Petascale Opportunities and Challenges for Earth System Modeling

Presented to the 14th ECMWF Workshop on the use of HPC in Meteorology

2 November 2010

Per Nyberg
Director, Marketing and Business Development
Earth Sciences Segment
nyberg@cray.com

Image Courtesy of Jamison Daniel, National Center for Computational Sciences, Oak Ridge National Laboratory.
Simulation of CCSM3 at T341 resolution on ORNL Cray XT5

Topics

- Cray's Presence in the Earth System Modeling Community
  - Recent Customer Updates
- State-of-the-Art Modeling on Cray Systems
- Extreme Scale Challenges for Earth System Modeling
- Cray XE6 Technology Update

Cray's Presence and Experience in the Earth System Modeling Community

- Earth System Modeling (ESM) represents a significant portion of the computing done on Cray Systems worldwide:
  - Dedicated operational and research centers.
  - Multi-disciplinary research centers.
  - From Teraflops to Petaflops.
  - NOAA/ORNL system is the largest in the world dedicated to climate research.
  - KMA will be one of the largest operational NWP systems in the world in early 2011.
- Cray Petascale systems have been key in:
  - Enabling transformational science
  - As development platforms for preparing climate and weather models for extreme scale capabilities.

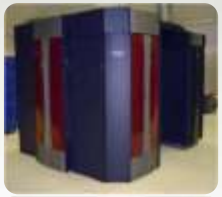
Leading Weather and Climate Centers Worldwide

Brazilian Center for Weather Forecasts and Climate Studies (CPTEC)

CPTEC

dmi

Danish Meteorological Institute



AEMet
Agencia Estatal de Meteorología

KMA

Korea Meteorological Administration

NOAA Climate Modeling and Research System GFDL/NCEP/ESRL

noaaresearch

University of Bergen

Bergen Center for Computational Science

Naval Oceanographic Office

Meteo Swiss

Finnish Meteorological Institute

NCAR

National Center for Atmospheric Research

CRAY
THE SUPERCOMPUTER COMPANY

Leading Research at Multi-disciplinary National Leadership Centers

OAK RIDGE
National Laboratory

Oak Ridge
Climate Change Science Institute

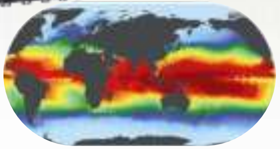
NOAA

NASA

Los Alamos
NATIONAL LABORATORY
EST. 1943

Sandia National Laboratories

Climate, Ocean and Sea Ice Modeling (COSIM)



UK Engineering and Physical Sciences Research Council

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

HECToR

CSC

UNIVERSITY OF HELSINKI

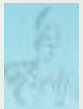
CSC Finland IT Center for Science

National Institute of Computational Sciences

NSF

CAPS

NCAR



Project Athena

NERSC

BERKELEY LAB

ESD
EARTH SCIENCES DIVISION

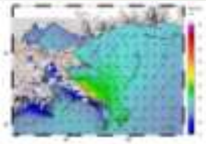
SCRIPPS INSTITUTION OF OCEANOGRAPHY
GLOBAL DISCOVERIES FOR TOMORROW'S WORLD

NOAA

CSCS
Swiss National Supercomputing Centre

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

C2SM
Center for Climate Systems Modeling



U.S. Army Engineer Research and Development Center

* > Petaflop Systems

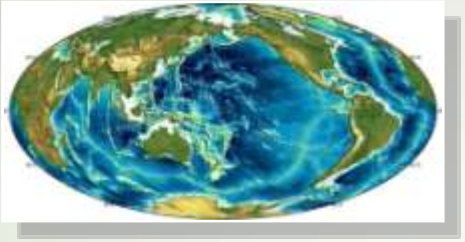
NOAA and ORNL Climate Modeling and Research System

- Systems to be delivered to Oak Ridge National Laboratory (ORNL) for use by National Oceanic and Atmospheric Administration (NOAA) and ORNL for advanced climate modeling and research
- Multi-phase, multi-year contract.
- At each phase the system will be the largest in the world dedicated to climate modeling.
- First phase installed and in production.
- Final phase system in 2011 will exceed 1 Petaflop.

Korea Meteorological Administration

KMA KOREA METEOROLOGICAL ADMINISTRATION

- >$40M five year contract for fully integrated capabilities:
  - Dual operational systems with failover capability
  - Multi-tier, multi-Petabyte global, centralized storage
  - Data management (Backup, archive, virtual tape library)
  - Pre/post and login servers
  - LAN and WAN networking
  - Control centre
- Two phase delivery with final system of 754 Tflops operational by early 2011.
- Installation in newly constructed KMA National Meteorological Supercomputer Center in Oh-Chang.
- Key capability in KMA's transition to new operational Unified Model based NWP and climate suite.
- Continuation of the KMA-Cray Earth System Research Center.



UM NWP operational system

3rd SC Facility

KMA NMSC

Phase 2 XE6 Systems Installed

KMA KOREA METEOROLOGICAL ADMINISTRATION

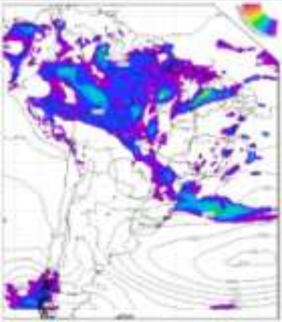
Phase 1 Cray XT5

Cray Supplied Control Room

Phase 2 Cray XE6 Systems (one system in each hall)

INPE/CPTEC Brazil

- National Institute for Space Research (INPE) Center for Weather Forecasts and Climate Studies (CPTEC) is the national weather service of Brazil.
- Mission to provide Brazil with studies on climate change and state-of-the-art weather, seasonal climate and environmental forecasts.
- Cray XT6 with a peak performance of ~250 TF.
- Installation is ongoing.



Courtesy: INPE/CPTEC

XT5m “mini” Wins - Finnish Meteorological Institute and National Center for Atmospheric Research

NCAR fmi.fi

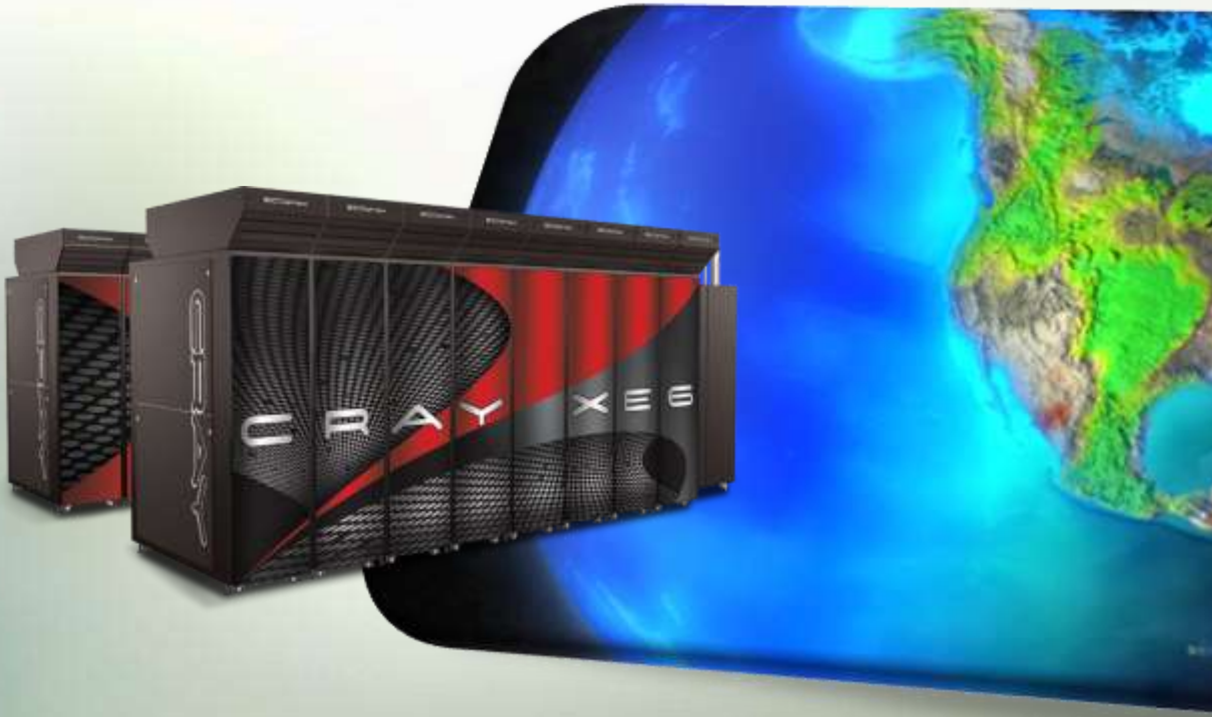
- FMI
  - Two identical Cray XT5m systems (total 4 cabinets)
  - One for operational NWP
  - One for research
  - Installed in Sept 2009
  - Peak Performance ~34 TFLOPS

- NCAR
  - XT5m system installed in April 2010.
  - Testing of Cray technologies within NCAR environment.
  - Development platform for NCAR community who use NERSC, ORNL and NSF Cray XT systems.
  - Usage for special projects such as high resolution regional climate modeling.

University of Stuttgart High Performance Computing Center Stuttgart (HLRS)

- >$60M contract signed on 26 October 2010.
- Multi-year, multi-phase contract .
- Includes the delivery of a Cray XE6 and the future delivery of Cray's next-generation "Cascade".
- Scientific users from all disciplines.
- Large focus on engineering with industrial users from Automotive and Aerospace industry (Porsche, Daimler, …)

State-of-the-Art Earth System Modeling on Cray Systems



With Petascale Capabilities Global GCMs are Becoming Policy Relevant Applications Tools

Fully coupled biogeochemistry-physical climate simulation on Jaguar: David Erickson (ORNL) and Steven Pawson (NASA / GMAO / GSFC)

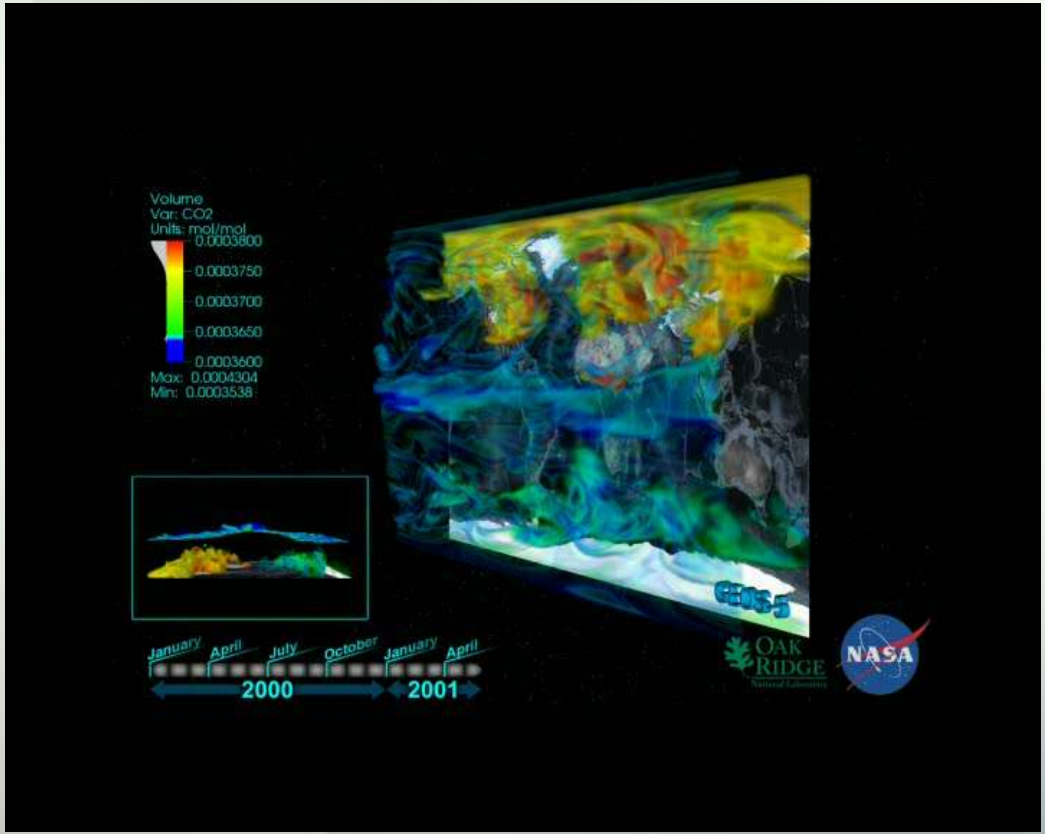


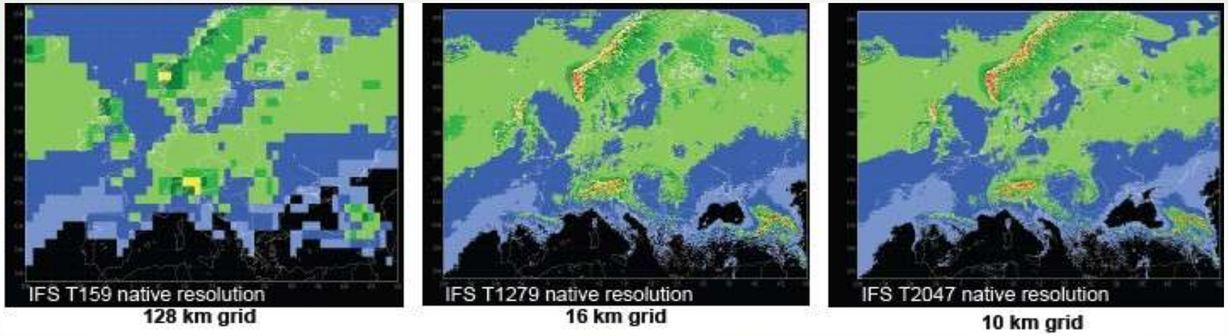
Image Courtesy of Jamison Daniel, ORNL

- Petascale capabilities begin to offer the ability to provide actionable insights to facilitate reliable decision-making for regional, national and global priorities.
- Example: International accords will require the need to accurately estimate greenhouse gas emissions and monitor their changes over time.

Project Athena

- The World Modeling Summit (WMS) in May 2008 called for a revolution in climate modeling to more rapidly advance improvement in climate model resolution, accuracy and reliability.

“Routine” atmospheric resolution



IFS T159 native resolution
128 km grid

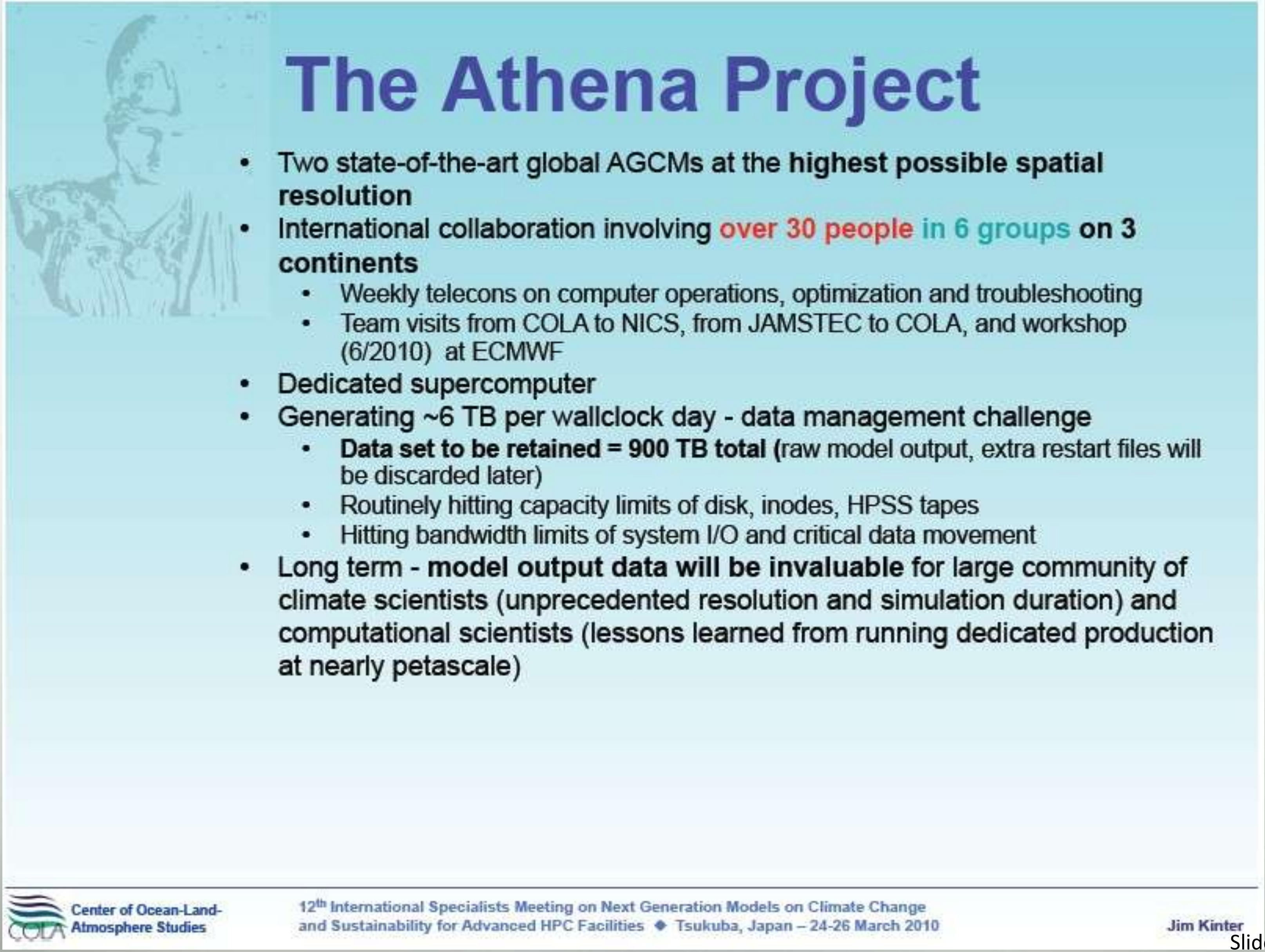
IFS T1279 native resolution
16 km grid

IFS T2047 native resolution
10 km grid

Desired “routine” atmospheric resolution

- The WMS recommended petascale supercomputers dedicated to climate modeling to provide:
  - Sufficient computational capability
  - Controlled environment to support long runs
  - Management and analysis of very large (petabyte) data sets.
- The NSF recognized the importance of the problem and offered to dedicate the NICS XT4 “Athena” system over a six-month period in 2009-2010 as a resource to meet the challenge.
- An international collaboration was formed among groups in the U.S., Japan and the U.K.…

International High Resolution Climate Modeling Project



The Athena Project

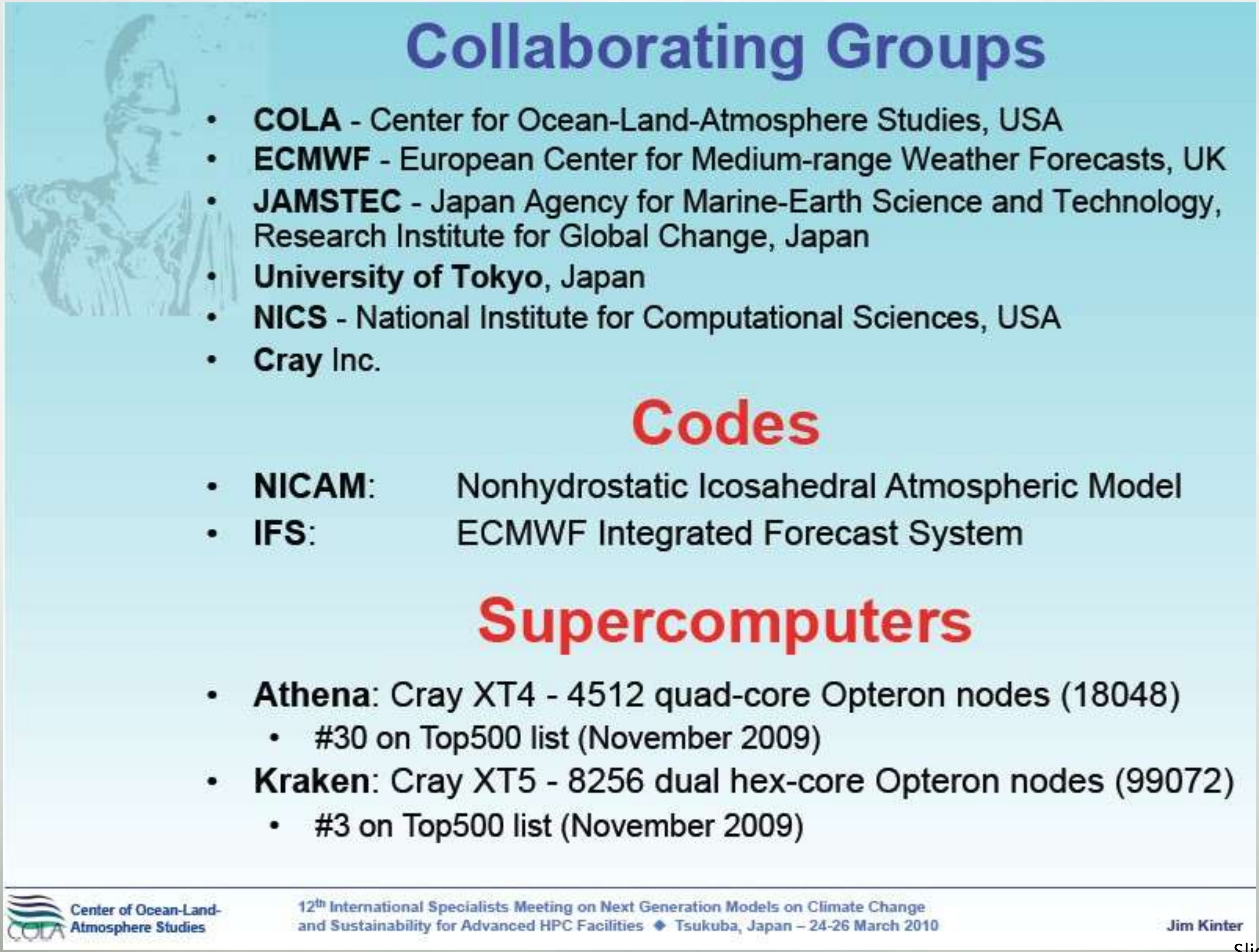
- Two state-of-the-art global AGCMs at the **highest possible spatial resolution**
- International collaboration involving over 30 people in 6 groups **on 3 continents**
  - Weekly telecons on computer operations, optimization and troubleshooting
  - Team visits from COLA to NICS, from JAMSTEC to COLA, and workshop (6/2010) at ECMWF
- Dedicated supercomputer
- Generating ~6 TB per wallclock day - data management challenge
  - **Data set to be retained = 900 TB total (**raw model output, extra restart files will be discarded later)
  - Routinely hitting capacity limits of disk, inodes, HPSS tapes
  - Hitting bandwidth limits of system I/O and critical data movement
- Long term - **model output data will be invaluable** for large community of climate scientists (unprecedented resolution and simulation duration) and computational scientists (lessons learned from running dedicated production at nearly petascale)

Center of Ocean-Land-Atmosphere Studies

12th International Specialists Meeting on Next Generation Models on Climate Change and Sustainability for Advanced HPC Facilities ◆ Tsukuba, Japan – 24-26 March 2010

Jim Kinter

International High Resolution Climate Modeling Project



Collaborating Groups

- **COLA** - Center for Ocean-Land-Atmosphere Studies, USA
- **ECMWF** - European Center for Medium-range Weather Forecasts, UK
- **JAMSTEC** - Japan Agency for Marine-Earth Science and Technology, Research Institute for Global Change, Japan
- **University of Tokyo**, Japan
- **NICS** - National Institute for Computational Sciences, USA
- **Cray** Inc.

Codes

- **NICAM**: Nonhydrostatic Icosahedral Atmospheric Model
- **IFS**: ECMWF Integrated Forecast System

Supercomputers

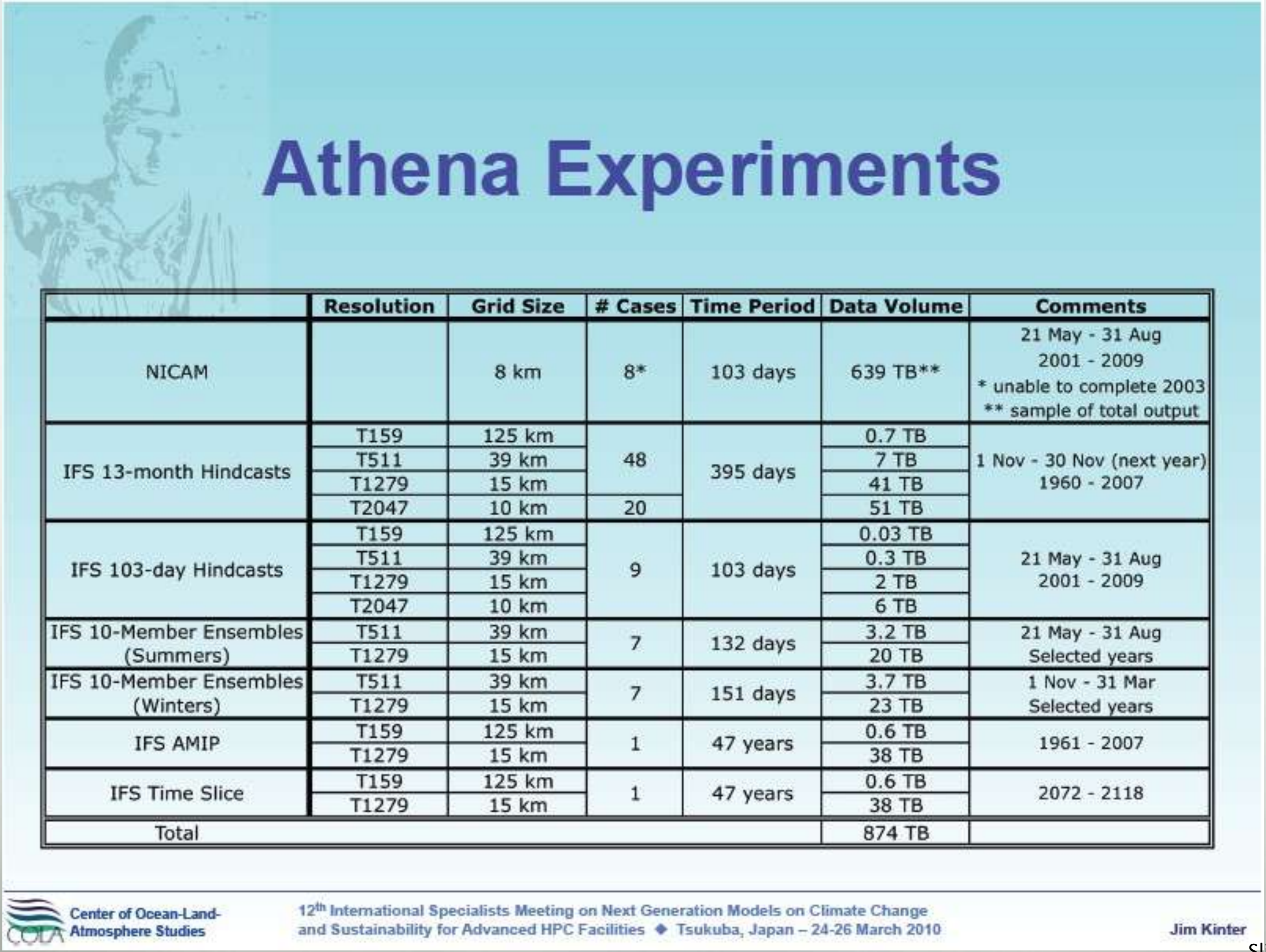
- **Athena**: Cray XT4 - 4512 quad-core Opteron nodes (18048)
  - #30 on Top500 list (November 2009)
- **Kraken**: Cray XT5 - 8256 dual hex-core Opteron nodes (99072)
  - #3 on Top500 list (November 2009)

Center of Ocean-Land-Atmosphere Studies

12th International Specialists Meeting on Next Generation Models on Climate Change and Sustainability for Advanced HPC Facilities ◆ Tsukuba, Japan – 24-26 March 2010

Jim Kinter

International High Resolution Climate Modeling Project



Athena Experiments

	Resolution	Grid Size	# Cases	Time Period	Data Volume	Comments
NICAM		8 km	8*	103 days	639 TB**	21 May - 31 Aug 2001 - 2009 * unable to complete 2003 ** sample of total output
IFS 13-month Hindcasts	T159	125 km	48	395 days	0.7 TB	1 Nov - 30 Nov (next year) 1960 - 2007
	T511	39 km			7 TB	
	T1279	15 km			41 TB	
	T2047	10 km	20		51 TB	
IFS 103-day Hindcasts	T159	125 km	9	103 days	0.03 TB	21 May - 31 Aug 2001 - 2009
	T511	39 km			0.3 TB	
	T1279	15 km			2 TB	
	T2047	10 km			6 TB	
IFS 10-Member Ensembles (Summers)	T511	39 km	7	132 days	3.2 TB	21 May - 31 Aug Selected years
	T1279	15 km			20 TB	
IFS 10-Member Ensembles (Winters)	T511	39 km	7	151 days	3.7 TB	1 Nov - 31 Mar Selected years
	T1279	15 km			23 TB	
IFS AMIP	T159	125 km	1	47 years	0.6 TB	1961 - 2007
	T1279	15 km			38 TB	
IFS Time Slice	T159	125 km	1	47 years	0.6 TB	2072 - 2118
	T1279	15 km			38 TB	
Total					874 TB	

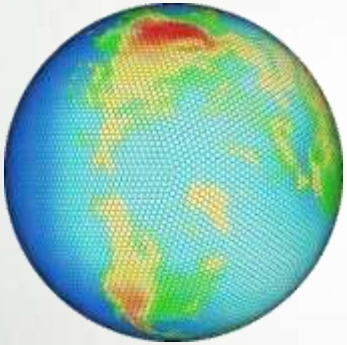
Center of Ocean-Land-Atmosphere Studies

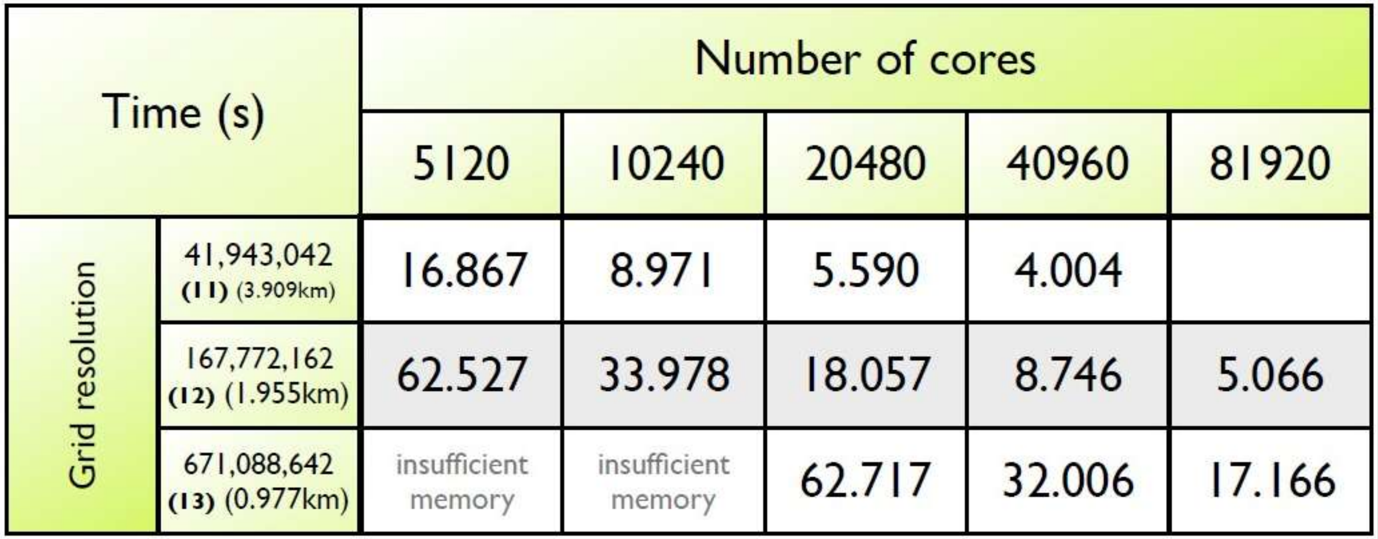
12th International Specialists Meeting on Next Generation Models on Climate Change and Sustainability for Advanced HPC Facilities ◆ Tsukuba, Japan – 24-26 March 2010

Jim Kinter

Global Cloud Resolving Model Development

- David Randall (Colorado State University) global cloud resolving model development using a geodesic grid.
- Development work is being done primarily on the NERSC Cray XT4.
- The model scaled to 80,000 processors on ORNL Jaguar Cray XT5 at resolution of 0.977km.

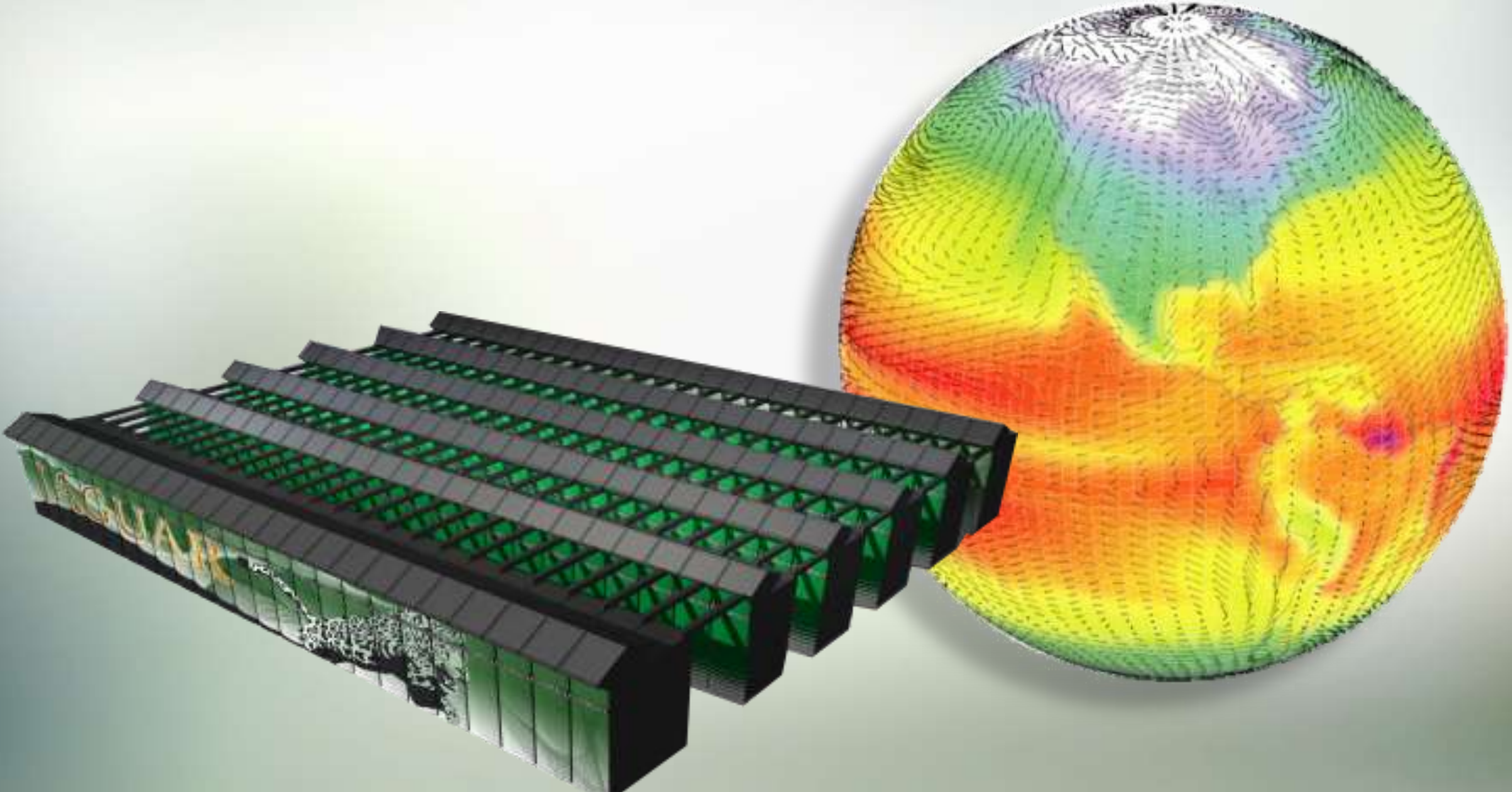




Time (s)		Number of cores				
		5120	10240	20480	40960	81920
Grid resolution	41,943,042 (11) (3.909km)	16.867	8.971	5.590	4.004	
	167,772,162 (12) (1.955km)	62.527	33.978	18.057	8.746	5.066
	671,088,642 (13) (0.977km)	insufficient memory	insufficient memory	62.717	32.006	17.166

Scaling of 3D-multigrid on Jaguar XT5 (20 V-cycles, 128 layers)
Courtesy: Ross Heikes, CSU

Extreme Scale Challenges for Earth System Modeling



Extreme Scale Challenges for Earth System Modeling

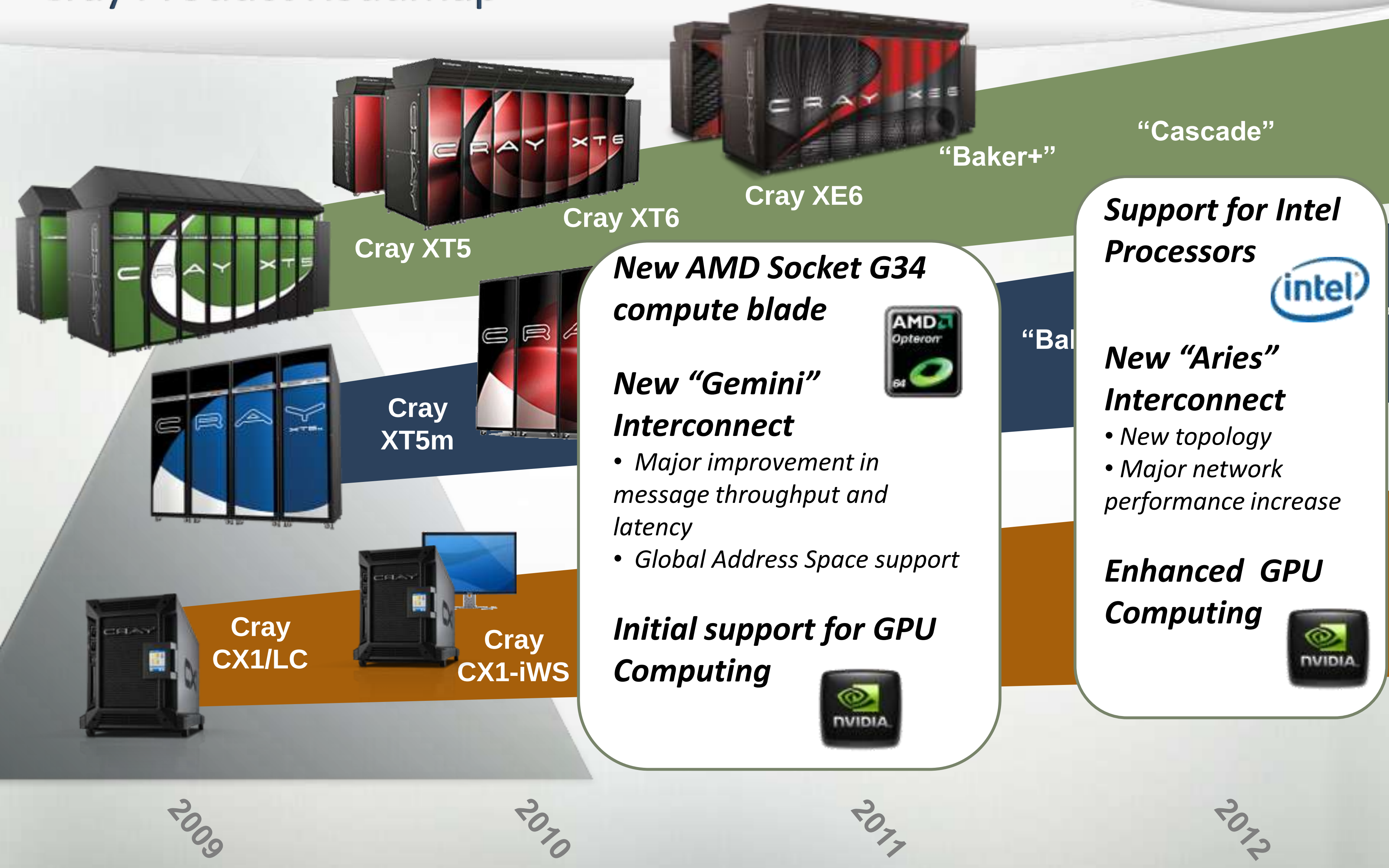
- **A collection of interrelated science and technology challenges.**
- **ESMs have become extremely complex multi-scale, multi-physics applications:**
  - Each with 100’s of person-years of scientific and software engineering investment.
  - Concern that a disruptive shift in hardware technology in the next 5–10 years that could require a complete change in the approach to data analysis, programmability, and interactive computing.
- **Petascale is not routine yet for many models.**
  - There also remains a large number of models and application areas that have not yet reached even the Terascale level.
- **Those that can scale have benefited from a focused, iterative multi-year algorithmic optimization effort:**
  - Optimization strategies do not remain stagnant and must take advantage of evolving hardware and software technologies.
  - Ongoing access to scalable, leadership class systems and support is essential.

Extreme Scale Challenges for Earth System Modeling

- **Persistent tension between programmability, portability, performance, resiliency and the unknown.**
- **Challenges include:**
  - Concern over long-term viability of current programming models (ie: MPI+Fortran) and implementation of new ones currently undetermined.
  - Fault tolerance and resiliency strategies:
    - Both to survive runtime errors and to reduce likelihood of undetectable errors that could compromise the value of large data sets.
  - Greater emphasis on resource conscious programming.
  - Applicability of accelerator technologies (eg: GPUs).
    - A potential disruptive technology.
    - Programming challenges with a potential to disrupt science progress.
    - Needs and objectives of operational centers and research centers can differ greatly.
  - Overall data management and better leverage of tens of $B investment in creating observations and model data sets.

Cray Technology Directions

Cray Product Roadmap



New AMD Socket G34 compute blade

New “Gemini” Interconnect

- *Major improvement in message throughput and latency*
- *Global Address Space support*

Initial support for GPU Computing

Support for Intel Processors

New “Aries” Interconnect

- *New topology*
- *Major network performance increase*

Enhanced GPU Computing

… Realizing our Adaptive Supercomputing Vision

The Cray XE6

Scalable Performance

Gemini Interconnect for Multicore era

CLE3.x with ESM

Sustained Petaflops

1M+ cores

Improved Msg. Latency

Production Efficiency

ECOphlex Cooling

Network Resiliency

Warm Swap Blades

NodeKARE

Can Upgrade XT5/6

Adaptive Supercomputing

CLE3.x with CCM

X86/Linux Env.

Mature Software Ecosystem

Multiple File Systems

CLE3, An Adaptive Linux OS designed specifically for HPC

CRAY LINUX ENVIRONMENT CLE3

ESM – *Extreme Scalability Mode*

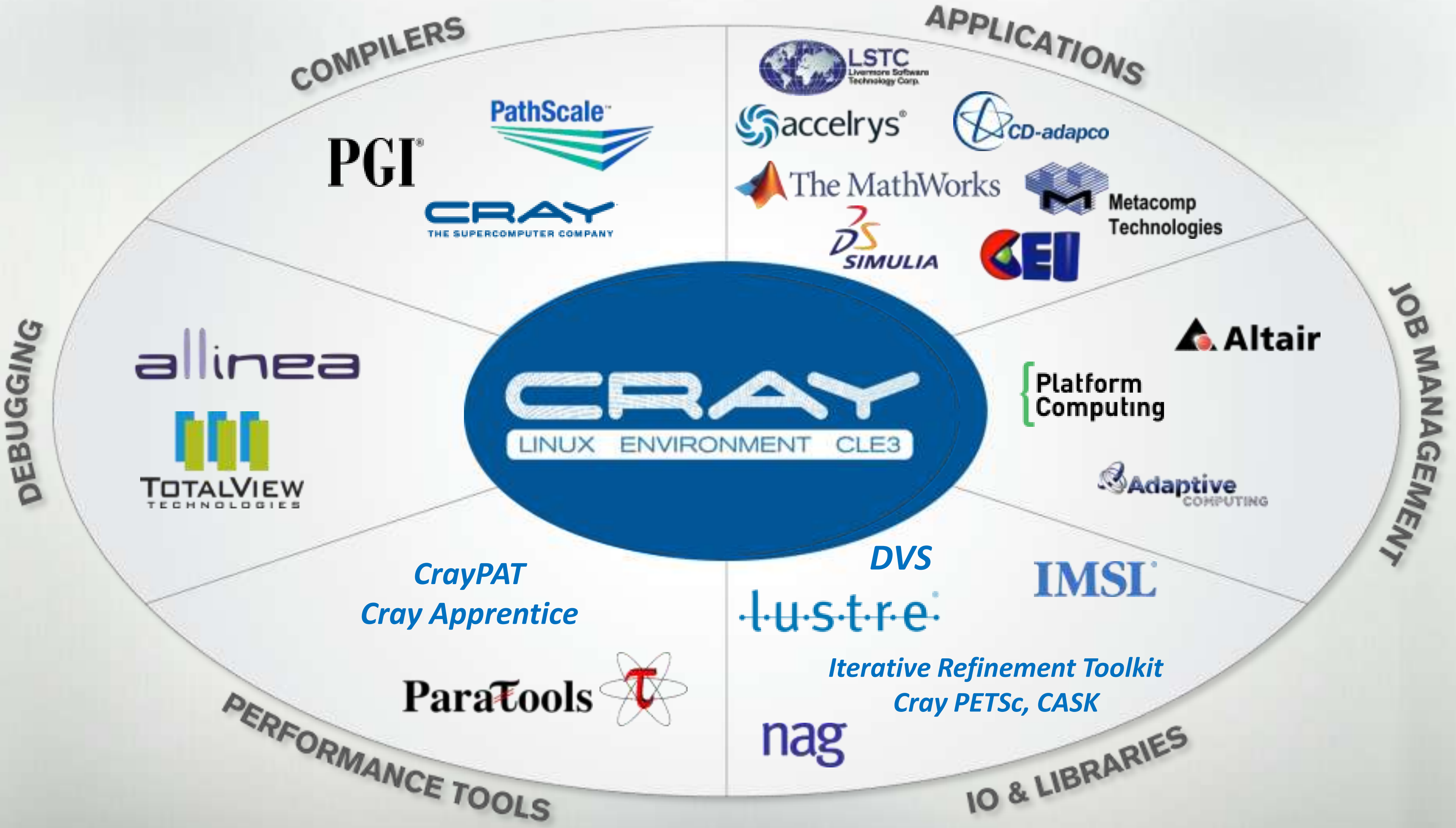
- No compromise *scalability*
- Low-Noise Kernel for scalability
- Native Comm. & Optimized MPI
- Application-specific performance tuning and scaling

CCM –*Cluster Compatibility Mode*

- No compromise *compatibility*
- Fully standard x86/Linux
- Standardized Communication Layer
- Out-of-the-box ISV Installation
- ISV applications simply install and run

CLE3 run mode is set by the user on a job-by-job basis to provide full flexibility

Cray Software Ecosystem



COMPILERS

PGI
PathScale
CRAY THE SUPERCOMPUTER COMPANY

APPLICATIONS

LSTC Livermore Software Technology Corp.
accelrys
CD-adapco
The MathWorks
Metacomp Technologies
SIMULIA
CEI

JOB MANAGEMENT

Altair
Platform Computing
Adaptive COMPUTING

IO & LIBRARIES

DVS
lustre
IMSL
Iterative Refinement Toolkit
Cray PETSc, CASK
nag

PERFORMANCE TOOLS

CrayPAT
Cray Apprentice
ParaTools

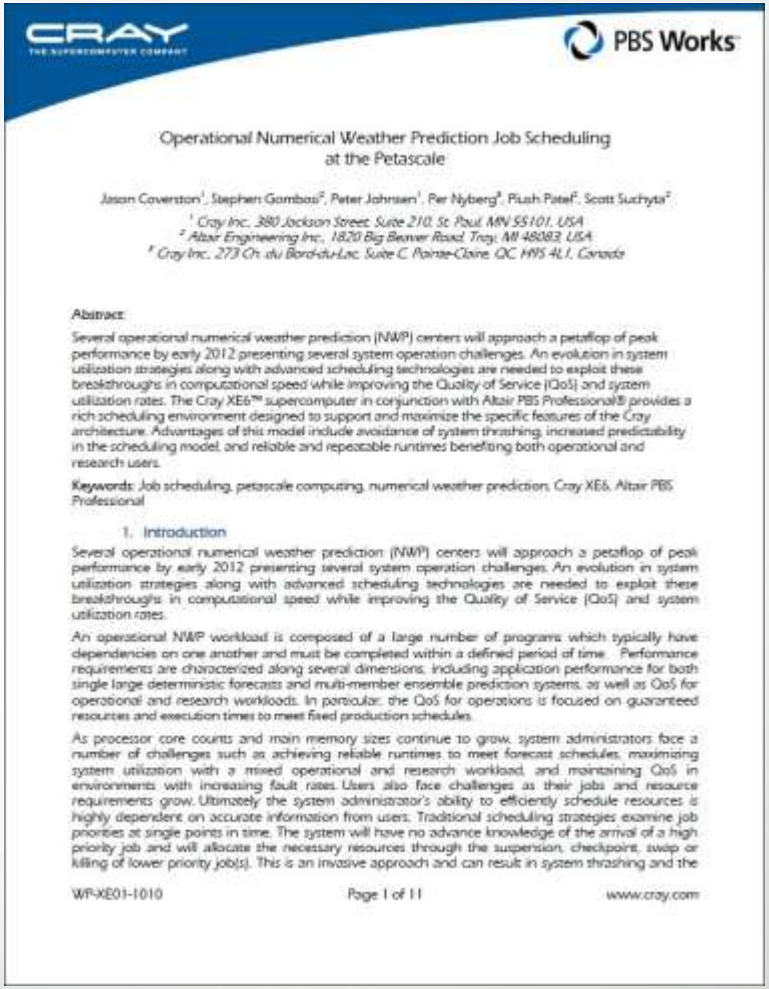
DEBUGGING

allinea
TOTALVIEW TECHNOLOGIES

CRAY LINUX ENVIRONMENT CLE3

NWP Job Scheduling

- The Cray XE6 provides a rich scheduling environment that is designed to support and maximize the specific features of the architecture.
- Fundamental scheduling strategy is to avoid the possibility of system thrashing and process level intervention, providing:
  - Increased predictability in the scheduling model
  - Reliable and repeatable runtimes
  - Maximum system efficiency



Operational Numerical Weather Prediction Job Scheduling
at the Petascale

Jason Coverston[1], Stephen Gombosi[2], Peter Johnsen[1], Per Nyberg[3], Piush Patel[2], Scott Suchyta[2]

[1] Cray Inc., 380 Jackson Street, Suite 210, St. Paul, MN 55101, USA
[2] Altair Engineering Inc., 1820 Big Beaver Road, Troy, MI 48083, USA
[3] Cray Inc., 273 Ch. du Bord-du-Lac, Suite C, Pointe-Claire, QC H9S 4L1, Canada

Abstract

Several operational numerical weather prediction (NWP) centers will approach a petaflop of peak performance by early 2012 presenting several system operation challenges. An evolution in system utilization strategies along with advanced scheduling technologies are needed to exploit these breakthroughs in computational speed while improving the Quality of Service (QoS) and system utilization rates. The Cray XE6™ supercomputer in conjunction with Altair PBS Professional® provides a rich scheduling environment designed to support and maximize the specific features of the Cray architecture. Advantages of this model include avoidance of system thrashing, increased predictability in the scheduling model, and reliable and repeatable runtimes benefiting both operational and research users.

Keywords: Job scheduling, petascale computing, numerical weather prediction, Cray XE6, Altair PBS Professional

1. Introduction

Several operational numerical weather prediction (NWP) centers will approach a petaflop of peak performance by early 2012 presenting several system operation challenges. An evolution in system utilization strategies along with advanced scheduling technologies are needed to exploit these breakthroughs in computational speed while improving the Quality of Service (QoS) and system utilization rates.

An operational NWP workload is composed of a large number of programs which typically have dependencies on one another and must be completed within a defined period of time. Performance requirements are characterized along several dimensions, including application performance for both single large deterministic forecasts and multi-member ensemble prediction systems, as well as QoS for operational and research workloads. In particular, the QoS for operations is focused on guaranteed resources and execution times to meet fixed production schedules.

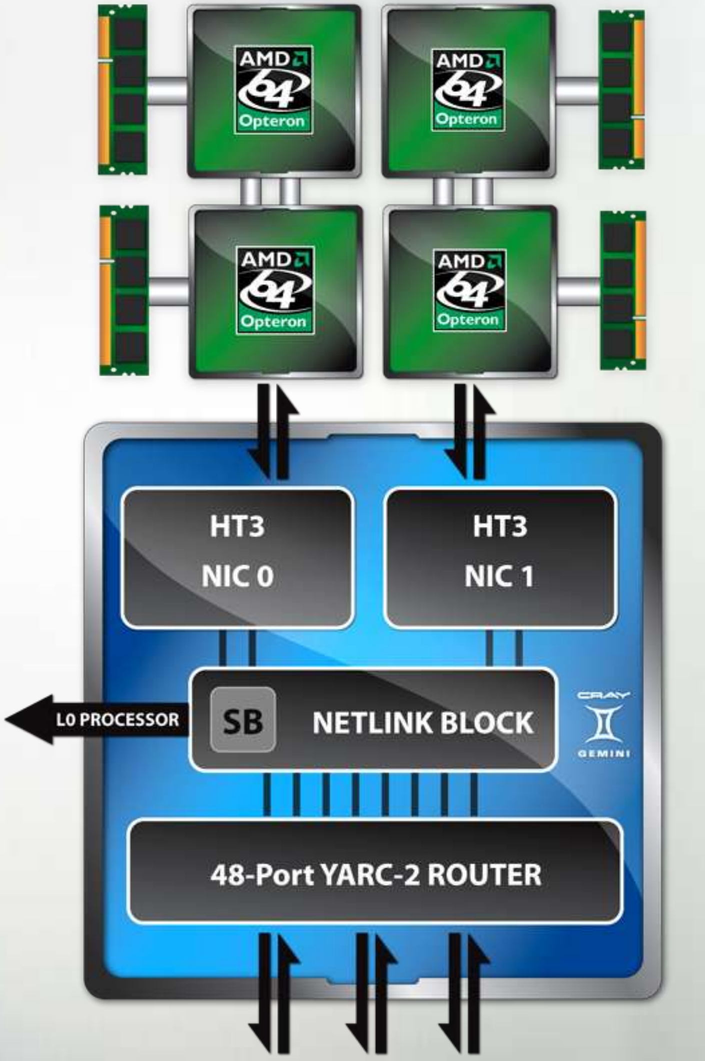
As processor core counts and main memory sizes continue to grow, system administrators face a number of challenges such as achieving reliable runtimes to meet forecast schedules, maximizing system utilization with a mixed operational and research workload, and maintaining QoS in environments with increasing fault rates. Users also face challenges as their jobs and resource requirements grow. Ultimately the system administrator's ability to efficiently schedule resources is highly dependent on accurate information from users. Traditional scheduling strategies examine job priorities at single points in time. The system will have no advance knowledge of the arrival of a high priority job and will allocate the necessary resources through the suspension, checkpoint, swap or killing of lower priority job(s). This is an invasive approach and can result in system thrashing and the

WP-XE01-1010 Page 1 of 11 www.cray.com

Cray-Altair Whitepaper: "Operational Numerical Weather Prediction Job Scheduling at the Petascale"
http://www.cray.com/Products/XE/Resources.aspx

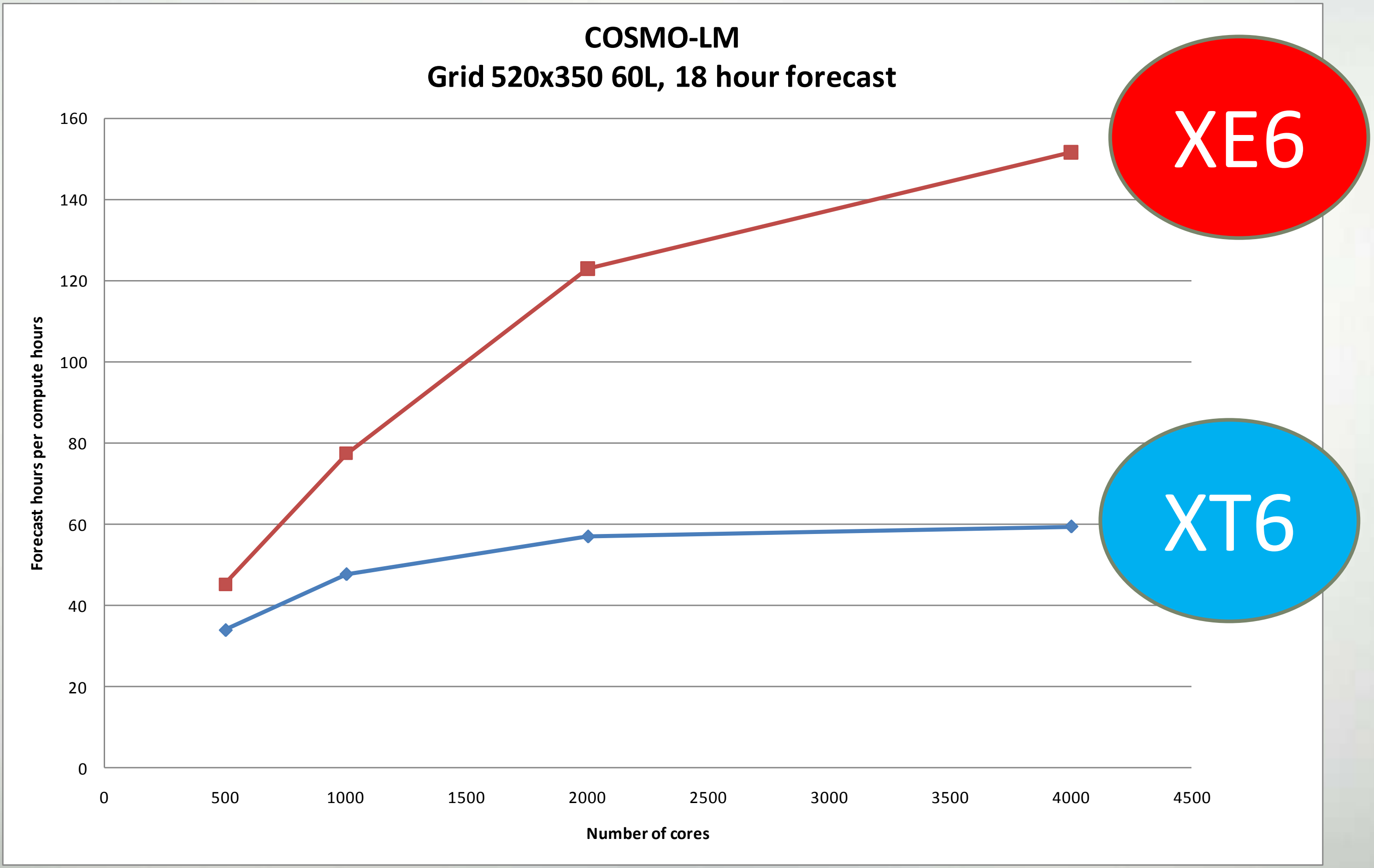
Cray Gemini Network ASIC

- MPI Support
  - Millions of independent messages/sec/NIC
  - BTE for large messages
  - FMA stores for small messages
  - One-sided MPI
- Advanced Synchronization and Communication Features
  - Globally addressable memory
  - Atomic memory operations
  - Pipelined global loads and stores
  - Efficient support for UPC, CAF, and Global Arrays
- Embedded high-performance router
  - Adaptive routing
  - Scales to over 100,000 endpoints
  - Advanced resiliency features

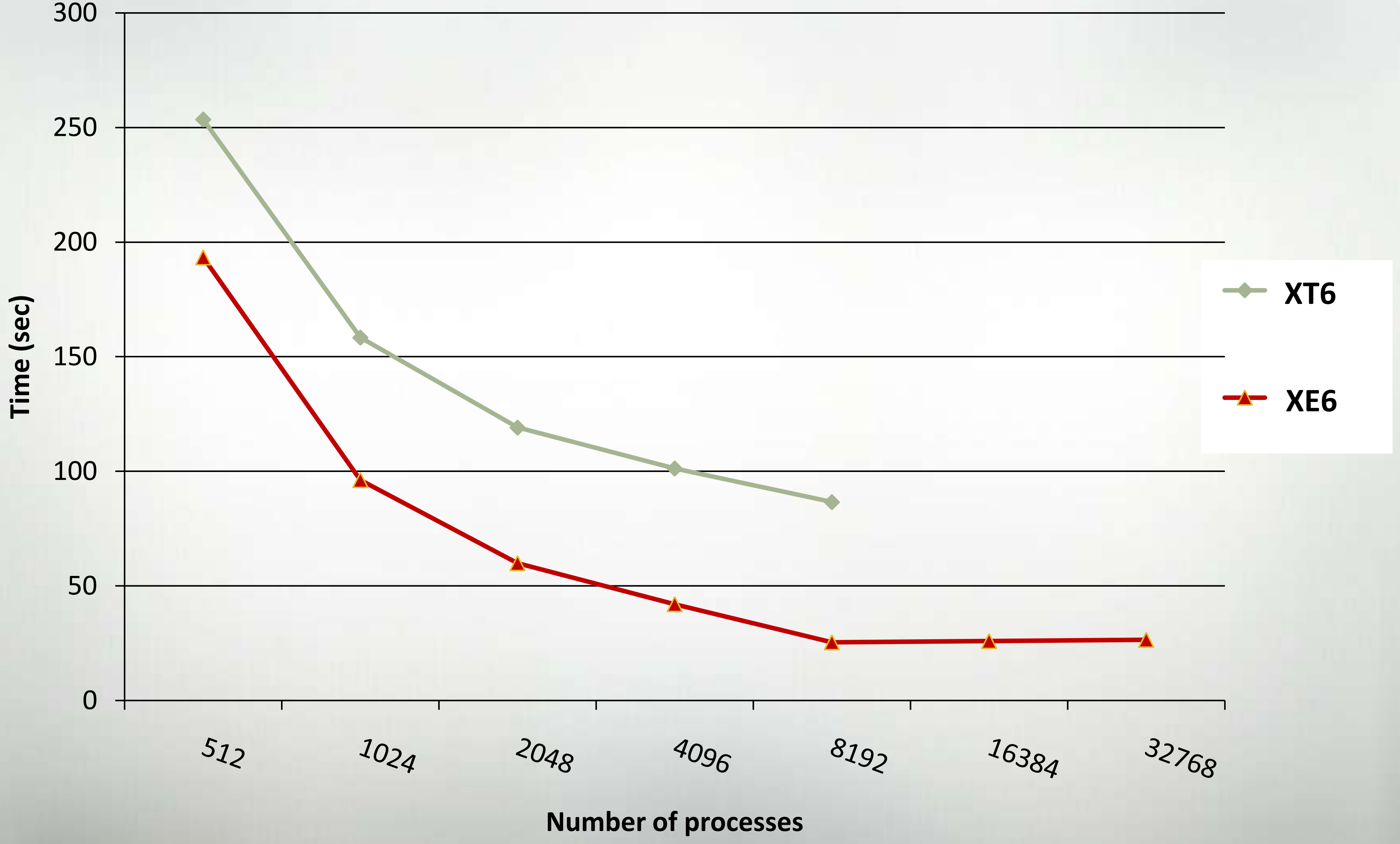


Scalability and Simulation Rate

- Forecast Hours per compute Hours



POP 0.1 grid Total (Timer 27) Comparison
XT6 g34 2.0-2.2 GHz w/ SeaStar vs
XE6 g34 1.9-2.1 GHz w/ Gemini



Summary

- Cray’s MPP technologies are playing a key role in supporting the weather and climate communities:
  - Enabling unprecedented simulations.
  - Supporting the development of next generation modeling capabilities.
- Extreme scale computing will require the successful solution to a collection of interrelated science and technology challenges.
- Cray’s research and development efforts are a multi-pronged approach to address the range of necessary technologies for HPC from current Petascale to emerging Exascale:
  - Performance and Scalability
  - System software and resiliency
  - Programming environments
  - Alternative processing types
  - Facilities and total cost of ownership

Thank you for your attention.