

Shared Memory Systems for Petascale Computing

Ilene Carpenter, Ph.D.

Applications Engineering Manager

ilene@sgi.com

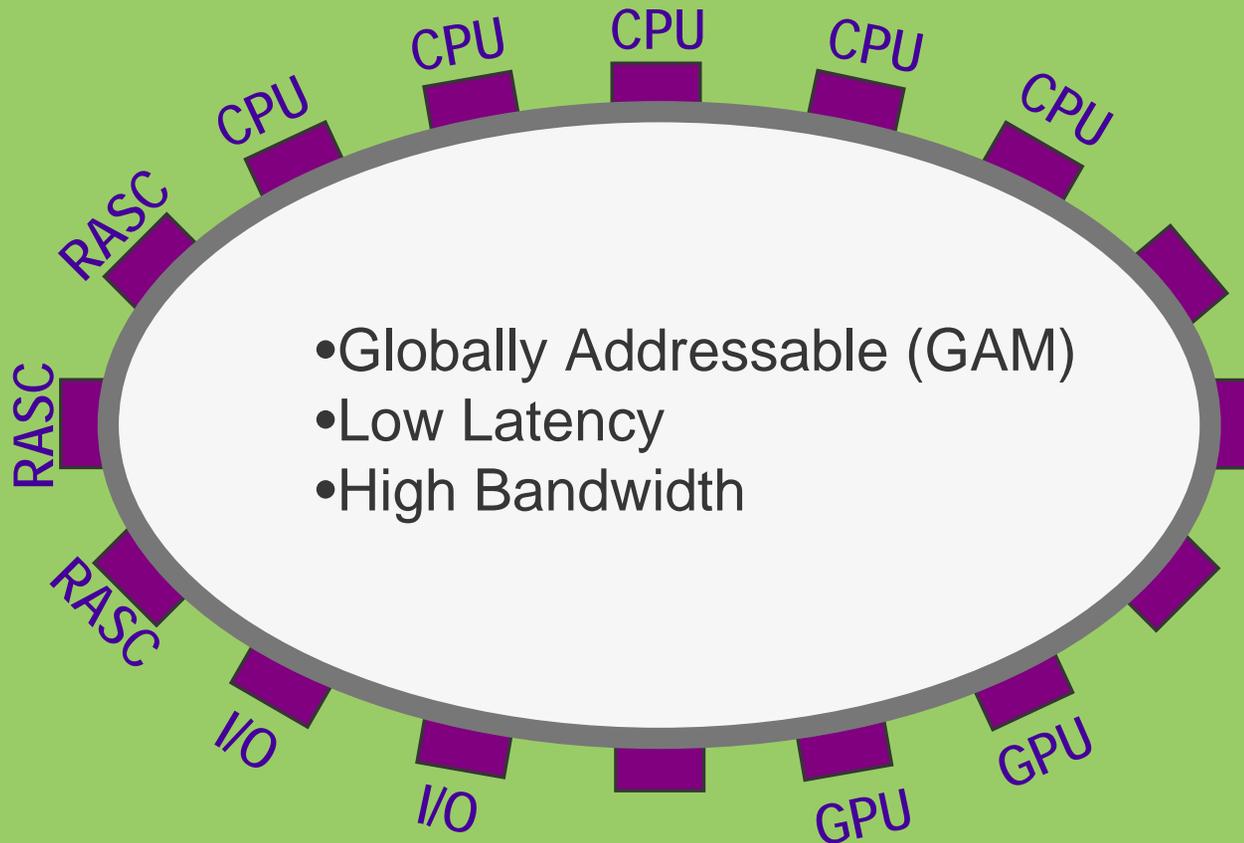
Overview

- **SGI Philosophy, Current Altix Systems**
- **Future plans**

Philosophy: deliver high productivity

- Globally addressable memory
 - Fully cache coherent to a few hundred processors
 - Multiple coherency domains in very large systems
- A small number of large core-count Linux kernels
 - Already dealt with issues that result from large number of threads in single address space that multi-core chips are bringing to clusters and MPP systems.
- Robust, coherent shared filesystems integrated with ILM (DMF) – I/O is a primary consideration in system design
- Large memory size per Linux kernel including very large memories for relatively few processors
 - runs very large models for applications that don't scale well
 - holds large databases in memory
 - analyze very large datasets with serial or moderately parallel tools

Memory-Centric Architecture

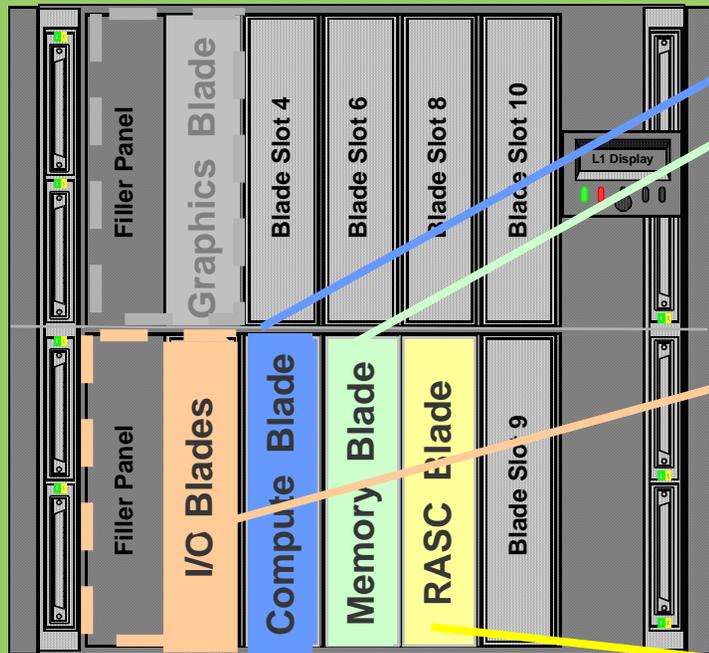


Selected Large System Installations

- NASA Columbia – 10,240p Altix 3700 and Altix 3700 BX2 systems
 - 2048p single NL fabric with 4 512p partitions + 16x512 IB cluster
 - Madison 9M
- LRZ 8192p Altix 4700
 - 16x512 core nodes, single NL fabric, Madison 9M
- TU Dresden – 1024 core Altix 4700, Montecito
- NOAA GFDL - 2560p Altix 3700 and Altix 3700 BX2 systems (Madison) + ~2500 cores Altix 4700 recently installed, Montecito dual-core
- APAC – 1936p Altix 3700/BX2, multiple partitions

SGI® Altix® Blade Options

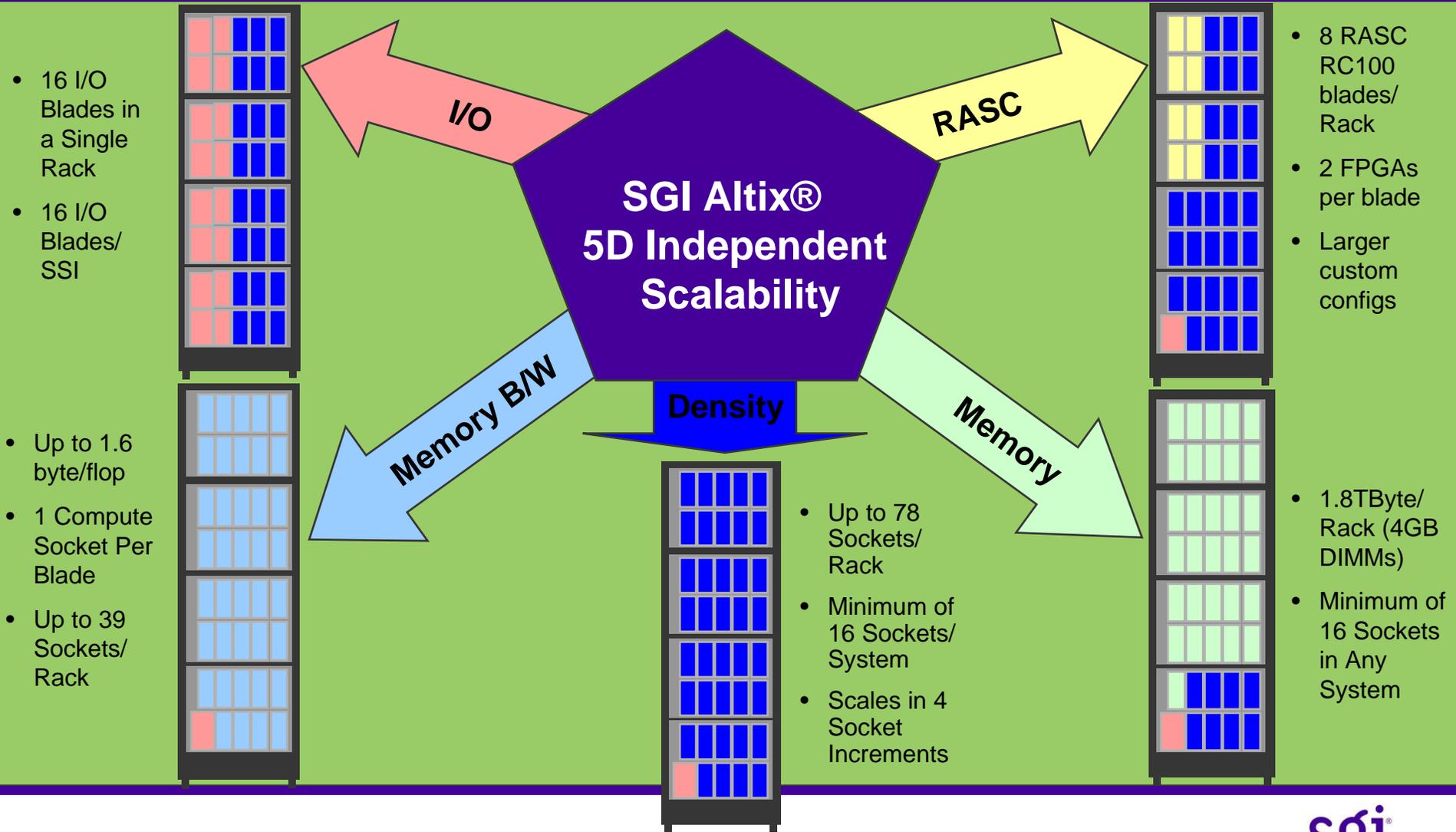
More Choices,
Better Solutions



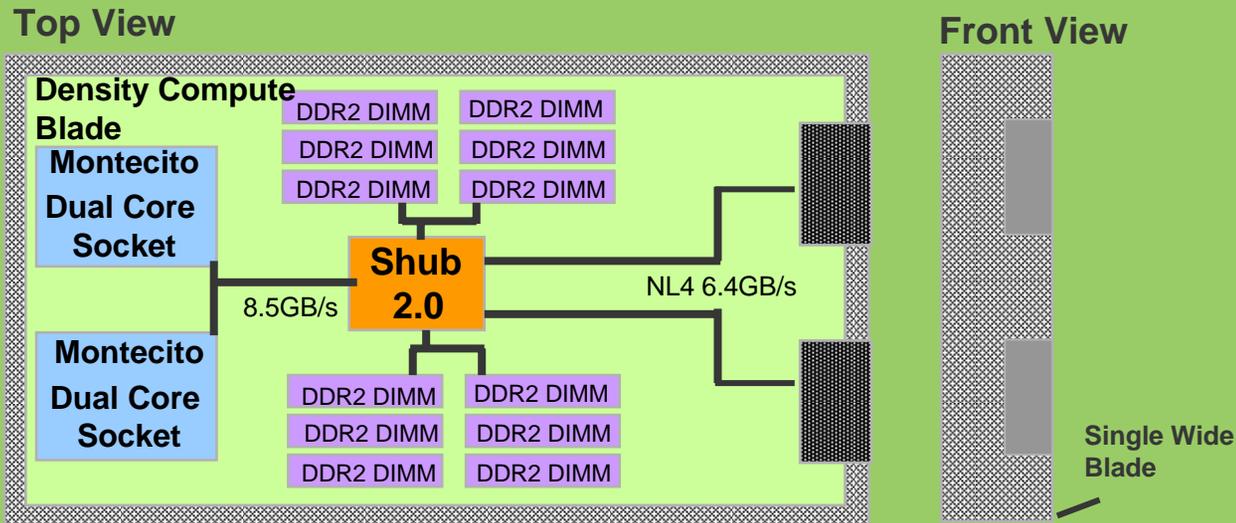
- Compute Blade
- Memory Blade
- Base I/O Blade
- 2 Slot PCI-X Blade
- 3 Slot PCI-X Blade
- 2 Slot PCI-e Blade
- 4 Slot PCI-X/PCI-e Blade
- RC100(RASC™) Blade

Independent Scaling

Optimum Balance for Any Workload



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

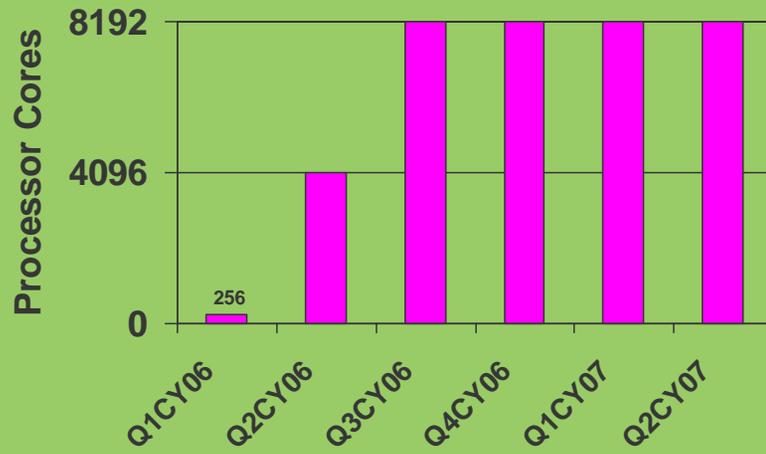
NOTE: Altix 4700 also available in high-bandwidth configuration – 1 socket per blade

Altix 4700

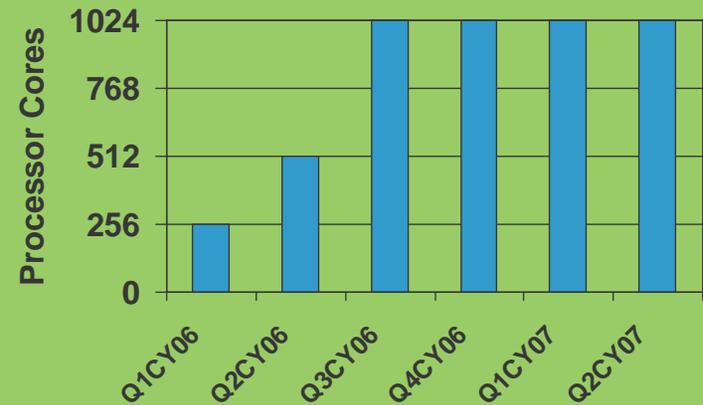
- Max size of NUMALink fabric
 - currently is 4096 SHUBs (16384 cores in max density config)
- Max cache coherency domain size
 - 1024 SHUBs (4096 cores in max density config)
- Max SSI size (single Linux kernel)
 - currently is 1024 cores
- Multi-paradigm computing
 - RASC blades
 - Graphics
 - Compute

Altix 4700 System Scalability: Scaling Higher

Altix 4700 System NL Scalability Roadmap



Altix 4700 System SSI Scalability Roadmap



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

SGI ProPack™ HPC Accelerator

**SGI ProPack™
for Linux®**

**Standard Linux
Distribution**

- **HPC libraries, products and extensions not available in standard Linux® distribution**
- **Includes open and closed-source software:**
 - 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**

SGI® ProPack™ for Linux 5

**SGI ProPack for Linux 5
(optional)**

**SUSE Linux Enterprise Server 10
(standard Linux distribution)**

- 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

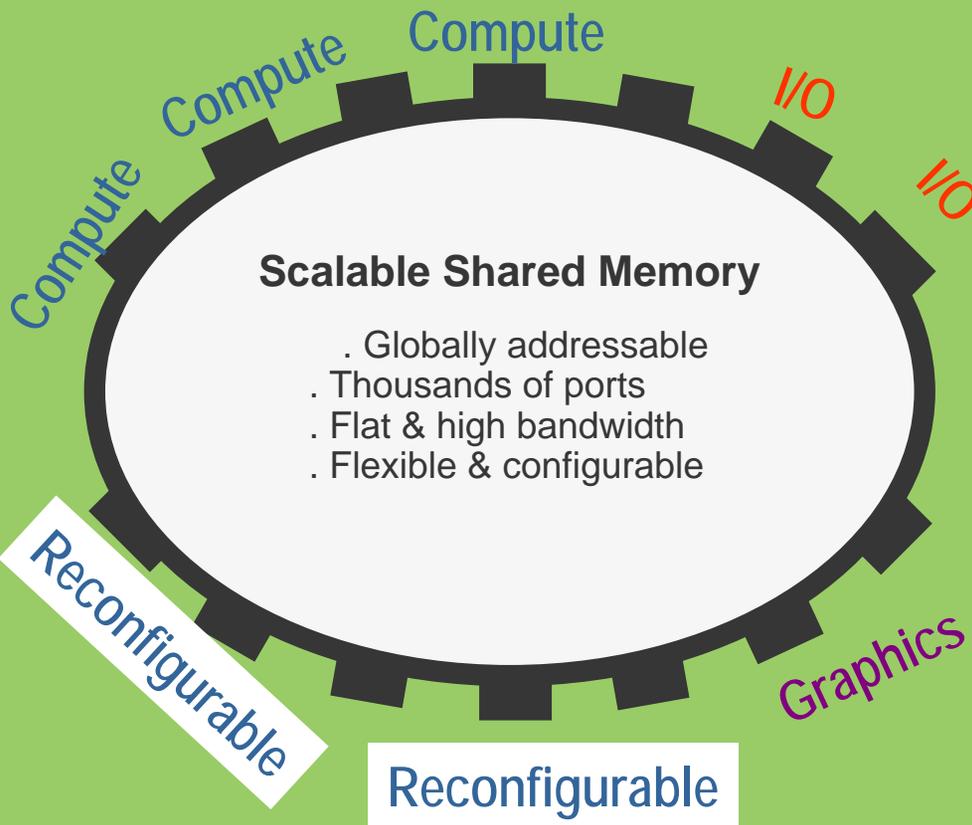
Petascale Strategies

- 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) for each node 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
 - High performance I/O
 - Simpler system administration
- Need extreme synchronization capabilities (HW and kernel)
- Expand elements of multi-paradigm computing
- Improve RAS features to enable high productivity

Next generation NUMALink system: Ultraviolet

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

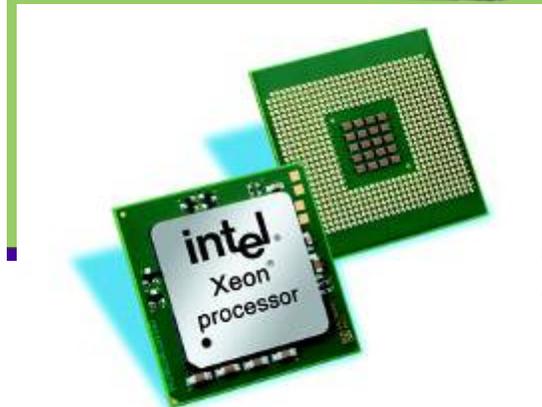
Ultraviolet Project



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

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

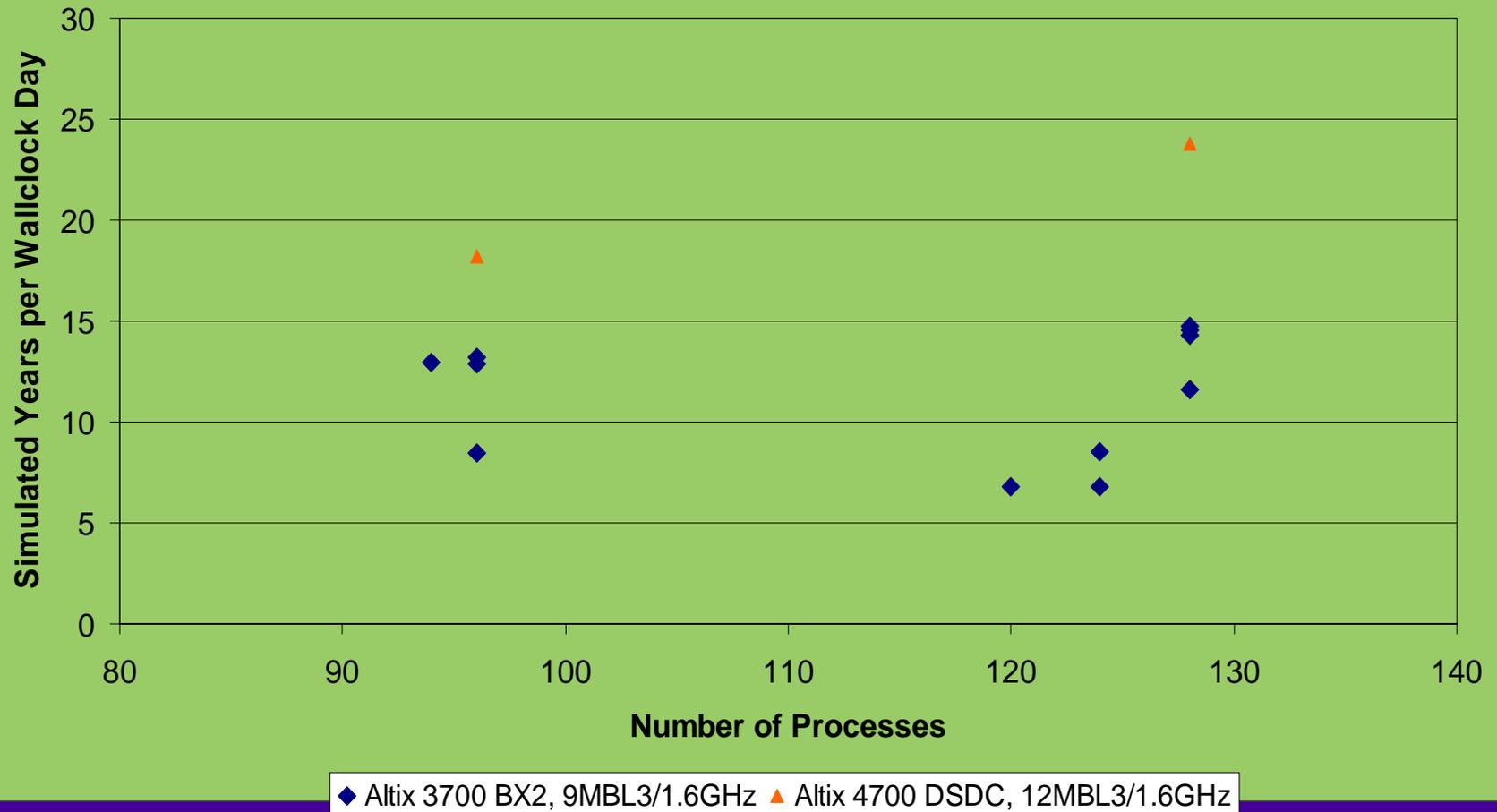
FFIO	Linkless version, set as environment variable to accelerate I/O calls. Drives dramatic performance enhancement in I/O intensive cluster configurations.
Intel Runtime Libraries	Developer and runtime modules from Intel for x86-64 environment.
CPUSETS	Used directly by cluster workload manager, provides 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.

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 systems we have run on

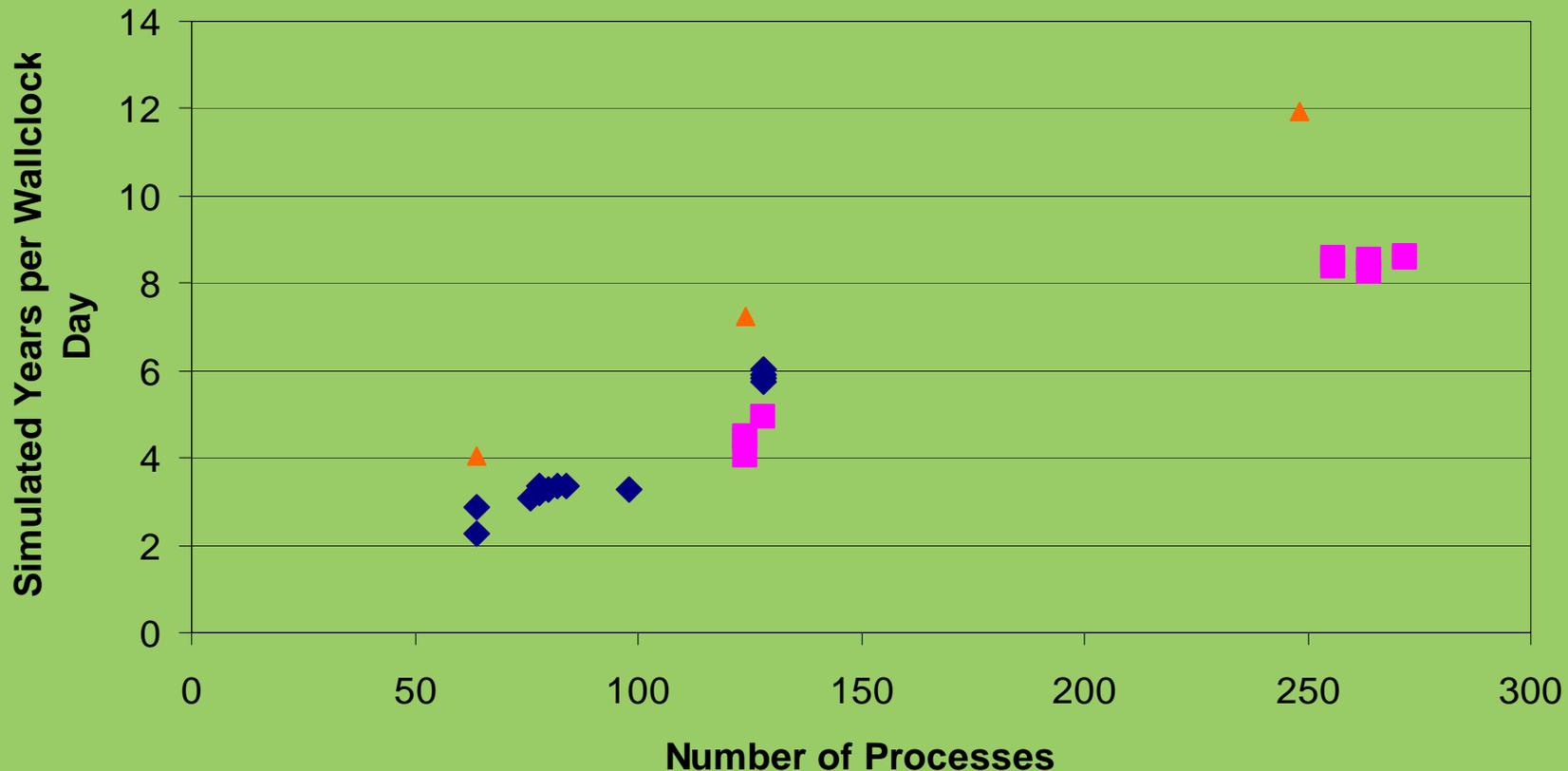
CCSM 3.0 on the SGI Altix 3700 BX2 & 4700

CCSM3 T42_gx1v3 Load Balancing Experiments



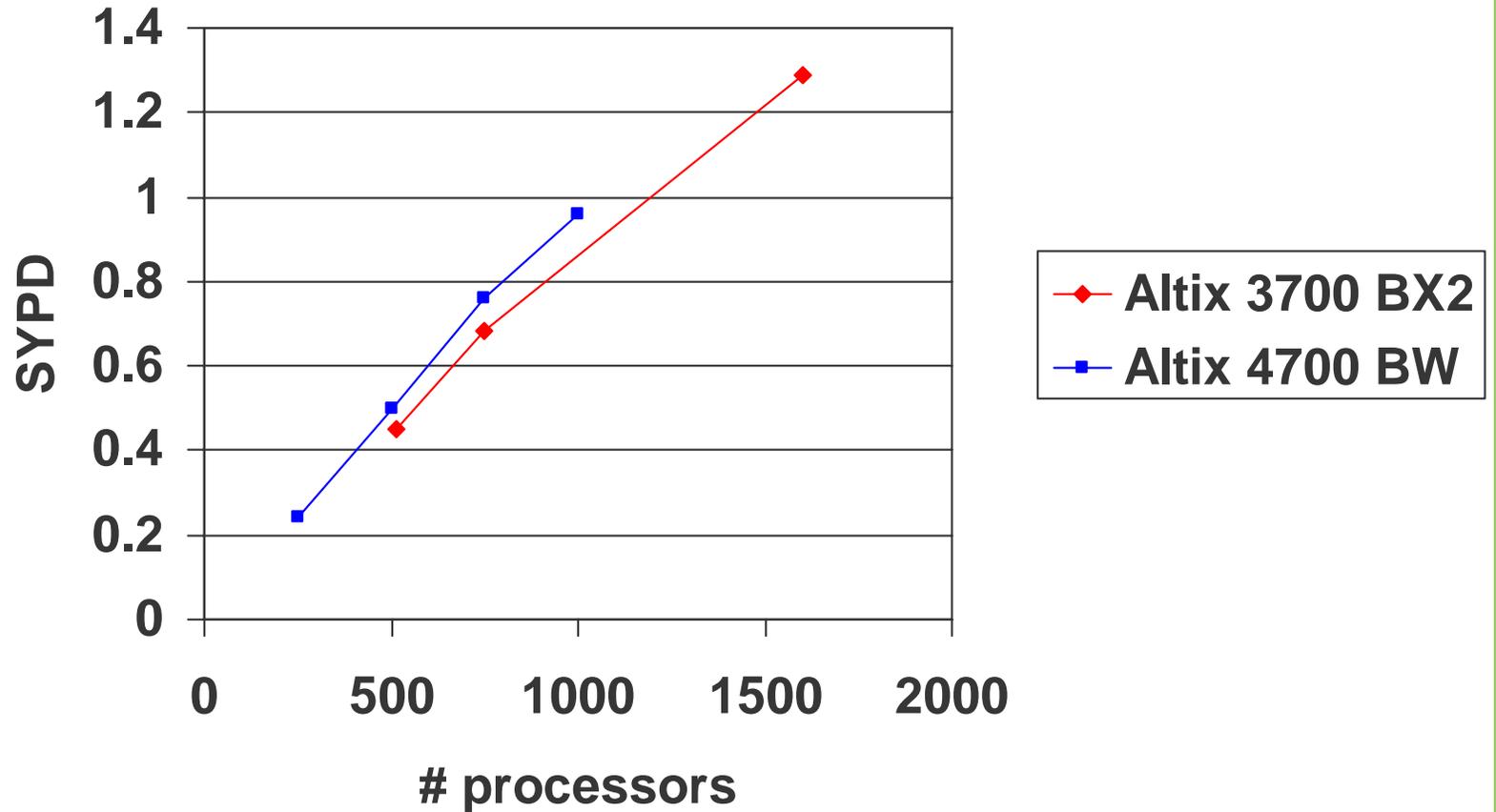
CCSM 3.0 on the SGI Altix 3700 BX2 & 4700

CCSM3 T85_gx1v3 Load Balancing Experiments



- ◆ Altix 3700 BX2, 9MBL3/1.6GHz
- ▲ Altix 4700 DSDC, 12MBL3/1.6GHz
- Altix 3700 BX2, 6MBL3/1.6GHz

POP 0.1 degree global



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 very dense systems for petascale computing
- Continued development of large GAM systems
- Expand target markets to include enterprise, especially large data management

sggi[®]