High-performance Computing for Earth Sciences

Ilene Carpenter, Ph.D.
Applications Engineering Manager
ilene@sgi.com

Overview

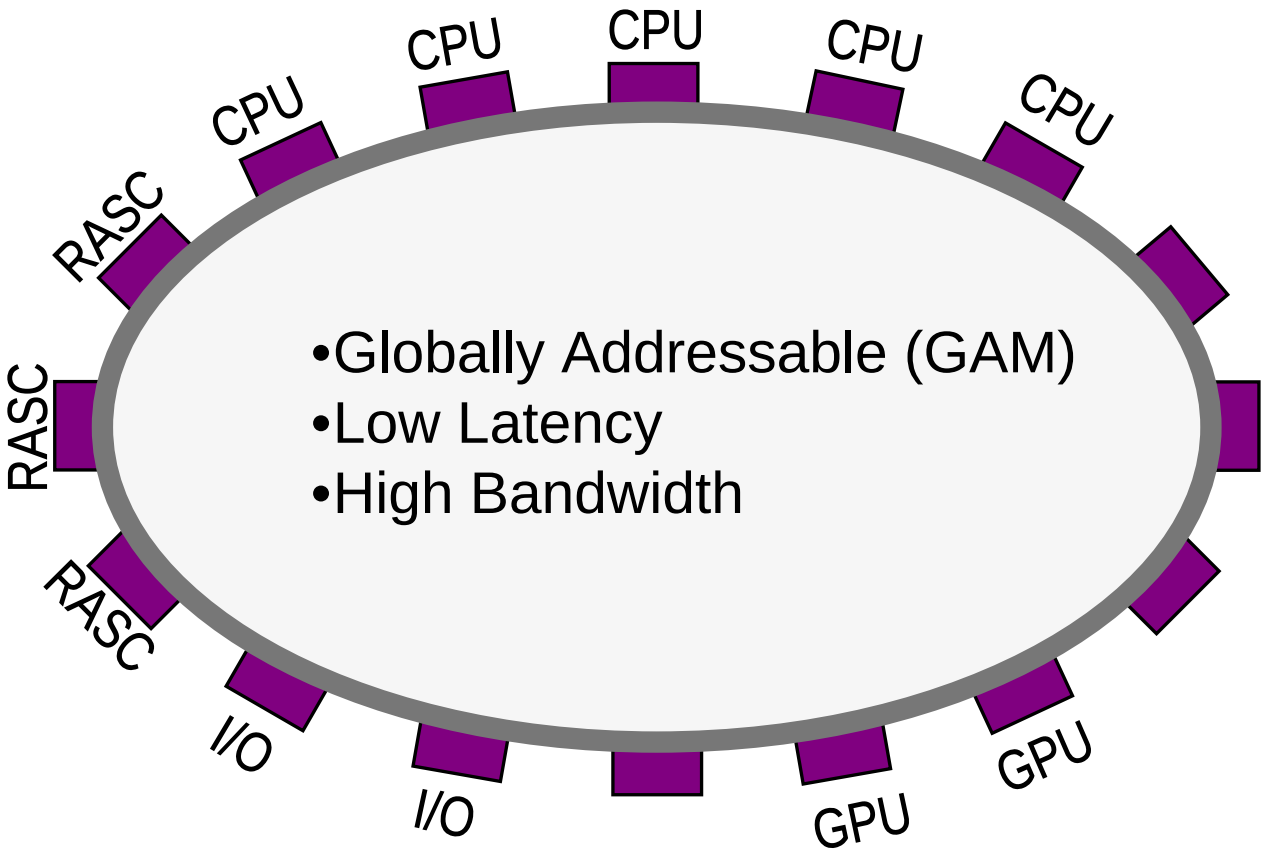
- SGI Philosophy, Current Altix Systems
- Future plans
- Application performance on current systems

SGI philosophy

Deliver high productivity and efficiency through technologies including

- Globally addressable memory
  - Fully cache coherent to a few hundred processors, multiple coherency domains in very large systems
- A small-moderate number of large core-count Linux kernels
- Robust shared filesystem (CXFS) integrated with ILM (DMF)
- Large memory size per Linux kernel, not limited by # of processors, which enables very large memories for relatively few processors. This is very useful for
  - running very large models for applications that don't scale well
  - holding large databases in memory
  - analyzing very large datasets with serial or moderately parallel tools
  - improving I/O performance

Memory-Centric Architecture

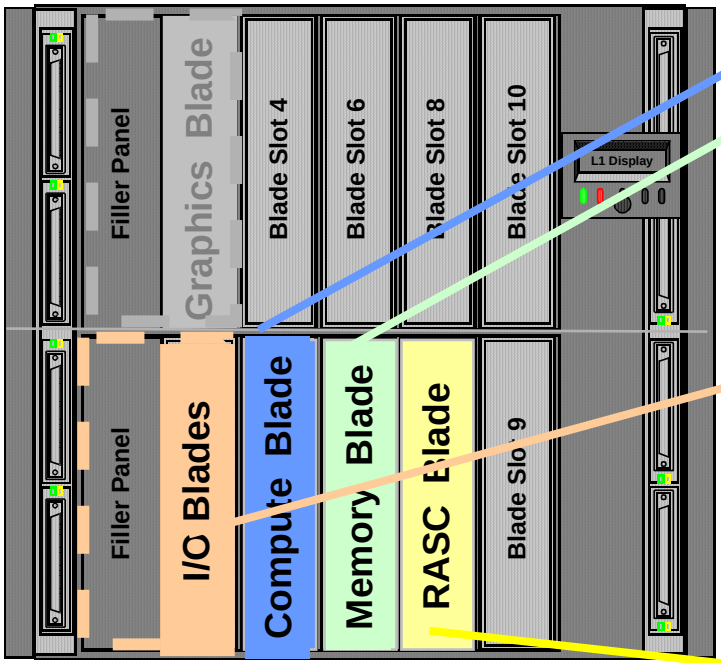


Selected Large System Installations

- NASA Columbia – 10,240p Altix constellation, 20 nodes
  - 2048p single NL fabric with 4 512p partitions + 16x512
  - Madison 9M
- LRZ 8192p Altix 4700
  - 16x512 core nodes, single NL fabric, Madison 9M
- TU Dresden – 2048 core Altix 4700, Montecito
- NOAA GFDL - 2560p Altix 3700 and Altix 3700 BX2 systems (Madison) + ~2560 cores Altix 4700 recently installed, Montecito
- APAC – 1936p Altix 3700/BX2, multiple partitions

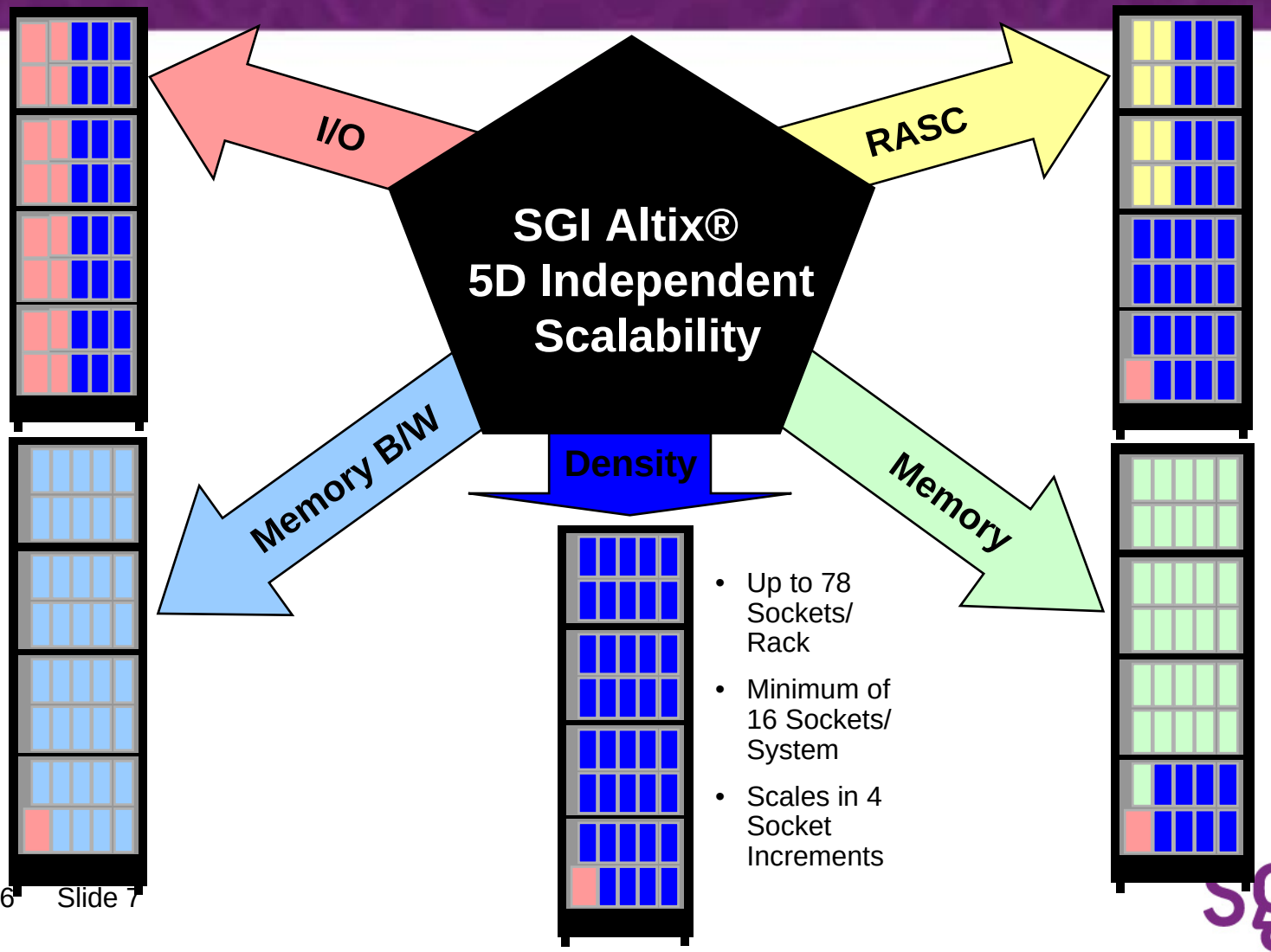
SGI® Altix® Blade Options

More Choices,
Better Solutions



- Compute Blade
- Memory Blade
- Base I/O Blade
- 2 Slot PCI-X Blade
- 3 Slot PCI-X Blade
- 2 Slot PCI-e Blade
- 4 Slot PCI-X/PCI-e Blade
- RC100(RASC™) Blade

Independent Scaling
Optimum Balance for Any Workload



- 16 I/O Blades in a Single Rack
- 16 I/O Blades/ SSI

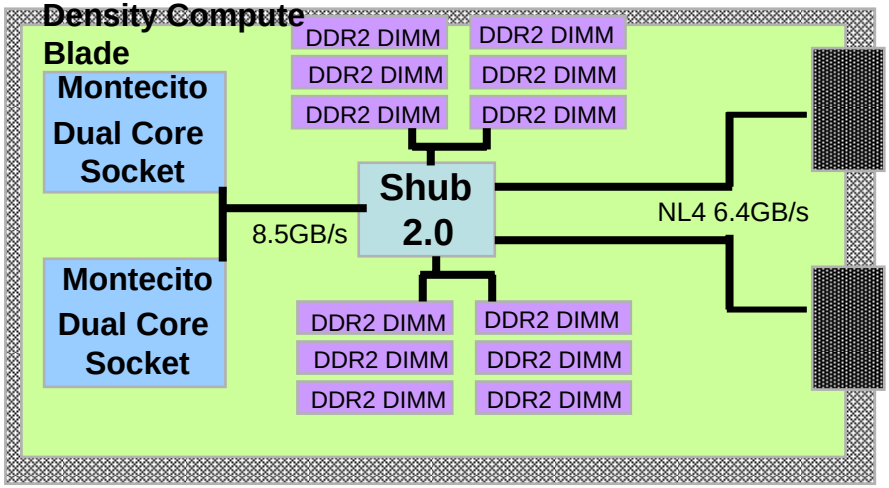
- 8 RASC RC100 blades/ Rack
- 2 FPGAs per blade
- Larger custom configs

- Up to 1.6 byte/flop
- 1 Compute Socket Per Blade
- Up to 39 Sockets/ Rack

- Up to 78 Sockets/ Rack
- Minimum of 16 Sockets/ System
- Scales in 4 Socket Increments

- 1.8TByte/ Rack (4GB DIMMs)
- Minimum of 16 Sockets in Any System

Compute Blade: Excellent Performance Density



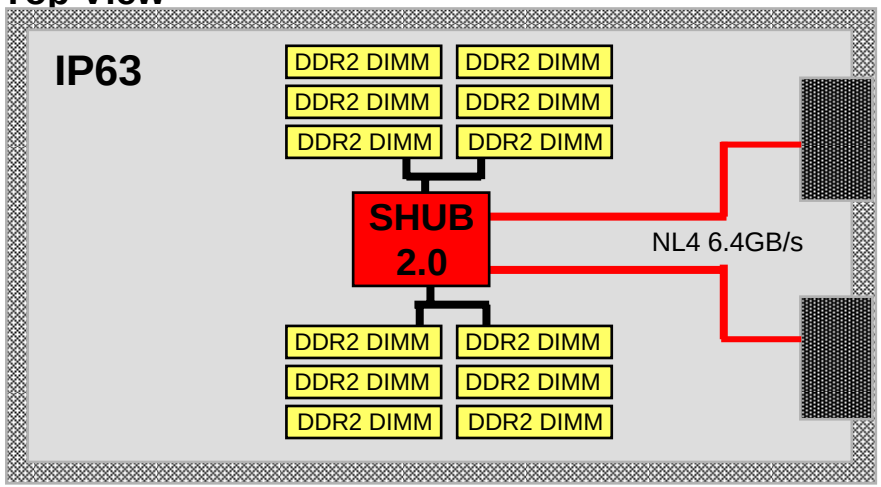
Best $/FLOP, Best Density:

- 2 Processor Sockets Per Blade
- Up to 76 Sockets Per Tall Rack
- Montecito and Montvale compatible
- Memory Sizes: 0.5GB – 6GB/core
  - Greater memory expansion available

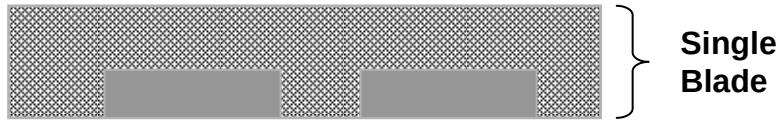
NOTE: Altix 4700 also available in high-bandwidth configuration – 1 socket per blade

Altix 4700 Memory Blade

Top View



Front View



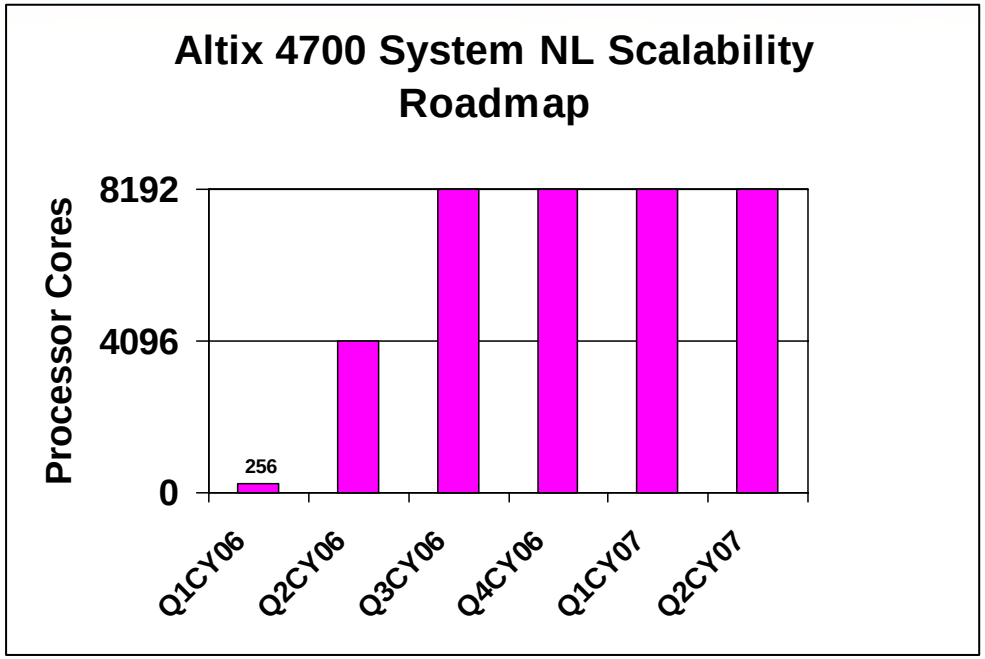
M2 Memory Blade:

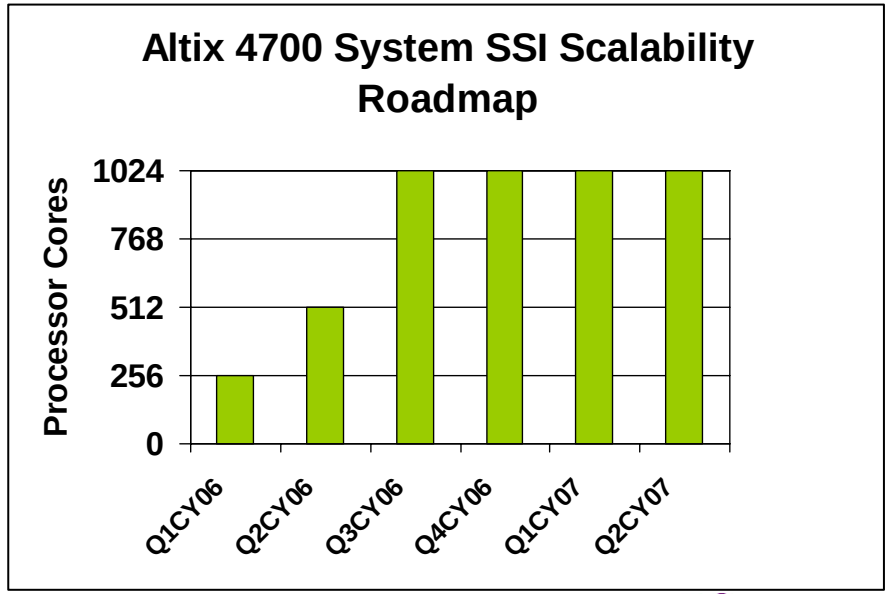
- Scale Memory Independently with 12 DDR2 DIMM Slots Per Blade
- 128 TB Global Addressable Memory with 2GB DIMMs, 200TB with 4GB DIMMs

Altix 4700

- Max size of NUMALink fabric
  - currently is 4096 SHUBs (16384 cores in max density config)
- Max cache coherency domain size
  - 1024 SHUBs (4096 cores in max density config)
- Max SSI size (single Linux kernel)
  - currently is 1024 cores
- Multi-paradigm computing
  - RASC blades
  - Graphics
  - Compute

Altix 4700 System Scalability: Scaling Higher

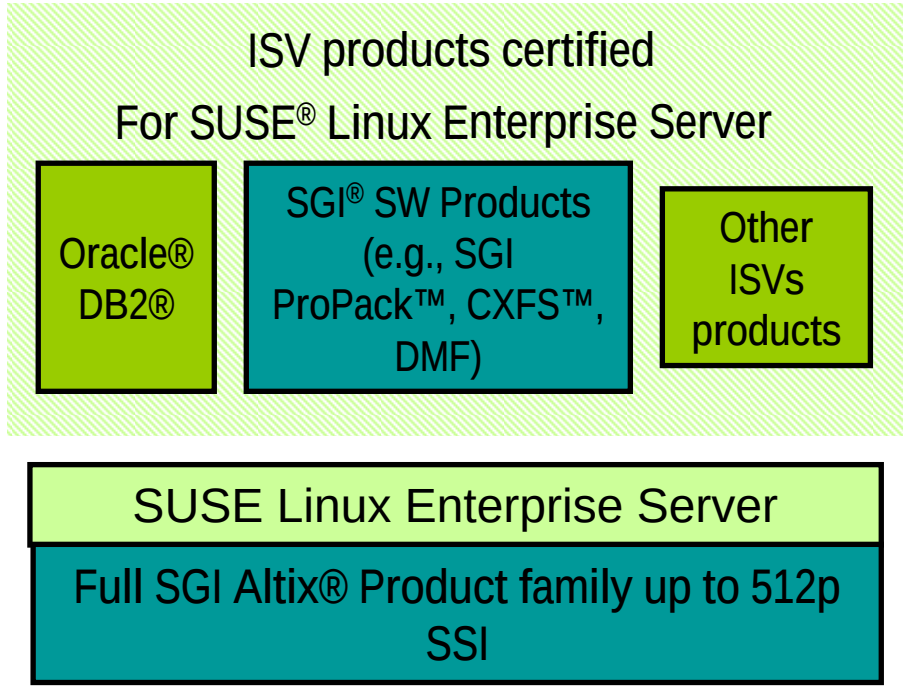




Leveraging Experience in Large Scale Systems to Enhance Reliability & Functionality of Smaller Systems!

SGI Linux Strategy

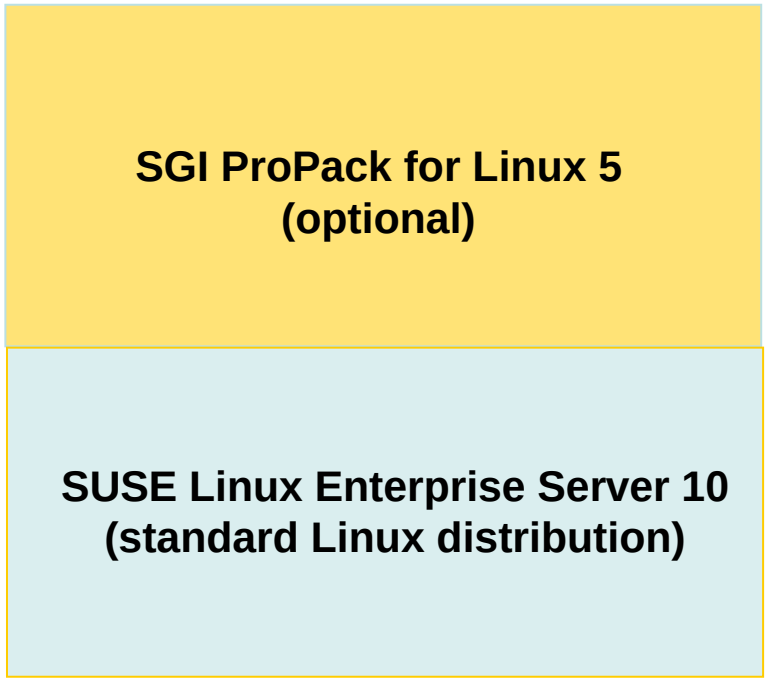
- Contribute HPC expertise to Linux® community
  - Extend Functionality with Community-Accepted Patches to Increase Scalability & Enhance Overall Performance
  - Fix bugs important to HPC customers
- Altix® 4000 Platform Based on Industry Standard Distribution
  - Certified Novell® SUSE LINUX Enterprise Server 9, based on 2.6 kernel
  - SGI ProPack™ 4 for Performance Enhancements & Additional Functionality



SGI ProPack™ HPC Accelerator

- **HPC libraries, products and extensions not available in standard Linux® distribution**
- **Includes open and closed-source software:**
  - SCSL, MPT
  - XVM
  - Performance Co-Pilot™
  - CPUsets and dplace
  - CSA (comprehensive system accounting)
  - FFIO libraries
  - DMF and CXFS™
  - Graphics support
- **Novell® SUSE LINUX Enterprise Server 9**
- **Base and common open-source apps**
  - Kernel platform support
  - Commands, libraries, 100s of RPMs, etc.

SGI® ProPack™ for Linux 5



- Available as an option for SUSE Linux Enterprise Server Version 10
- ProPack on Altix XE includes specific features to drive performance and tuning in x86-64 cluster configurations

New SGI® Altix® XE Product Line

- New SGI line of advanced x86-64 workgroup servers and clusters
- Based on Intel® Dual-core Xeon® 5100 processor architecture
- Fully integrated, fully tested, customizable clusters
- Top performance:
  - 1333Mhz FSB
  - 10.6GB/s memory bandwidth per socket (2 cores)
  - 3 - 4flops/core (2add+2mult)

- **Leading energy efficient performance** – sub-80 watts/socket:
  - 3 GHz thermal design point (TDP) of 80W, others rated at 65W
- Full RoHS compliance
- Modular Systems Management (RAS)
- Industry standard Linux®:
  - SUSE® Linux® Enterprise Server
  - Red Hat Enterprise Linux® *

*** Anticipated availability in Q3CY07**

SGI® ProPack™ Features for Altix® XE

FFIO	Linkless version, set as environment variable to accelerate I/O calls. Drives dramatic performance enhancement in I/O intensive cluster configurations.
Intel Runtime Libraries	Developer and runtime modules from Intel for x86-64 environment.
CPUSETS	Used directly by cluster workload manager, provides ability to allocate specific CPU for system daemons, etc for improved performance, decreased CPU contention
ESP	Tool used by administrators to monitor system health.
XVM	Provides disk striping, mirroring – makes nodes “CXFS” ready.
NUMATOOLS	Used to specify CPU, memory usage characteristics & fine tuning – accessible by developers, users to tune application execution.
Performance Co-Pilot™	System monitoring tool; used to view processor activity, loads, etc.
Storage Administration Tools	Additional tools for managing disk resources – xscsi, udev, LSI commands. Not provided by standard Linux® OS.
Infiniband OpenFabric/Gridstack	Voltaire’s IB management tool.
Failover / Cluster Manager	Basic tool for cluster failover management
CXFS™ Client	Enables use of SGI® CXFS™ - high Performance, shared file system, provides data sharing, enhanced workflow, and reduced costs in data-intensive environments.

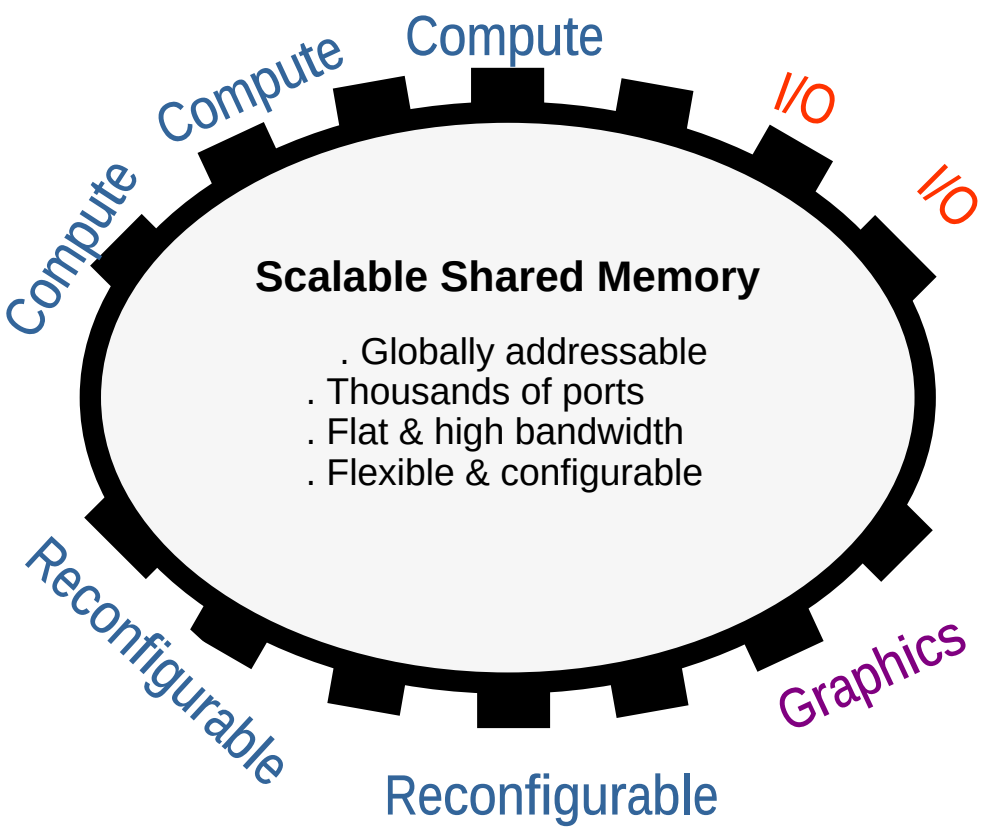
Petascale Strategies

- Continue to design systems with large globally addressable memory
  - Large cache-coherency domains
    - enable a variety of programming paradigms
    - not needed or desirable to extend to full system
  - Allows good performance with CAF and UPC
- Continue to use large SSI (single Linux kernel) on systems with large globally addressable memory to enable
  - Ease of use, higher productivity for users
  - Variety of programming models including very large memory for applications that don't scale to large numbers of cores
  - Simpler system administration
- Need extreme synchonization capabilities (HW and kernel)
- Add cluster/MPP products designed for petascale

Next generation GAM system: Ultraviolet

- Next generation SHUB
- Next generation NUMALink
- Global reference unit
- Features for extremely scalable synchronization
- Enhanced RAS features
- Expanded multi-paradigm options

Ultraviolet Project



- Multi-paradigm Computing
  - Vector
  - Scalar
  - PIM-type
  - Application-specific
- Reconfigurable

Momentum in NWP installations

Recent wins in NWP include -

- The Hungarian Meteorological Service (HMS): selected a 144 processor SGI Altix as its new NWP supercomputer.
- Brazil's National Institute of Meteorology (INMET): chose SGI for computation, visualization and storage.
- Belgium's Royal Meteorological Institute (KMI): deployed a 56 processor SGI Altix system
- The Finnish Meteorological Institute (FMI): installed a 304 processor Altix system
- The Netherlands Meteorological Institute (KNMI): installed a 224 processor Altix system

Customers with Altix systems for Weather Forecasting

- **Finnish Meteorological Institute – 304p**
- **KNMI (Netherlands) - 224p**
- **Hungarian Meteorological Service (144p)**
- **Catalan Meteorological Service (CESCA/MeteoCat) – 128p**
- **Yunnan Meteorology Bureau – 80p**
- **Desert Research Institute (DRI) – 72p**
- **Shanghai Meteorology Center – 64p**
- **INMET Brazil – 64p**
- **KMI (Belgium) – 56p**
- **NOAA NSSL – 32p**
- **Taiwan Central Weather Bureau – 28p**
- **BAMS – 24p**
- **China Met. Administration – 22p**
- **China Met. Administration, Institute of Arid Meteorology – 20p**
- **Meteorological Service of New Zealand – 20p**
- **Puertos del Estado (Spain) – 20p**
- **Roshydromet – 12p**
- **Romanian Met – 2p**

Customers with Altix systems for meteorology and climate research

- NOAA GFDL –2560p Altix 3700 + 2560c Altix 4700– MOM4, AM, CM2.1
- University of Oceanography of China, Tsing Dao – 224p
- Nanjing UIST- 128p + 8p
- U Tasmania/Antarctic CRC – 128p
- CMMACS – 80p Altix 3700 BX2 & 350 – MOM4
- Universidad Complutense – 64p
- Beijing Normal University, Climate Modeling Branch, State Lab of Remote Sensing Science – 56p
- First Institute of Oceanography (China) – 56p
- Georgia Tech – 48p
- Institute of Desert Meteorology, China – 32p
- Univ. of Florida- 32p
- Harvard University – 28p
- Univ. of Wisconsin CMISS – 24p
- MIT Dept of Earth, Atmosphere and Planetary Science – 20p
- Dalhousie University – 16p
- NIO, Goa, India – 16p
- Univ. of Utah – 16p
- Woods Hole Oceanographic Institute – 16p
- Univ. of Colorado Boulder - 12p
- APAT -8p
- Utrecht Univ – 8p
- Univ. of South Florida – 4p
- Florida Institute of Technology – 2p

Customers with Altix systems running climate, weather and environmental sciences applications

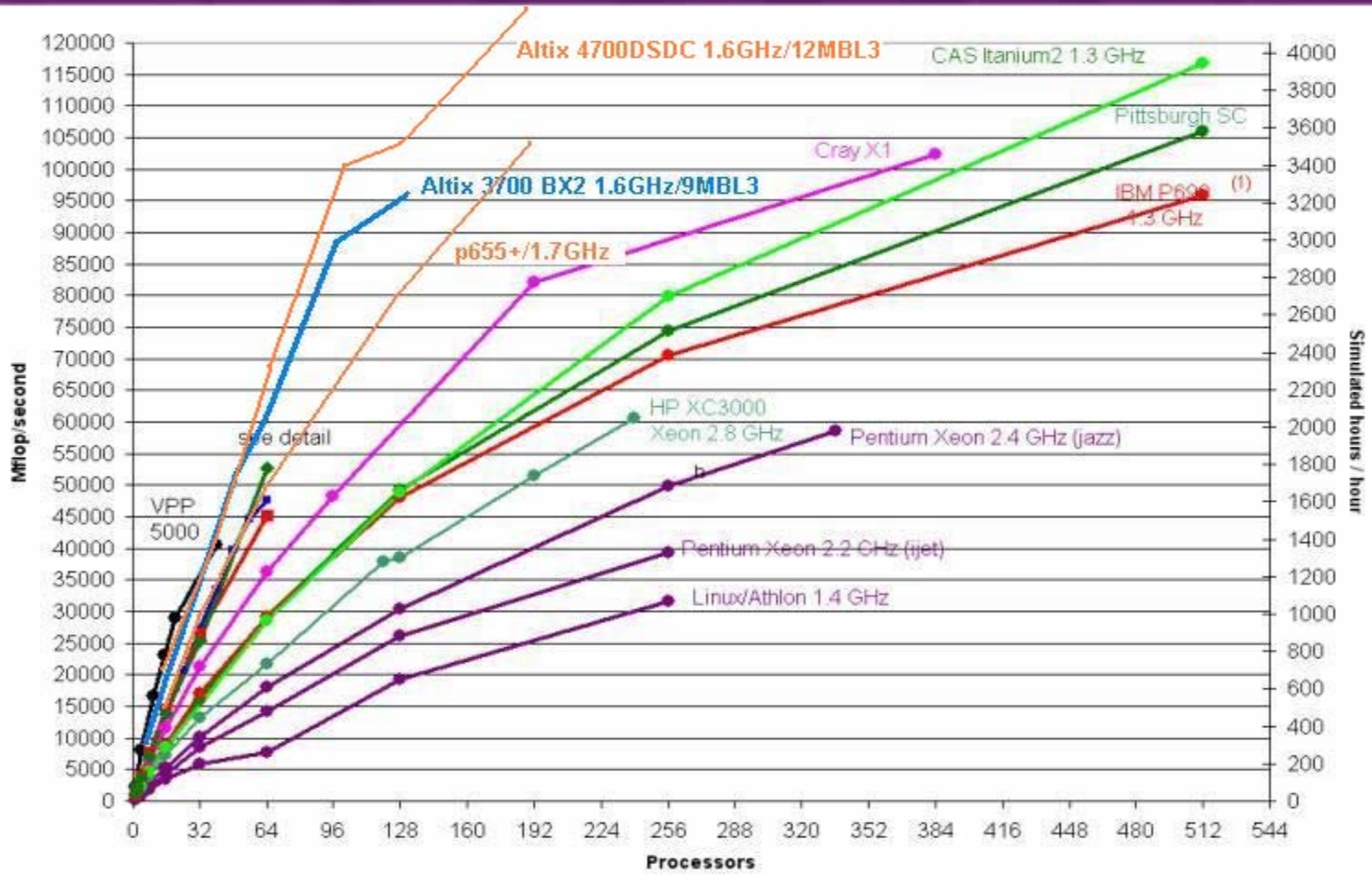
- NASA Ames– 10240p – ECCO, CAM, CCSM, fvGCM
- NASA Goddard - 2048p
- APAC – 1936p – CCSM, MOM
- NCSA – 1024p – WRF
- Univ. of Manchester/CSAR – 512p
- NRL – 384p (1x128, 1x256) - various
- ORNL – 256p – CCSM, CAM, POP
- LANL – 256p – POP, HYCOM
- U of Queensland– 208p
- JAMSTEC – 96p
- CSIRO – 64p

Climate and Weather Model Performance

- SGI MPT library takes advantage of shared memory to provide very low latency, high bandwidth MPI communication.
- I/O performance is balanced with compute performance when scaled to the largest Altix systems we have run on

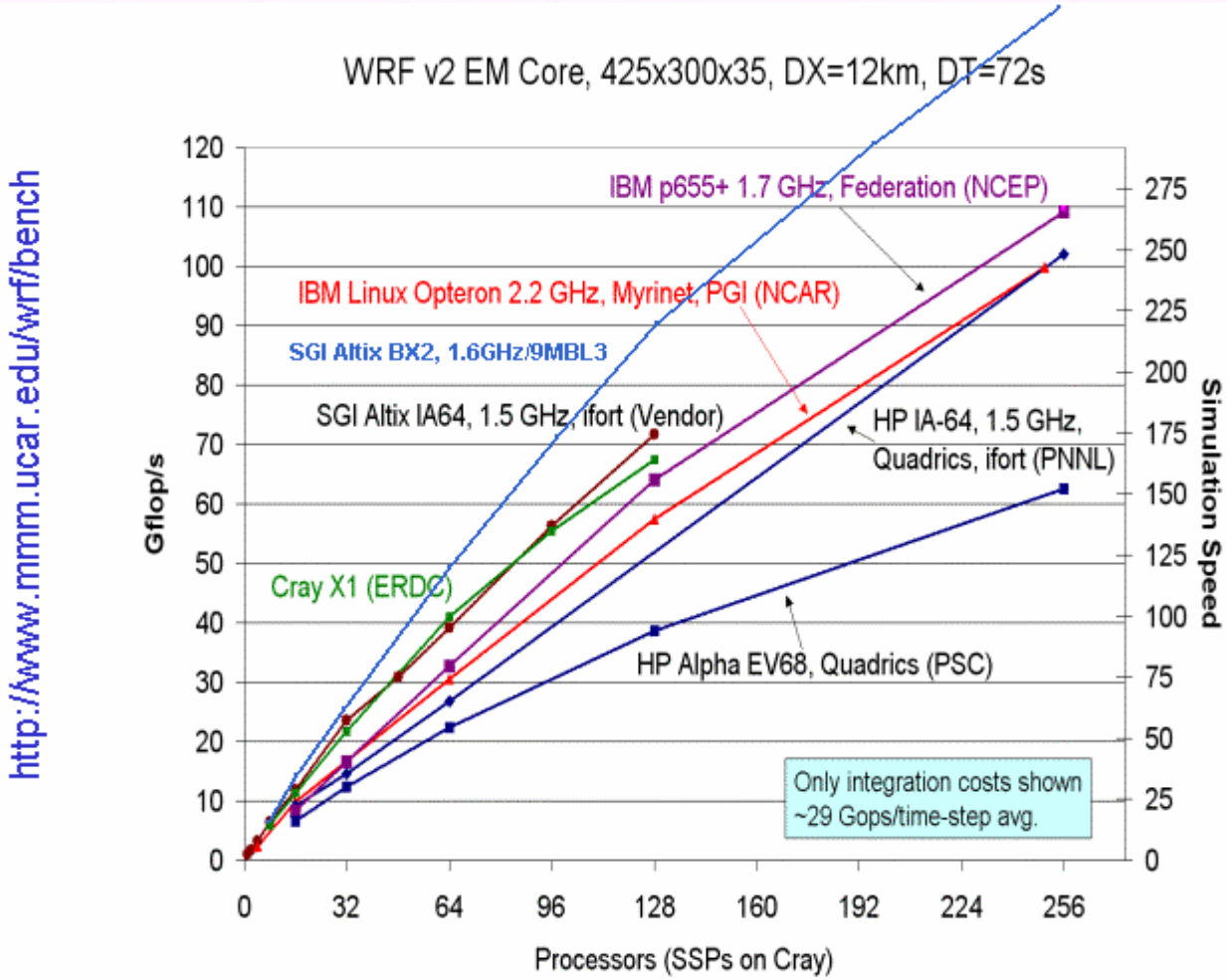
MM5 3.6.3 - Standard benchmark

Higher is Better



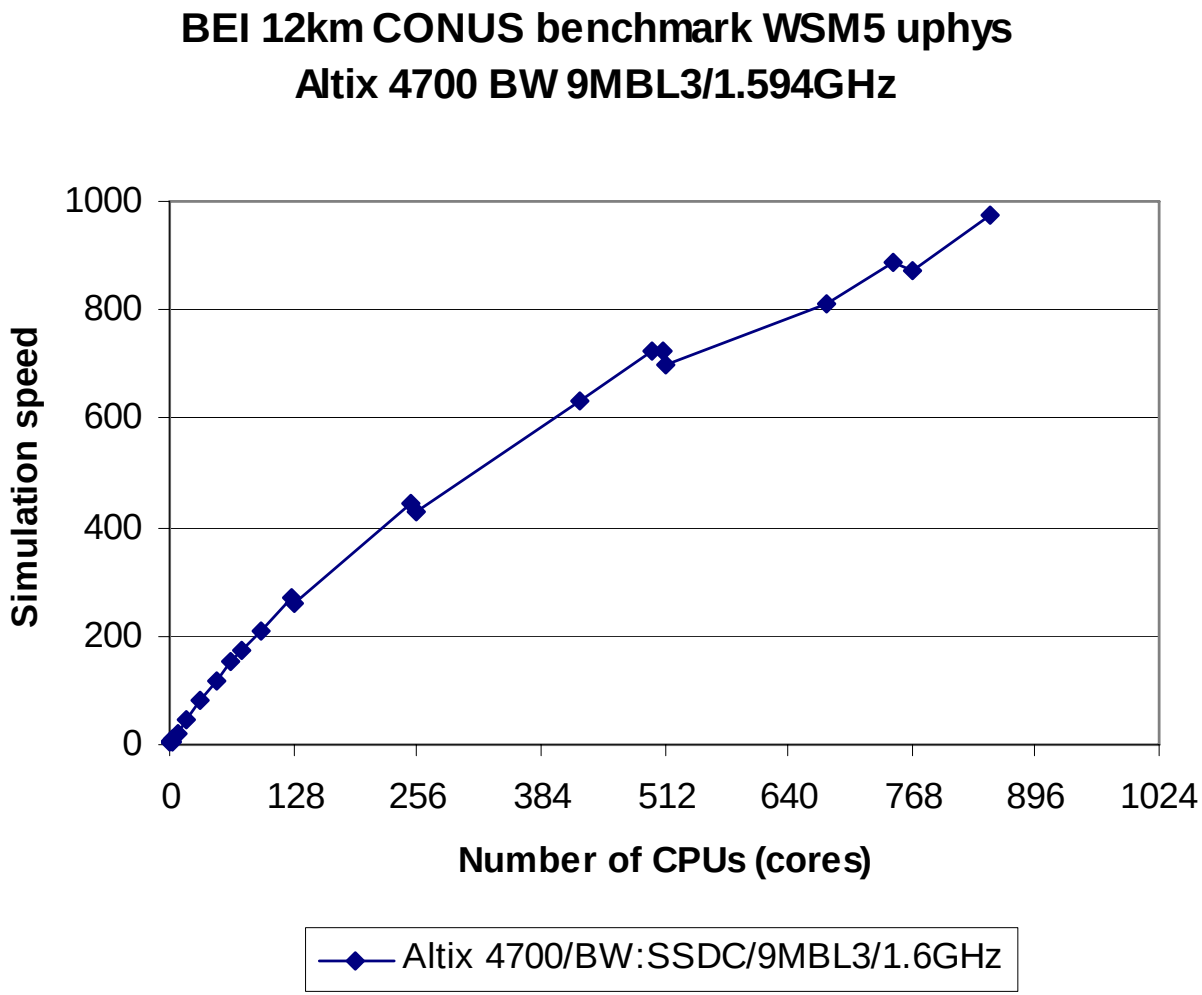
Source: http://www.mmm.ucar.edu/mm5/mpp/helpdesk/20040304a.html

WRF 2.0.2 - Scalability & Performance on Altix 3700

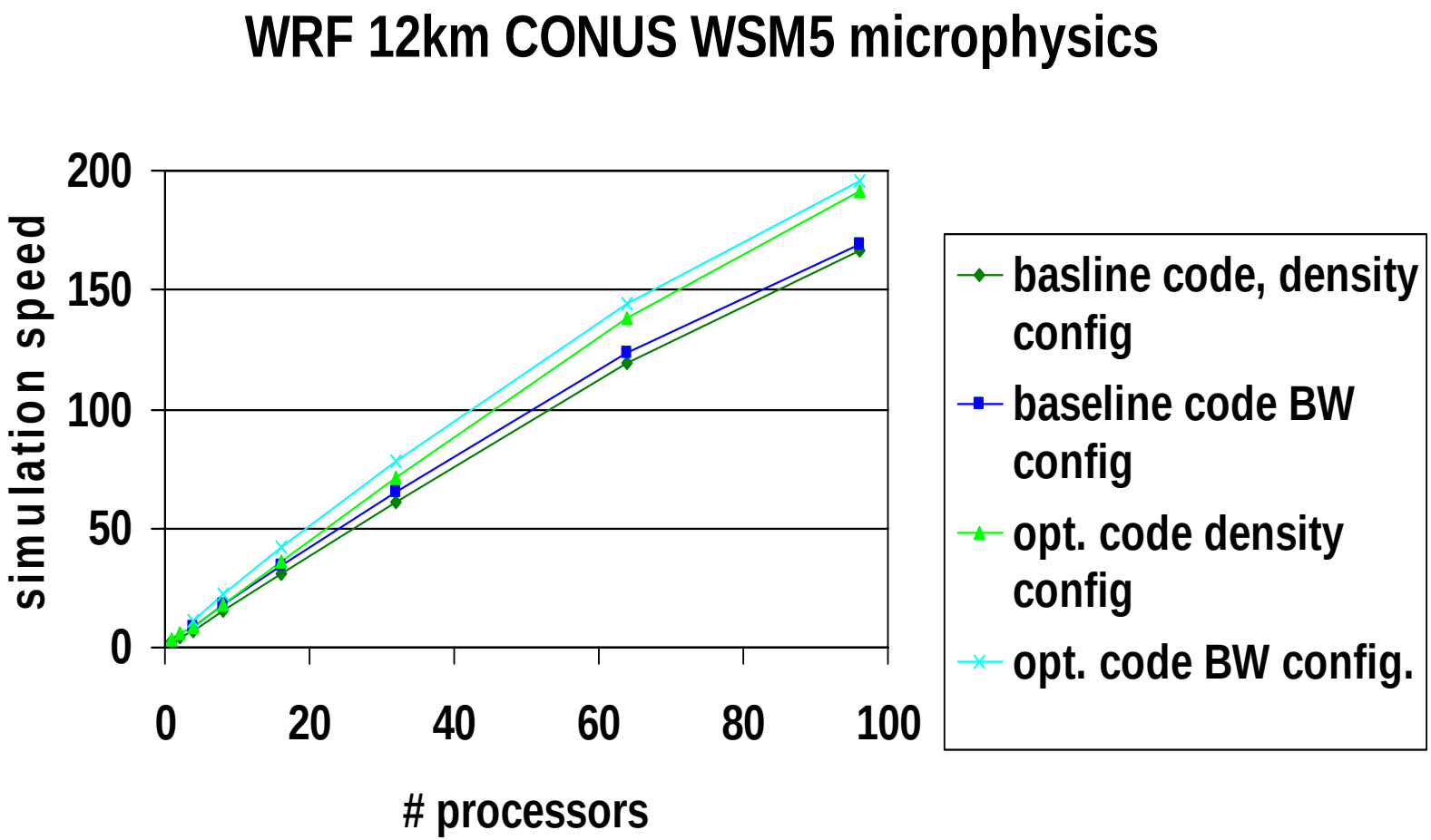


Source: http://box.mmm.ucar.edu/wrf/bench

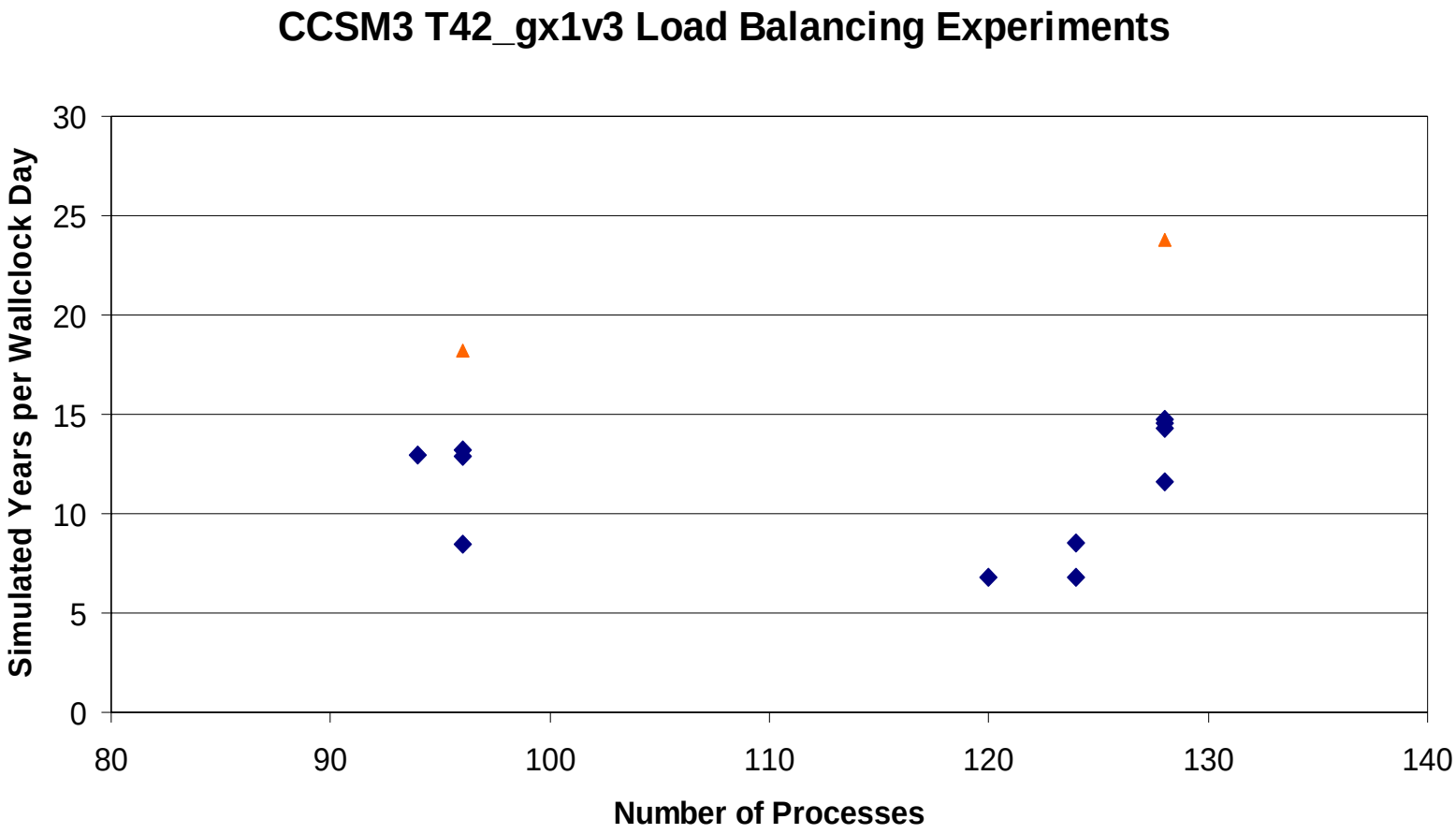
WRF 2.1.2: Altix 4700, 1.6GHz/9MBL3, 1024-core SSI



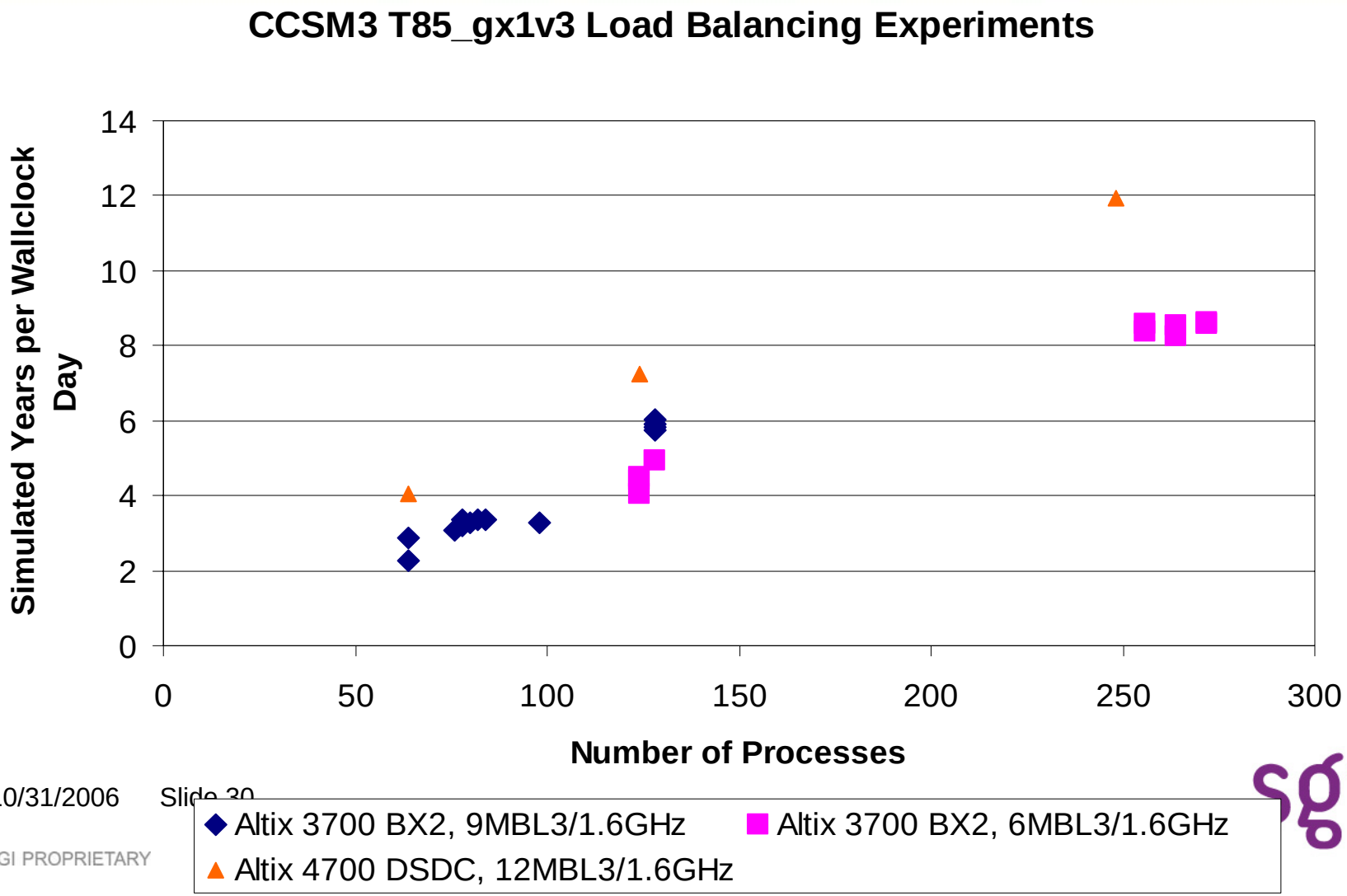
WRF 2.1.2: Altix 4700, 1.6GHz/12MBL3



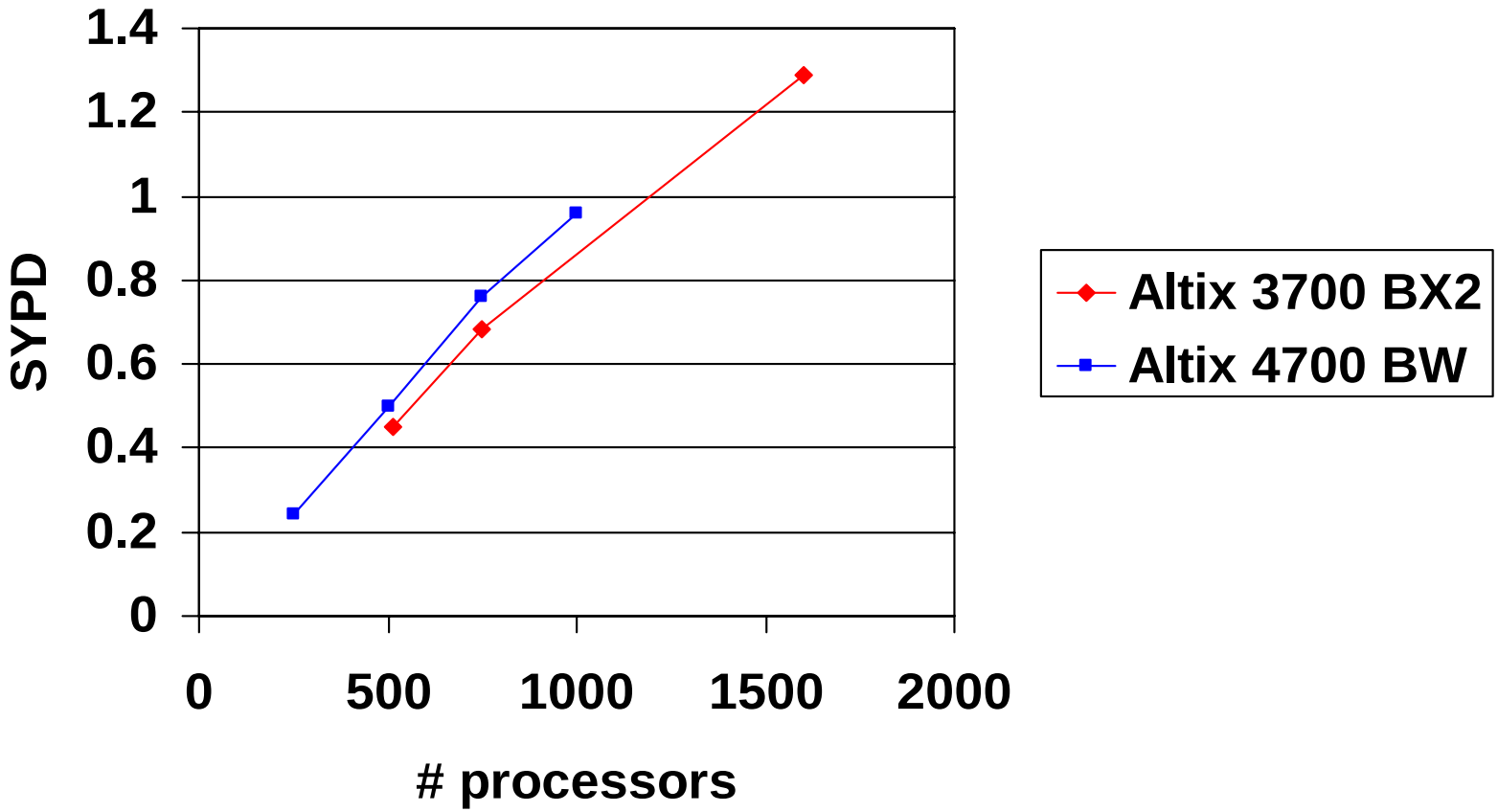
CCSM 3.0 on the SGI Altix 3700 BX2 & 4700



CCSM 3.0 on the SGI Altix 3700 BX2 & 4700



POP 0.1 degree global



Recently ported weather and climate applications

- ARPS
- BRAMS
- RSM
- IFS Forecast model and 4DVar (RAPS9 release)
- ROMS 2.2

Conclusions

SGI has emerged from Ch 11 with an expanded product line and expanded market focus:

- Addition of x86-64 based products
  - Altix XE clusters today
  - highly scalable systems in the future
- Continued development of large GAM systems
- Expand target markets to include enterprise, especially large data management


sgi®