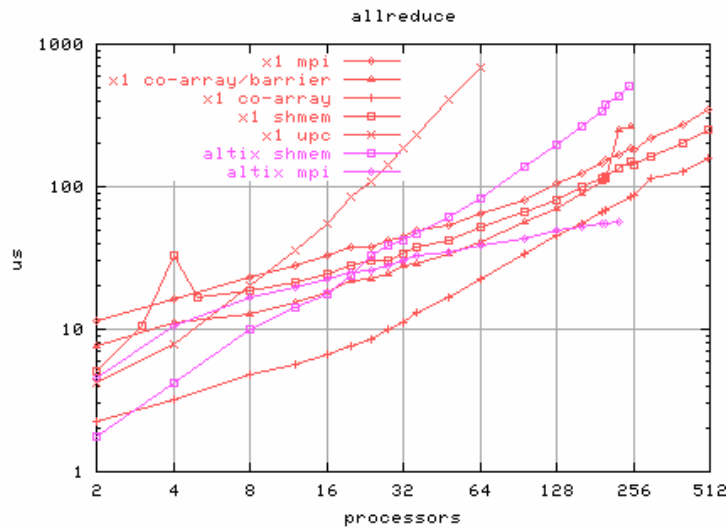
Halo Exchange Bandwidth on X1



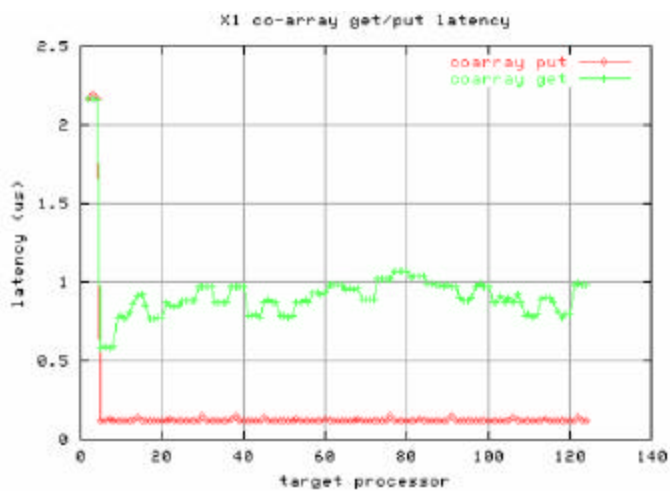
Halo Exchange MPI Bandwidth across Architectures



MPI_Allreduce Latency



Co-array Remote Access Latencies



ORNL/JV

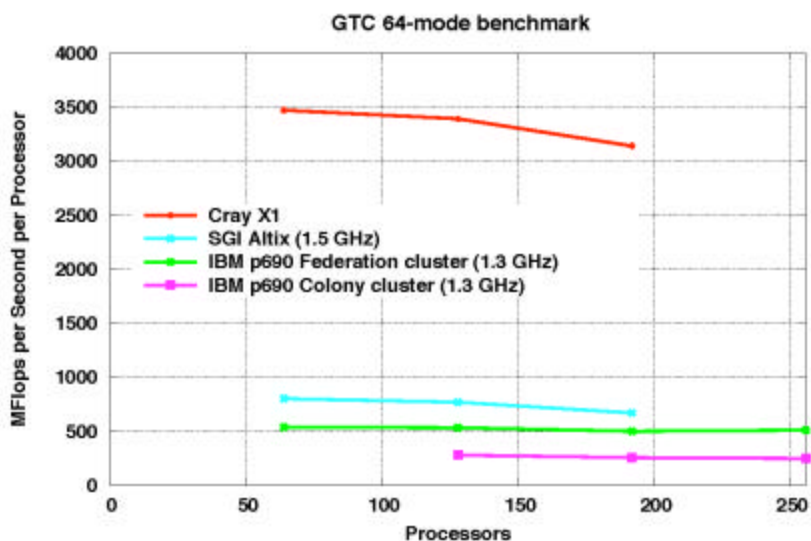
25

Application Performance

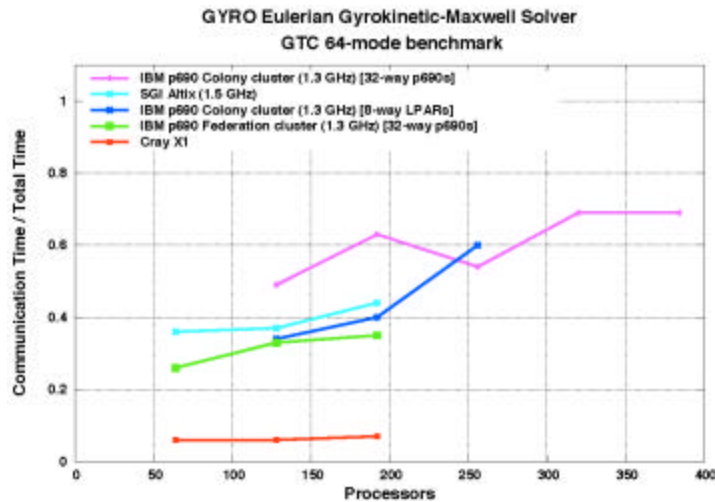
Gyro / Fusion

- ➔ Used to study plasma microturbulence in fusion research
- ➔ Is a Eulerian gyrokinetic-Maxwell solver developed by R.E. Waltz and J. Candy at General Atomics
- ➔ Uses the `MPI_ALLTOALLV` command to transpose the distributed data structures

Gyro Performance / Fusion



Gyro Communication Fraction



ORNL/JV

29

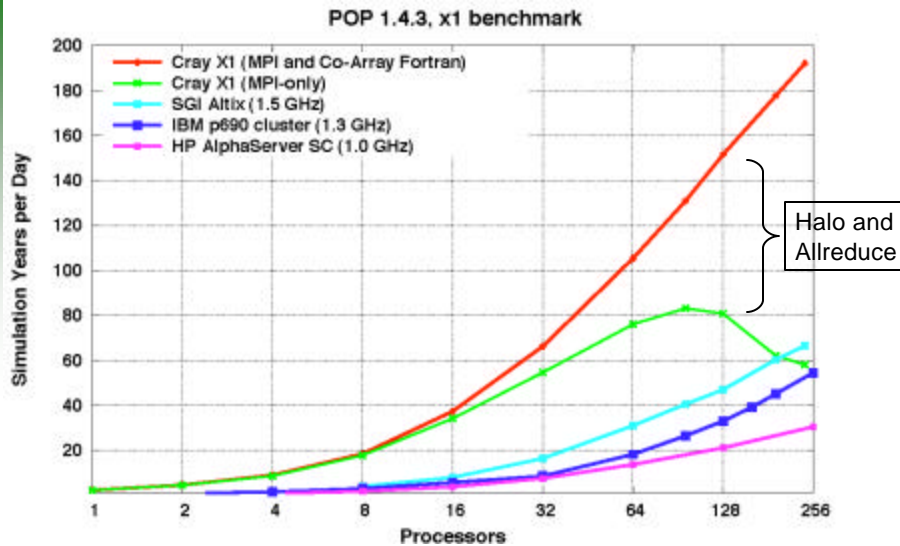
Parallel Ocean Program (POP) / Climate

- ➔ Is an ocean modeling code that is used as the ocean component in the Community System Climate Model (CCSM) coupled climate model
 - Baroclinic: 3D with limited nearest-neighbor communication; scales well
 - Barotropic: dominated by solution of 2D implicit system using conjugate gradient solves; scales poorly
 - Fixed size benchmark problem
- ➔ With communication dominated by a halo update and a `MPI_ALLREDUCE`
 - used in calculating residuals and inner products in a conjugate gradient linear system solve
- ➔ Domain decomposition determined by grid size and 2D virtual processor grid.

ORNL/JV

30

Parallel Ocean Program (POP) Performance / Climate



ORNL/JV

31

Application Summary

➤ Gyro

- Performance on non-vector systems is constrained by communication bandwidth
 - This is not true on the Cray X1
- There is also a phase of computation that does not yet scale beyond 64-way parallelism.
 - This has not limited performance on the non-vector systems, but it does limit performance on the X1

➤ POP

- Scalability is determined by communication latency
- Co-Array Fortran performance is excellent
 - Algorithms currently used for global reduction are not scalable
- MPI short message and collective performance is mediocre
 - MPI collectives should perform as well as the Co-Array Fortran implementation, and we expect the performance need for Co-Array Fortran to diminish in the near future

ORNL/JV

32

Conclusions

- ➔ **Motivation**
 - Current application requirements exceed contemporary computing capabilities
 - Cray X1 offered a 'new' system balance

- ➔ **Cray X1 Architecture Overview**
 - Nodes
 - Distributed Shared Memory Interconnect
 - Programmer's view

- ➔ **Performance Evaluation**
 - Microbenchmarks pinpoint differences across architectures
 - High memory bandwidth and interconnect bandwidth
 - Several applications show striking improvement

Acknowledgements

- ➔ These slides have been authored by a contractor of the U.S. Government under contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes

- ➔ Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the United States Department of Energy under Contract No. DE-AC05-00OR22725.