



# Update On SGI Technology

Michael Woodacre  
Chief Engineer  
[woodacre@sgi.com](mailto:woodacre@sgi.com)

# Lots has happened since 2008 workshop...

- Rackable takeover
- Launch of Altix ICE 8400
- Launch of Altix UV
- Launch of SGI Management Center
- Advances in data center technology
- Advances in storage technology
- Looking towards challenges of extreme scale computing

# SGI: accelerating results™

## VISION

Leader in technical computing:

- High Performance, Large Scale Infrastructure
- Open & Scalable Storage
- Integrated Software
- Expert Services

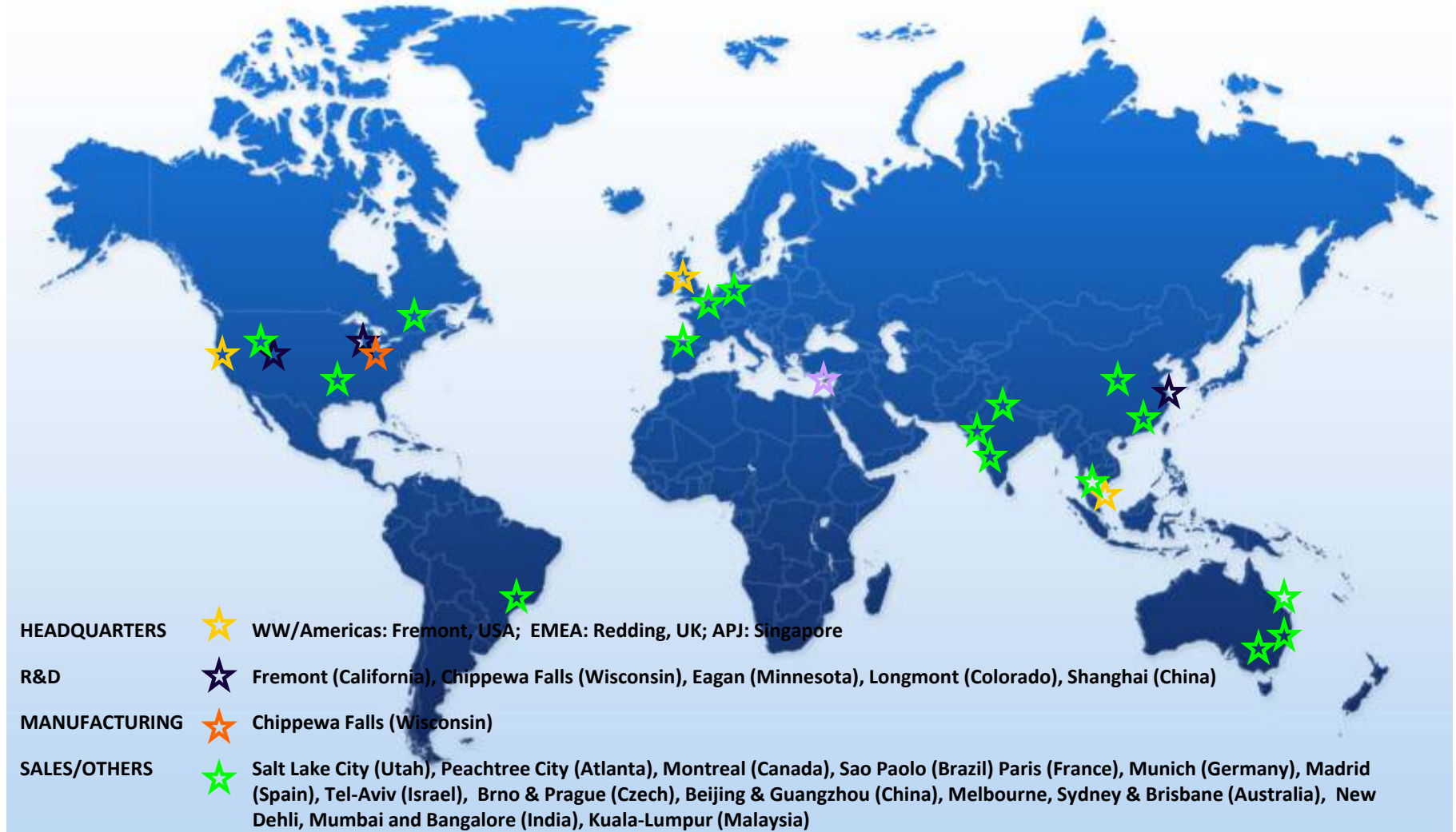
## KEY FACTS

- Public (NASDAQ: SGI)
- HQ: Silicon Valley
- Customers: 6,000
- Employees: 1,300+
- Patents: 700
- Global: 55 countries
- Financially Strong
- Debt Free
- \$450M of Assets

## GROWTH

- \$9B TAM in Server
- \$13B TAM in Storage
- Market is growing at 6.5%
- New market-leading products
- Strong technology cycle
- Increase our service contract attach rates
- Packaged PS Offerings

# Sgi Around the Globe



# SGI Has Trusted Answers To Industry Needs

Focused on the major platforms that enable Technical Computing

## COMPUTE

Scale Out  
Scale Up

## STORAGE

Block / File  
Entry, Mid,  
High- End

## SOFTWARE

Management  
Performance

## DATACENTER

Containers  
Infrastructure

- Leading solutions from cloud computing to big memory
- Emerging leadership in cloud and persistent data storage
- Integrated best-of-breed compute, storage and networking
- Open software platform
- Platforms managed with SGI Management Center
- Industry-leading custom engineering BTO process

# SGI Compute Strategy

Leading solutions from scalable entry to hyper-scale and cloud

## Large-scale Datacenter

CR FND XE



Cloud Inspired  
Hyper-scale  
Eco-Logical  
Density  
BTO

## Shared Memory

UV 1000

UV 100

UV 10



Large Memory  
Big Data  
Fast I/O

## Scalable Entry

O3 X2



Office friendly  
Self-contained  
Scalable  
Low IT needs

Origin®  
400



## High Performance InfiniBand

XE ICE



Capability  
Capacity  
Cost-Optimized  
Multi-Topology  
Choice

# SGI Storage Strategy

Leading solutions from cost to scalability and performance

## Cloud Systems

IS1000



IS2000



Short-lived data  
Cost optimized  
Redundancy  
Software RAID

## RAID Systems



Entry Level (IS220/5000)

Price/Performance Leader



Enterprise RAID (IS4000)

Balanced Price/Performance

HPC RAID (IS6000/15000)

Ultimate Performance/Throughput

## Integrated Storage Servers

NAS 50/100



IS3500



File serving  
App Appliance  
Cost or Performance  
optimized

## Persistent Data

COPAN™



Spectra  
Logic



Long life data  
Disk or Tape  
Large Capacity  
Eco-logical  
High Density

Software: File systems (Lustre, CXFS™, XFS®), DMF, LiveArc™



# Rackable / CloudRack



- 1 Industry's Most Flexible & Configurable Platform**  
Supports low/high wattage Intel and AMD through SSD
- 2 Built-to-Order**  
Configure the platform based on the customer's work load
- 3 Datacenter Optimized**  
Cooling, power, layout and facility costs are top of mind
- 4 FLOPS per SQ per Watt Optimized**  
High density and energy efficient are pre-requisites for scale
- 5 Cloud Inspired (public and private)**  
Amazon EC2/S3, eBay, BT, Microsoft, Intuit, Shopzilla, NSA



# Altix ICE



- 1** World's fastest distributed memory computer  
Base on SPECmpil. Up to dual IB channels per node.
- 2** Scalable  
Supports up to 131,072 nodes, 1 Million + Cores
- 3** Open  
Runs Standard Linux, Intel Xeon 5600 or AMD Opteron 6100 CPUs
- 4** New Topologies  
Hypercube, enhanced hypercube, fat-tree

# Altix UV



Altix UV 1000

- 1** World's fastest shared memory computer  
Base on SPECint and SPECfp, and STREAMS
- 2** Scalable  
Single system image up to 2048 cores and 16TB memory
- 3** Open  
Runs Standard Linux, Intel Xeon 7500 Processors
- 4** New Markets  
HPC, Large Databases, Scalable I/O, RISC replacement

# COPAN



COPAN 400M

- 1 Long-life persistent data storage**  
Disk is better than tape
- 2 Eco-logical**  
High density (up to 3x the capacity per sq ft)  
Energy Efficient (up to 10x the power savings)
- 3 Open**  
Runs Linux, Industry standard VTL/D2D packages,  
and uses standard SATA technology
- 4 Wide Appeal**  
Every data center needs one !

# Modular Data Centers



- 1** Self-contained datacenter  
Power distribution, cooling, safety
- 2** Eco-logical  
Achieving PUEs of 1.1 or better
- 3** Eco-nomical  
1/5 the cost of a traditional datacenter
- 4** Simple and easy to deploy  
Live in 5 days

# Services Complete Our Solutions

Delivering customer value and accelerating their time to results

Packaged Offerings

Consulting

Onsite

Deployment

HW Support

## Solutions

- Data Storage
- Containers
- Reality Centers
- Assessments
- In factory (CSP/I)
- Benchmarking
- Training / Education
- Platform Migration
- 3<sup>rd</sup> Party Product

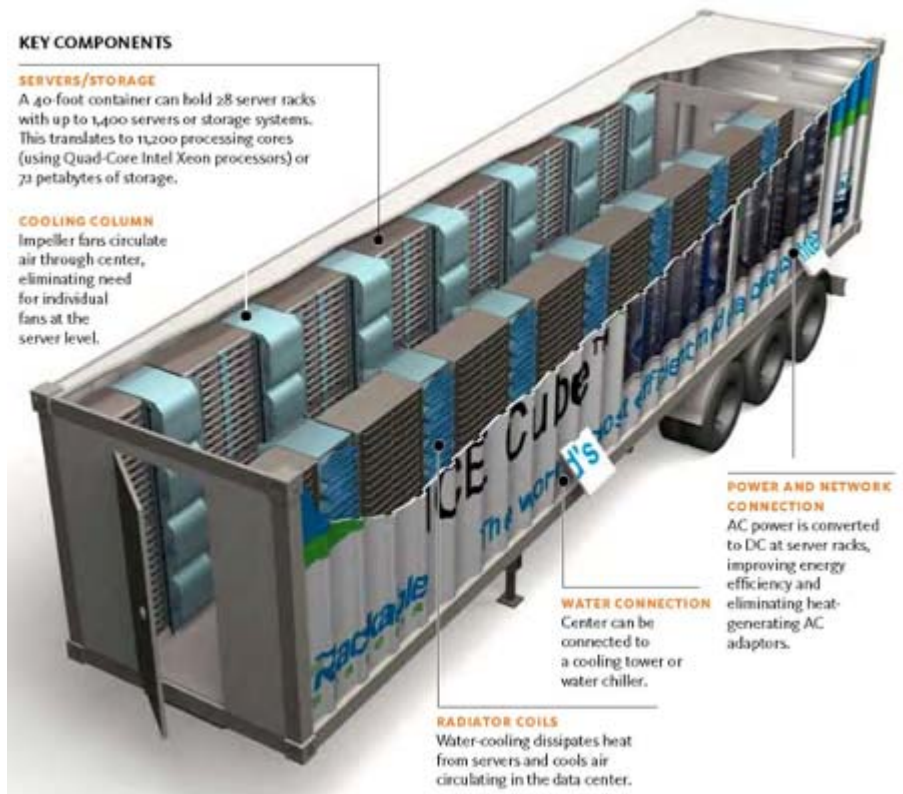
## Recognition

- Global Call Centers
- 400+ Professionals
- 26 Countries
- 24x7x52

# Introducing ICE Cube



## Thinking outside the box...with a Cube!



### KEY COMPONENTS

#### SERVERS/STORAGE

A 40-foot container can hold 28 server racks with up to 1,400 servers or storage systems. This translates to 11,200 processing cores (using Quad-Core Intel Xeon processors) or 71 petabytes of storage.

#### COOLING COLUMN

Impeller fans circulate air through center, eliminating need for individual fans at the server level.

#### POWER AND NETWORK CONNECTION

AC power is converted to DC at server racks, improving energy efficiency and eliminating heat-generating AC adaptors.

#### WATER CONNECTION

Center can be connected to a cooling tower or water chiller.

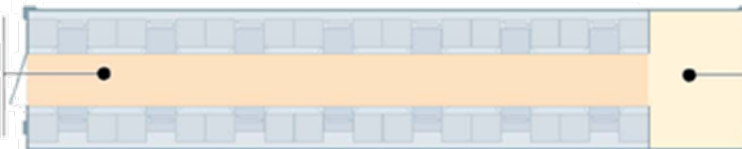
#### RADIATOR COILS

Water-cooling dissipates heat from servers and cools air circulating in the data center.

### BREATHING ROOM

#### CENTRAL AISLE

36-inch aisle provides easy access to systems.

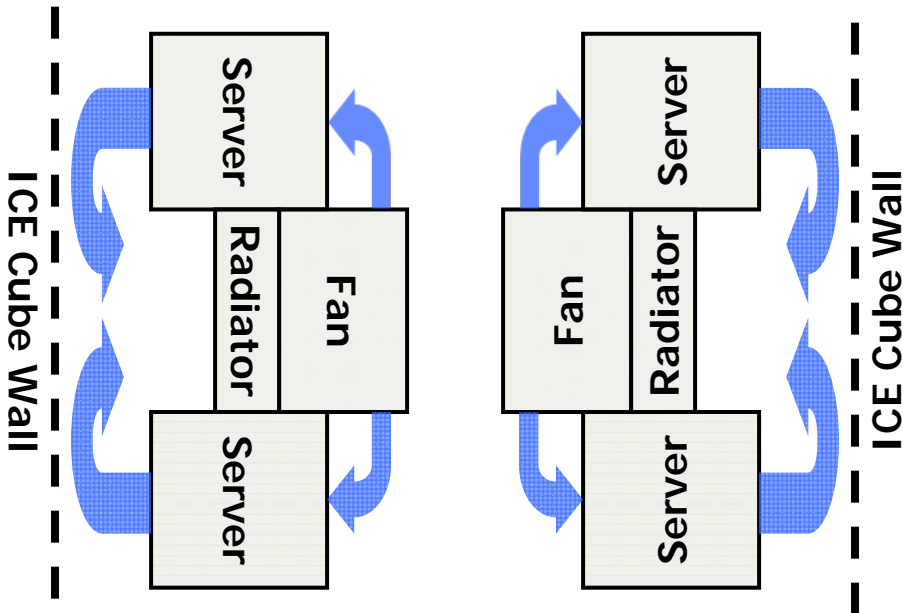


#### VESTIBULE

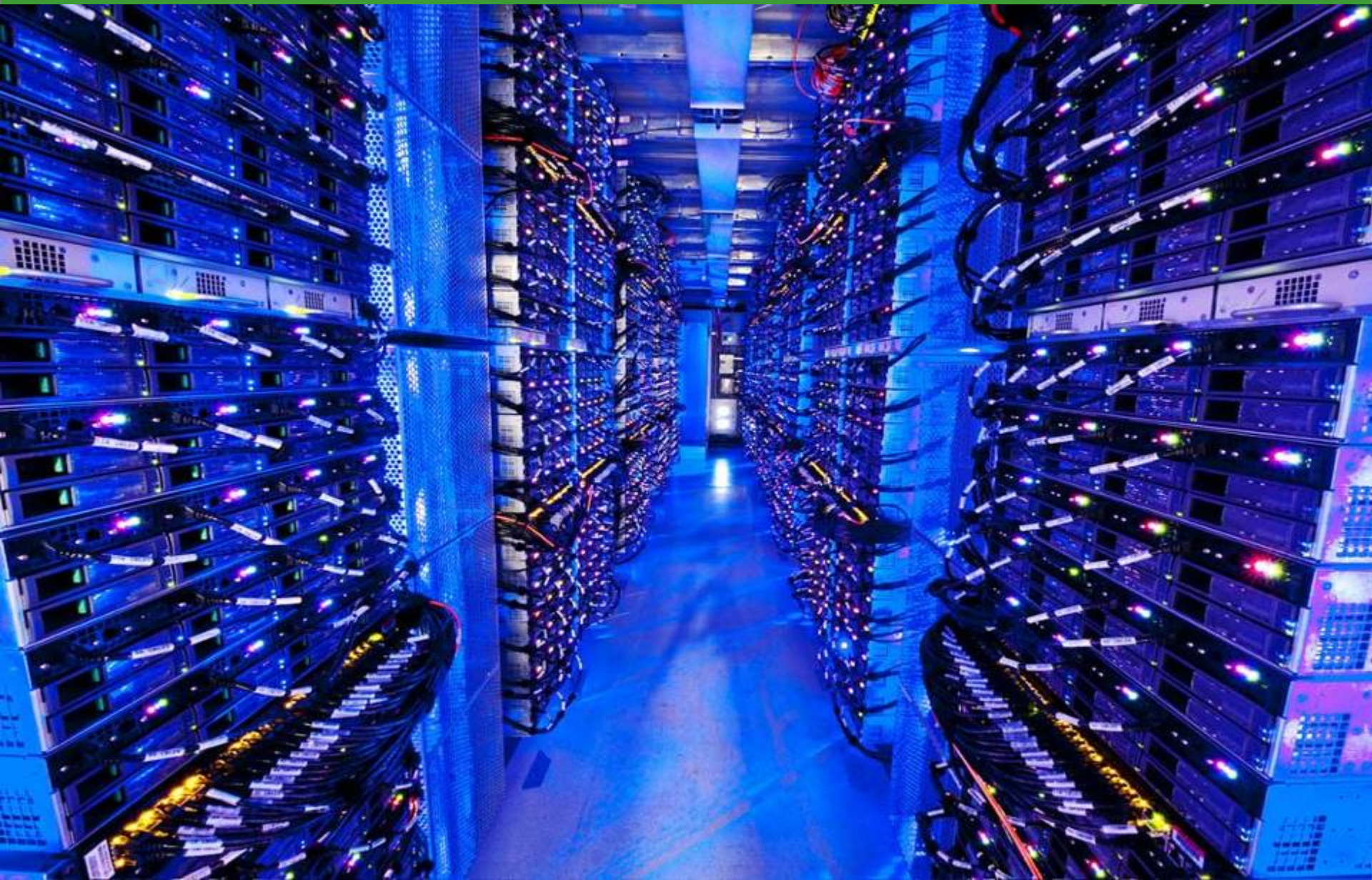
Large area provides space for IT administrators to work or store gear.

# Dual Row: Advanced Cooling Design

- ICE Cube has water supply and return lines
- Fans draw air through radiators between each rack
- Air is cooled immediately before passing through the servers
- Tight integration allows for higher water loop temp and reduced air handler power usage



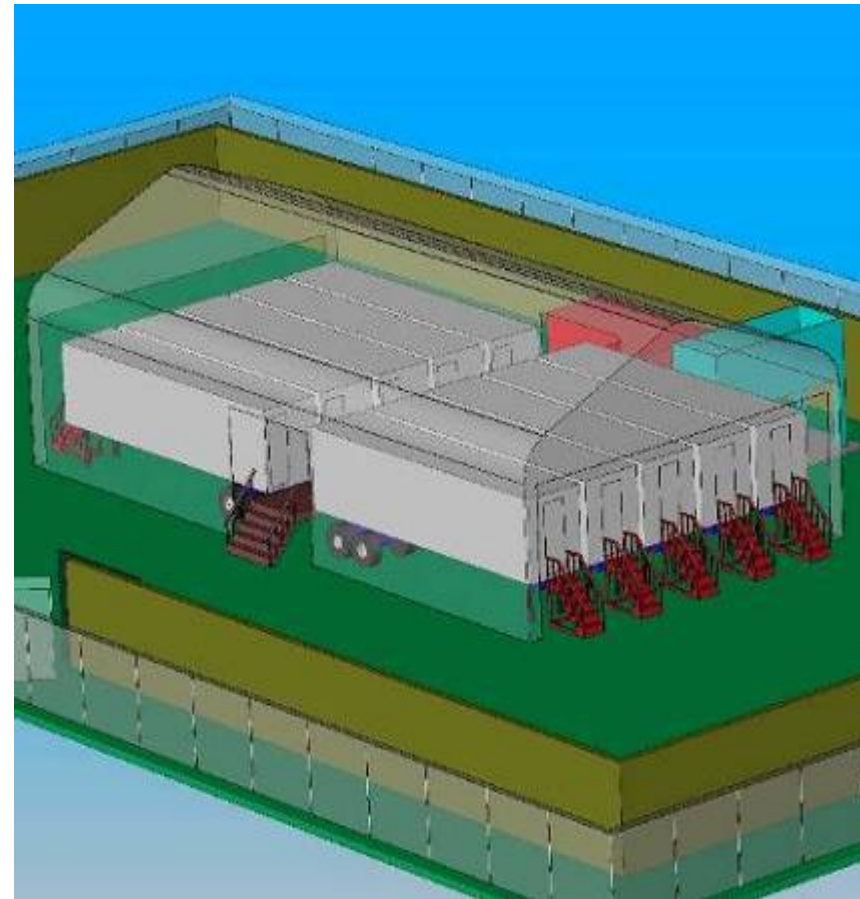
# A Look Inside



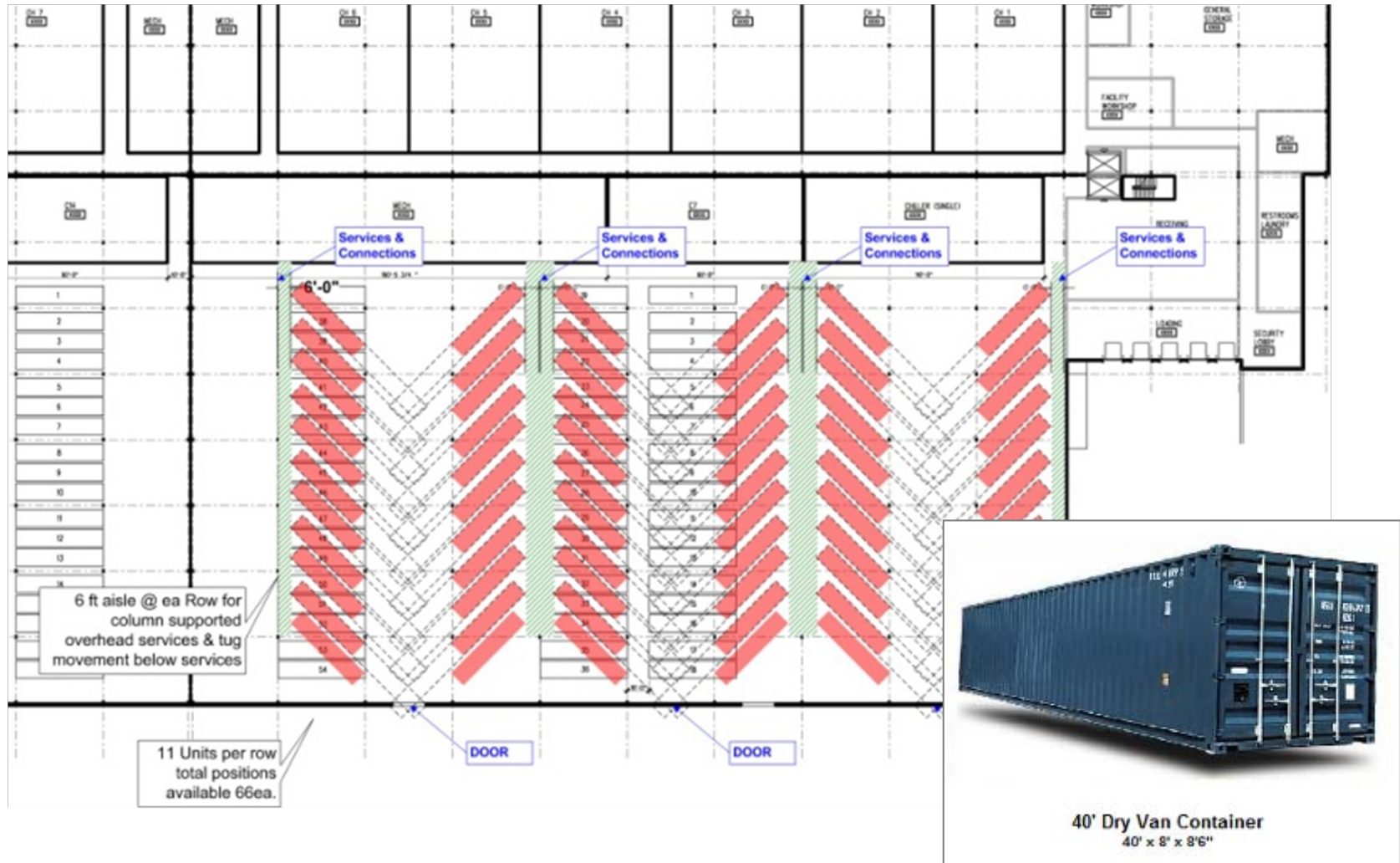


# Construction Site Scalability & Flexibility

- Deploy multiple ICE Cube containers in a data park (stackable)
- Increased geographic options
  - Nontraditional site locations
  - Redundancy through diversity
  - Harness regional strengths
- Rapid deployment
  - Seasonal usage
  - Business continuity
  - Disaster recovery
  - Leverage existing infrastructure
  - Redeploy as needed



# Modular Data Center Site of the Future



# Ideal Deployment Location Can Improve PUE

- Many locations have a ready source of  $<65^{\circ}\text{F}$  water. Big opportunity to cut cooling costs.
- Example: Lake Michigan Water Temp (right) is  $<65^{\circ}\text{F}$  most months. Rarely requires actively running a chiller.

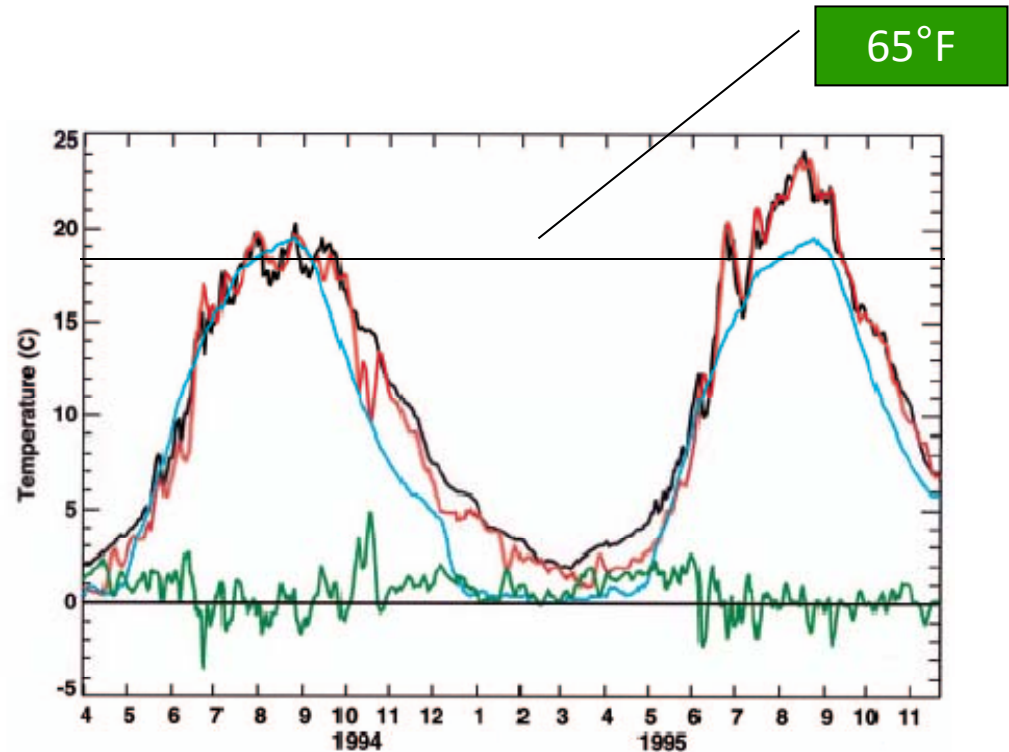


Plate 4. Time series of climatological (blue line), observed (red line), and modeled (black line) lake surface temperature in 1994–1995. Green line represents the difference between modeled and observed temperature.

# Dual Row: Hybrid Container in Production



# Universal Air Container: Outside



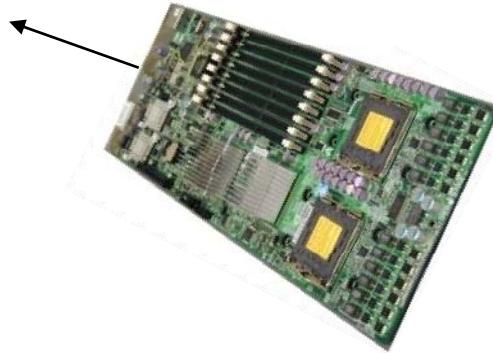
- Ships as three modules
  - IT Module, Adiabatic Cooler, Transformer
- Up to (8) 34.3kW self-cooled roll-in 44U racks (280kVA)
  - Up to (8) CloudRack C2 or (6) Altix ICE or (6) Altix UV 1000
- 2-Stage Adiabatic Cooler (20%, 40%, or 60%)
- Enables PUE < 1.07

# SGI® Altix® ICE 8400

## Designed for High-Performance Computing



**Performance Density: Up to 1536 Cores and  
14.13 TFlops per Rack / 8.3 ft<sup>2</sup> (0.77 m<sup>2</sup>)**



**SGI® Altix® ICE Compute Blade**  
Up to 12-Core, 96GB, 2-IB



**SGI® Altix® ICE Compute Blade**  
Up to 24-Core, 128GB, 2-IB



### Altix ICE Rack:

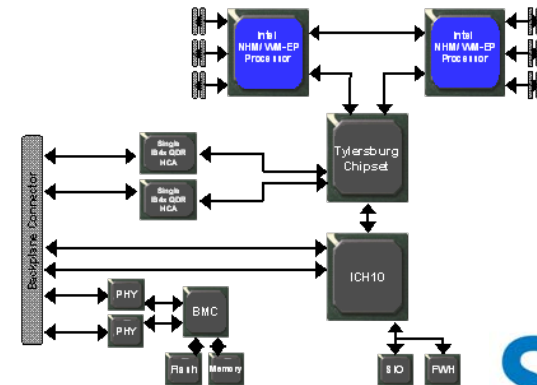
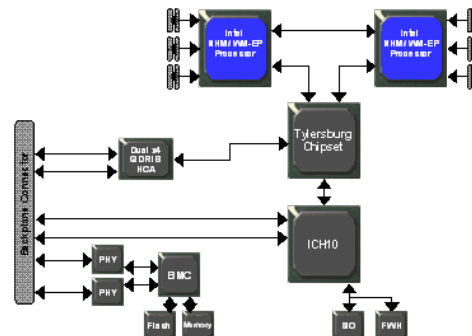
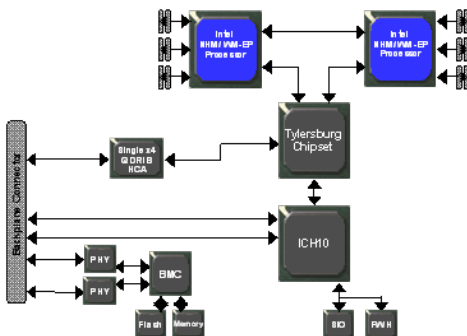
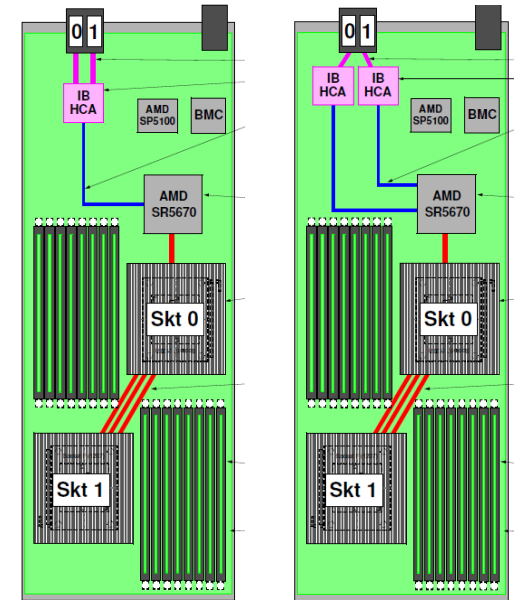
- 42U rack (30" W x 40" D)
- 4 **Cable-free blade** enclosures, each with up to 16 2-Socket nodes
- Up to 128 DP Intel® Xeon® or AMD Opteron™ 6100 sockets
  - Single-plane or Dual-plane IB QDR interconnect
- Minimal switch topology simplifies scaling to 1000s of nodes

World record benchmark result of 51.3 for Altix ICE 8400 on  
SPECmpil\_2007!

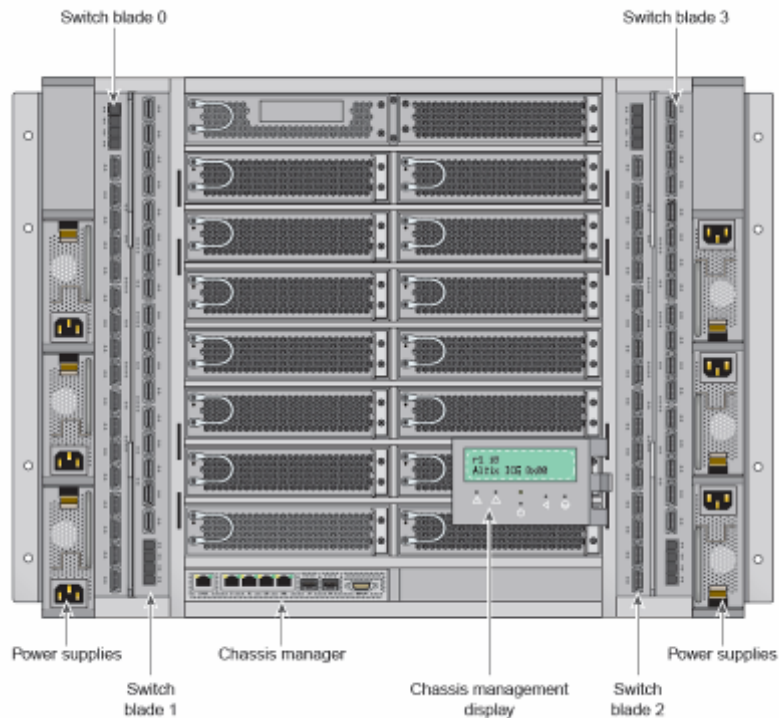


# Flexible Compute Blade Options

- Intel® Xeon® 5500/5600 or AMD Opteron™ 6100 processors
- Intel blades feature 12 DIMM slots and up to 768 cores/cabinet. Up to 130W processors are supported.
- AMD blades feature 16 DIMM slots and up to 1536 cores/cabinet\* Up to 105W processors are supported.
- Choice of three on-board Mellanox® ConnectX-2 InfiniBand HCA configurations
  - Single-port, dual-port or two single-port chipset(s)
- Option for 2.5" storage on the node (SSD and/or HDD)



# Flexibility in Networking Topologies



**Robust integrated switch blade design enables industry-leading bisectional bandwidth at ultra-low latency!**

- **Hypercube Topology:**
  - Lowest network infrastructure cost
  - Well suited for "nearest neighbor" type MPI communication patterns
- **Enhanced Hypercube Topology:**
  - Increased bisectional bandwidth per node at only a small increase in cost
  - Well suited for larger node count MPI jobs
- **All-to-All Topology:**
  - Maximum bandwidth at lowest latency for up to 128 nodes
  - Well suited for "all-to-all" MPI communication patterns.
- **Fat Tree Topology:**
  - Highest network infrastructure cost. Requires external switches.
  - Well suited for "all-to-all" type MPI communication patterns



# Hierarchical System Management



## System Administrative Controller



## Rack Leader Controller

Boot  
Roc



## IRU Chassis Mgmt Controller

IRU mgmt  
OS Synchronization



## Service Nodes



Options: "6016", XE270/500 & UV10  
Optional NVIDIA® GPU Support:  
Quadro® FX 3800/ 4800/ 5800,  
Tesla™ C1060/ C2050\*/ S1070/ S2050\*

- Isolate components, management and run-time functions. Easily hot swap components.
  - Previous three generations can all be cabled together under single system manager
  - Management framework scales seamlessly, allowing easy addition of enclosures and racks to an existing system
    - Service Nodes are "peers" in the system and be can be scaled independently of compute nodes matching customer requirements

# SGI Altix ICE- Industry Breakthrough Compute Rack Level 'Live' Integration

**NASA Ames Post - NAS TECHNICAL HIGHLIGHTS** February 8, 2010

## 'Live' Integration of Pleiades Rack Saves 2 Million Hours (excerpt)

The new 512-core *rack arrived in late December and installation was completed in early January*. Integration into the Pleiades system was accomplished by connecting the new rack's InfiniBand (IB) dual port fabric via 44 fibre cables-*while* Pleiades was *running a full production workload*.

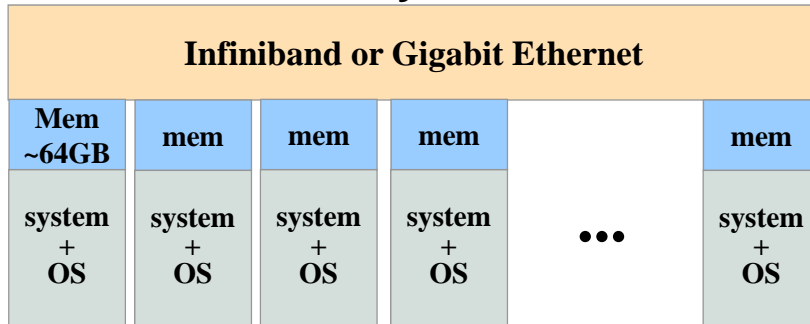
*This live integration saved 2 million hours in productivity* that have previously been lost each time a planned system outage occurs. When outages on Pleiades are planned, users get a one-week notice and system utilization plummets about three days before the actual shutdown. This drop in usage is partly due to the fact that batch jobs are only started if they can finish by the start of the planned outage. About half of Pleiades' computational hours are consumed by long-running jobs-most take five days to complete-further adding to the usage slowdown.

<http://www.nas.nasa.gov/News/TechHighlights/2010/2-8-10.html>

- **SGI's superior hypercube based IB network topologies not only enables adding nodes and switches but also now enables adding racks of nodes and switches without disturbing the existing production load.**
- **Competitor network topology offerings such as fat tree and 3D torus are either inherently limited or strictly incapable of supporting such a dynamic reconfiguration.**

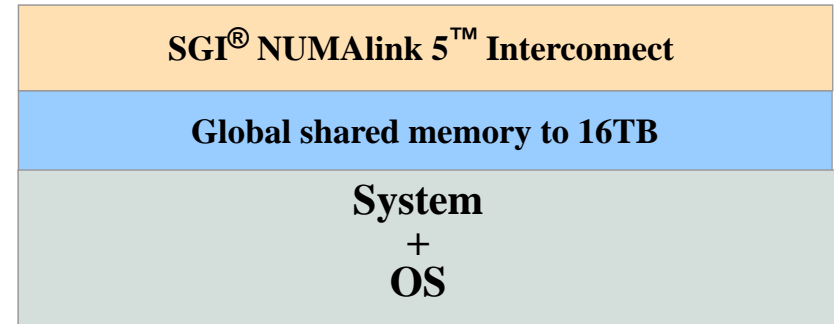
# SGI Altix UV Shared Memory Architecture

## Commodity Clusters



- Each system has own memory and OS
- Nodes communicate over commodity interconnect
- Cross-node communication creates potential bottlenecks
- Coding required for parallel code execution

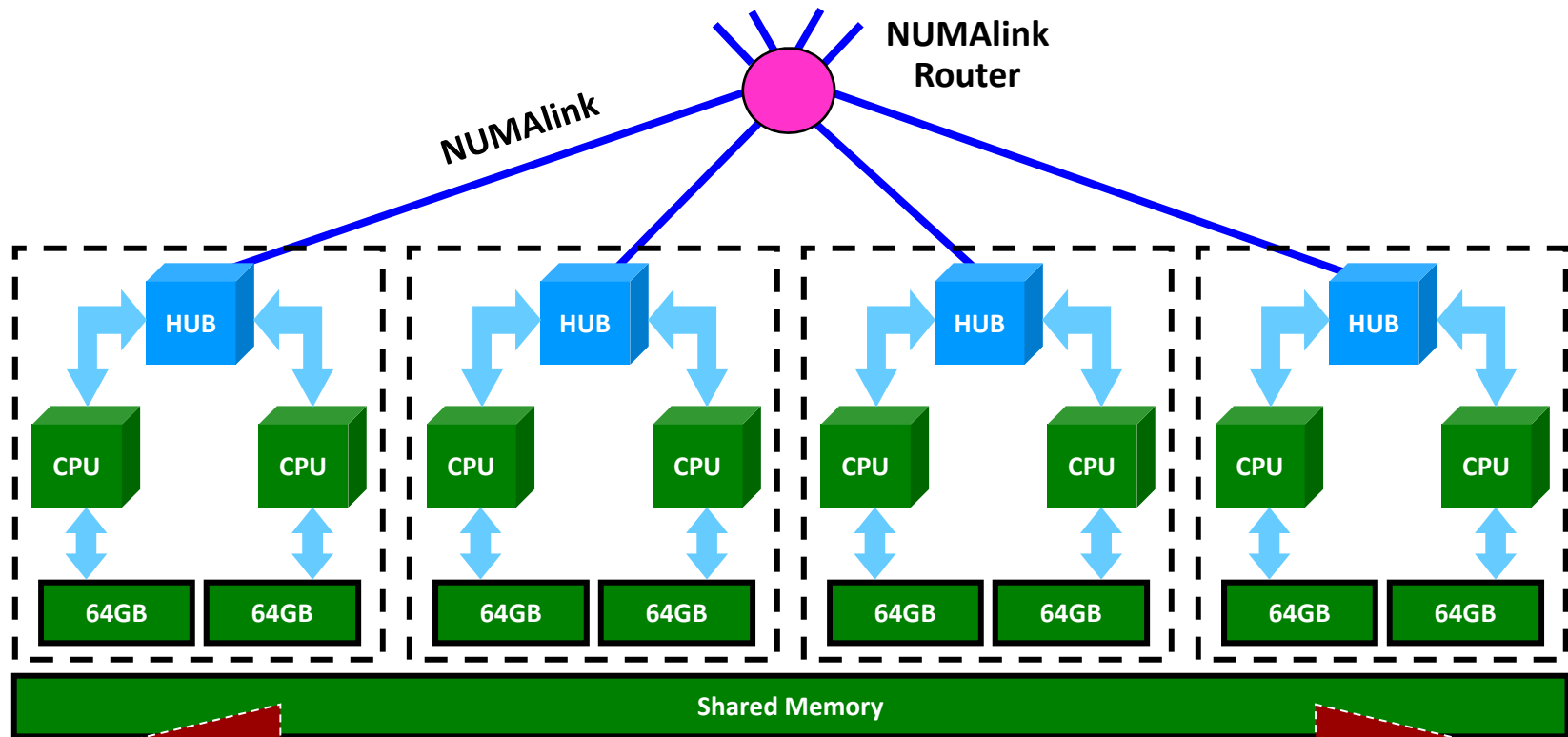
## SGI® Altix® UV Platform



- All nodes operate on one large shared memory space
- Eliminates data passing between nodes
- Big data sets fit entirely in memory
- Less memory per node required
- Simpler to program
- High Performance, Low Cost, Easy to Deploy

# Globally Shared Memory System

- NUMALink<sup>®</sup> 5 is the glue of Altix<sup>®</sup> UV 100/1000



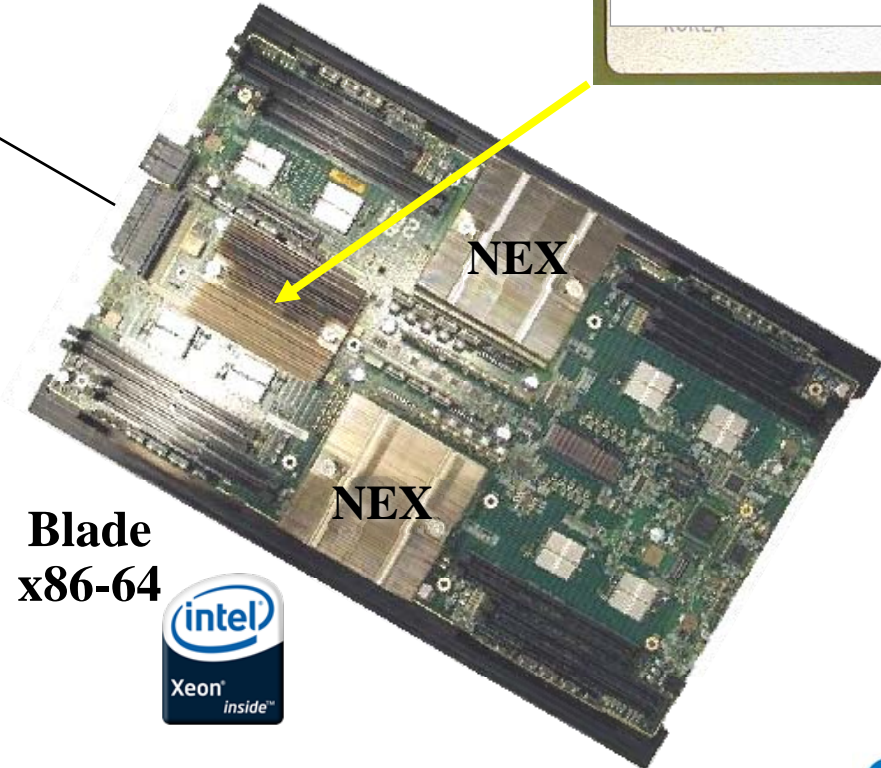
Up to 16 TB Global Shared Memory

# Altix® UV : 2 different ways of using it



**UV-NIC**

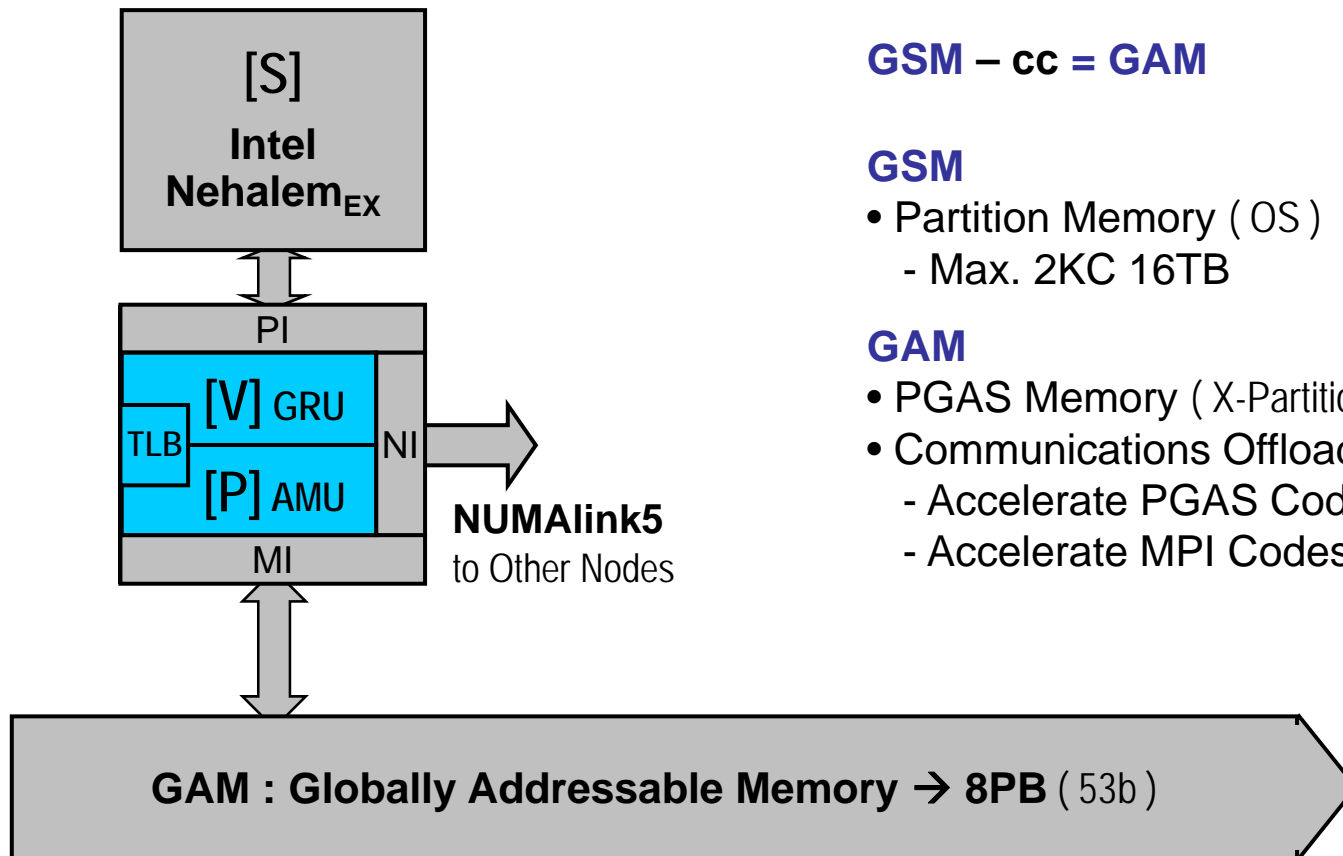
- 16TB GSM
- 8PB GAM



**Blade  
x86-64**



# UV Foundation : GAM + Communications Offload



**GSM – cc = GAM**

## **GSM**

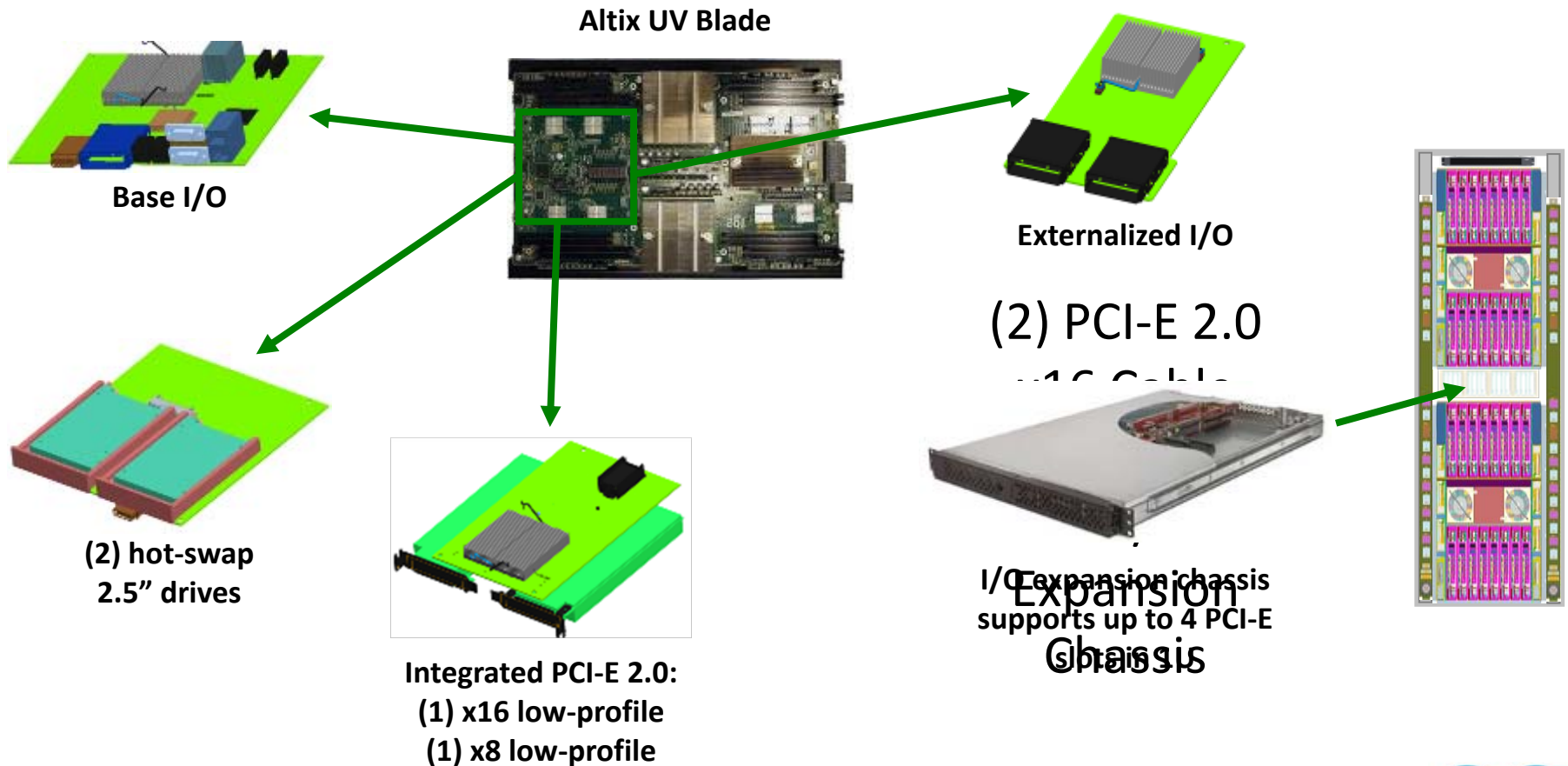
- Partition Memory ( OS )
  - Max. 2KC 16TB

## **GAM**

- PGAS Memory ( X-Partition )
- Communications Offload ( GRU + AMU )
  - Accelerate PGAS Codes
  - Accelerate MPI Codes ( MOE v.v. TOE )

# I/O Expansion Options

- Four I/O riser choices offer configuration flexibility

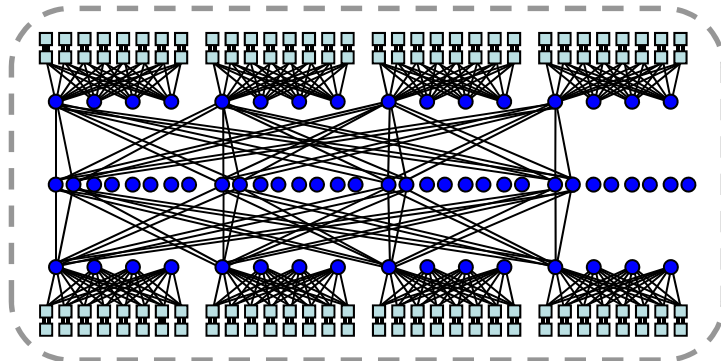


# Scalability: Architectural Limits

- Altix® UV's architecture supports scaling to Petaflop level
- 256-socket fat tree groups in 8 x 8 torus
  - 4-rack groups x 8D x 8W = 256 racks for 16,384 sockets is illustrated
- Upper limit on scaling is the Altix UV hub, capable of connecting 32,768 sockets

## Petaflop System

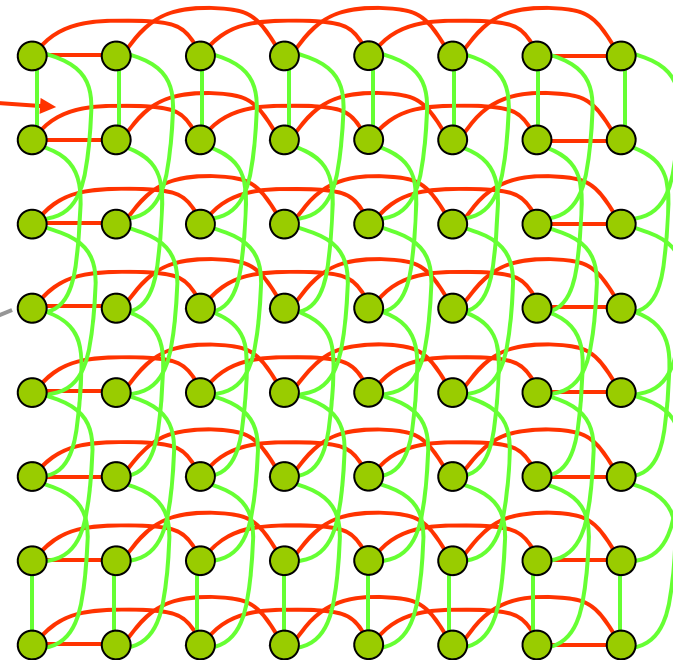
(8) L1Rs per plane  
(8 of 16) ports / L1R support Fat Tree  
(8 of 16) ports / L1R support 2-copies of Torus  
(16) copies of Torus per plane



256-Socket Fat Tree Building Block (4 racks)

Each Red & Green Torus link shown is (2) links / L1R

Green Links  
(Interleaved across the aisles)



Red Links (Interleaved down the ranks)

● = 4-Rack Group



# Open Platform

- Altix® UV runs standard x86 applications
  - No need for recompilation or access to source code
- Choice of Novell® SUSE® Linux Enterprise Server (SLES) or Red Hat® Enterprise Linux® operating systems
  - Run out-of-the-box, no modifications
- Altix UV blades provide PCI-E expansion slots compatible with industry-standard cards
  - E.g. storage, networking, graphics cards
- Altix UV supports a large range of storage options, including SGI® InfiniteStorage RAID, EBOD, SAN, NAS, tape and software such as DMF, CXFS® and LiveSAN™

# Application Development Advantages

- Scale problem size without decomposition or rework
  - Minimal penalty to fetch off-node data
- Freely exploit new and existing programming models in any combination or scale
- Ideal for code development and prototyping
  - Avoid the hindrance of cluster paradigms
  - Unified parallel C translator in development
- Enjoy Simplified Load Balancing
  - Direct a task to any processor as all data is accessible
- Application Fusion
  - Complex workflows in Global Addressable Memory

# Altix<sup>®</sup> UV is Ideal for Wide Range of Applications

- Ideal application characteristics include
  - I/O-Bound and memory-bound apps
  - Inter-processor communications intensive apps
  - In-Memory and Large (VLDB) Databases
  - Graphs Traversal, Sort and Inferences
  - MapReduce
  - Apps with asymmetric computational patterns
- A Single System Image (SSI) system like Altix<sup>®</sup> UV is often the perfect complement to large scale-out clusters with Altix UV being the “simulation supernode”

# Performance: World Records

## SPECint rate base2006:

**#1: SGI Altix UV 1000 1024c Xeon X7560 20600**

**#2: SGI Altix UV 1000 512c Xeon X7560 10400**

#3: SGI Altix 4700 Bandwidth System 1024c Itanium 9030

#4: Sun Blade 6048 Chassis 768c Opteron 8384 (cluster) 8840

#5: ScaleMP vSMP Foundation 128c Xeon X5570 3150

#6: SGI Altix 4700 Density System 256c Itanium 2890

## SPECfp rate base2006:

**#1: SGI Altix UV 1000 1024c Intel Xeon 7560 16000**

#2 SGI Altix 4700 Bandwidth System 1024c Itanium 10600

**#3: SGI Altix UV 1000 512c Xeon X7560 6840**

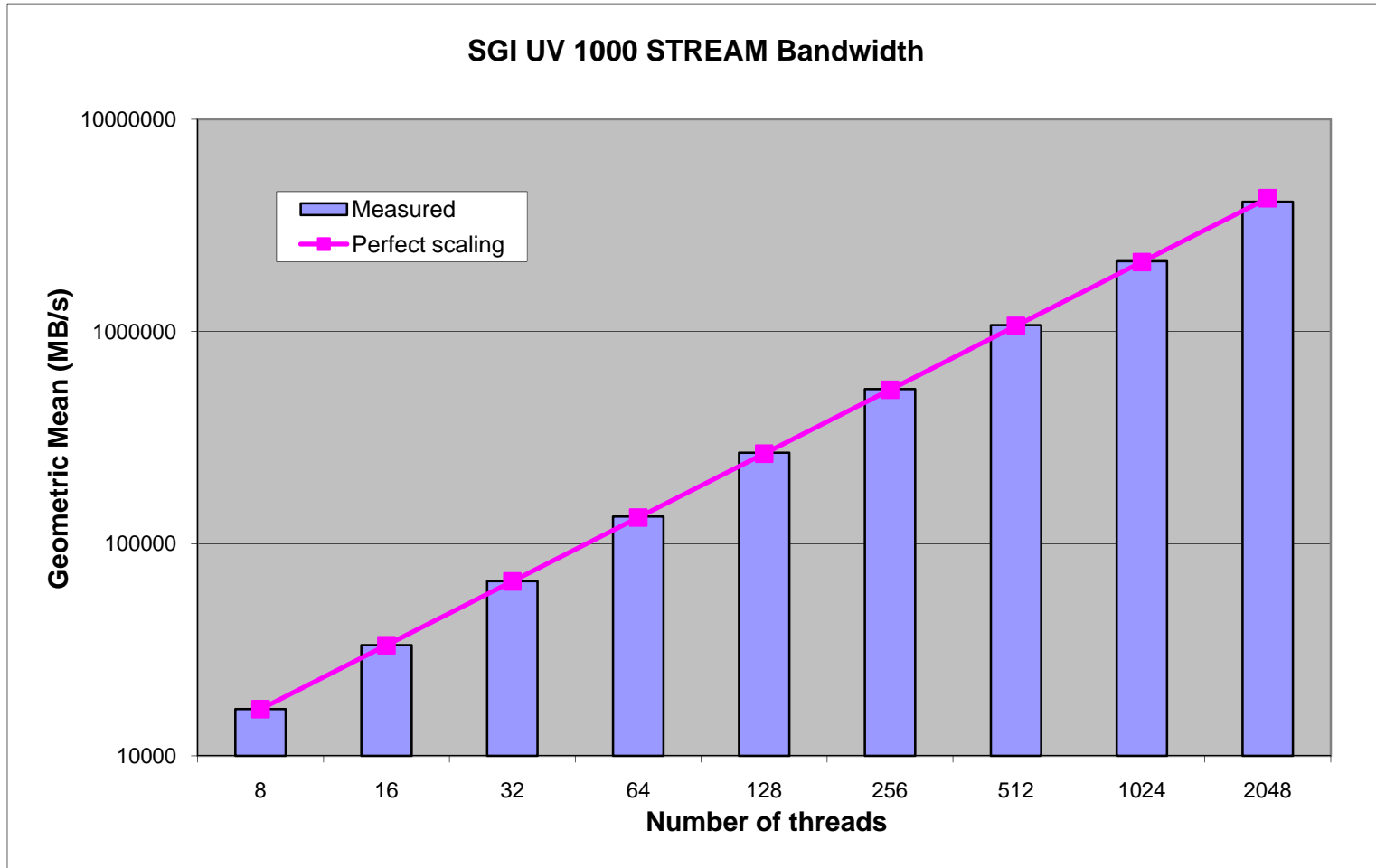
#4: Sun Blade 6048 Chassis 768c Opteron 8384 (cluster) 6500

#5: SGI Altix 4700 Bandwidth System 256c Itanium 3420

#6: ScaleMP vSMP Foundation 128c Xeon X5570 2550

Source: [www.spec.org](http://www.spec.org) (July, 2010)

# World Record Streams Memory Bandwidth



# SGI's World Record Result Summary

## Specjbb

- World record Multi-JVM performance of 12,665,917 BOPS with 128 JVMs using Oracle JRockit 1.6
  - <http://www.spec.org/jbb2005/results/res2010q3/jbb2005-20100616-00867.html>
- World record Single-JVM performance of 2,818,350 BOPS/JVM using Oracle Java HotSpot 1.6
- Above 1M BOPS on the smallest box ever!
  - Single-JVM performance of 1,080,399 BOPS/JVM using Oracle Java HotSpot 1.6 on the smallest box with only 48c (8 6c).

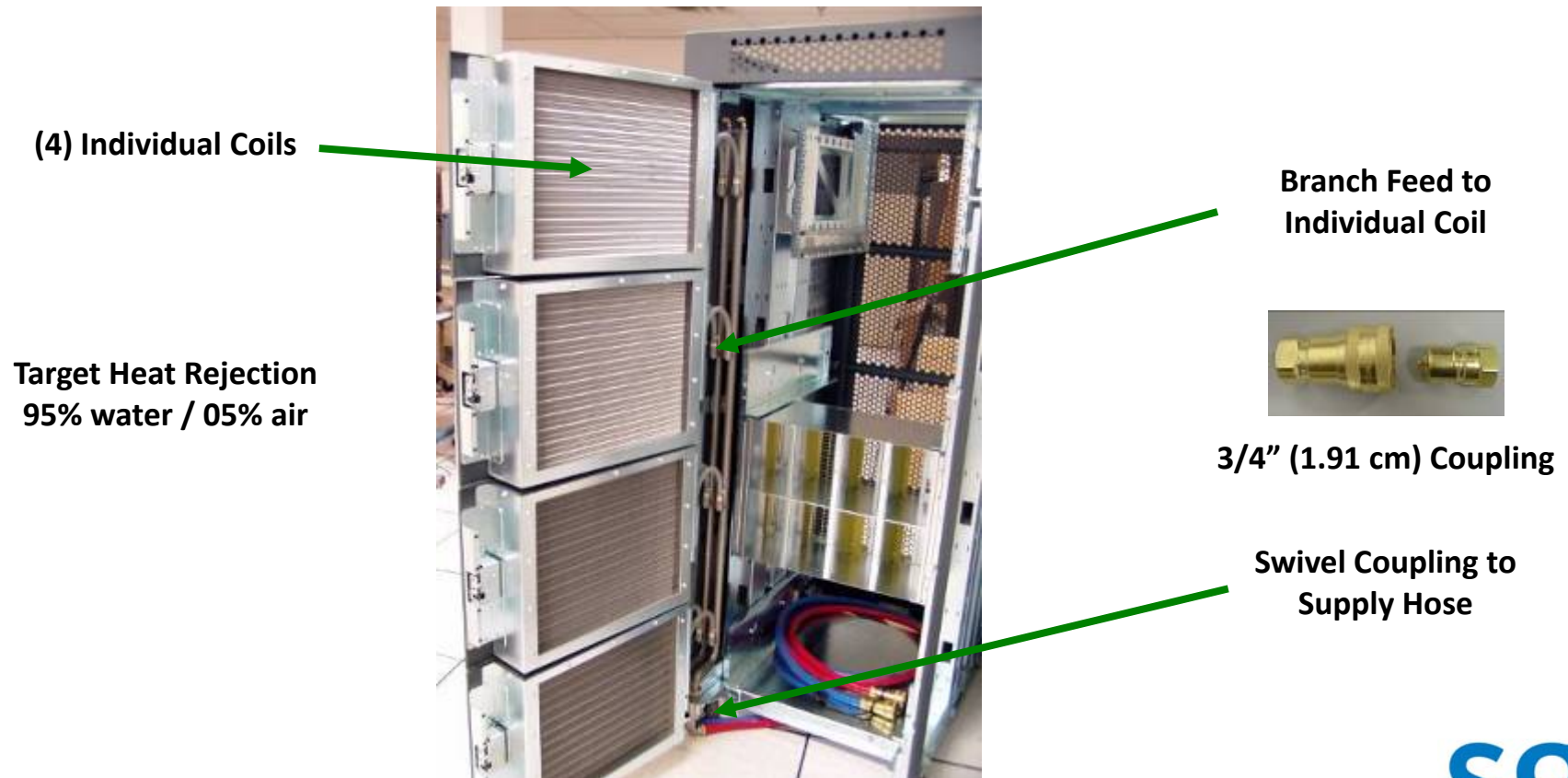
# Eco-Logical™: Energy Efficiency Features



- Leading performance/watt efficiency from SSI
  - Enables deployment of more compute capacity within the same power envelope
- 80 PLUS® Gold certified power supplies
  - 92% efficient at 50% load
- Linear airflow path minimizes fan power
- Variable speed fans controlled by chip temperature sensors
  - Fans at 50% speed draw only 12.5% of their full power
- Supports 2008 ASHRAE TC9.9 Expanded Recommended Environmental Envelope
  - 64.4–80.6°F (18–27°C) dry-bulb temp.
  - Attain reduced data center cooling costs

# Eco-Logical™: Water Chilled Door Option

- By “close-coupling” cooling to the heat source, data center cooling issues can be mitigated





# SGI-Specific RAS Features

Going above and beyond the base functionality originating from Intel® Xeon® 7500 processors (“Nehalem-EX”) and Intel® 7500 chipset (“Boxboro”), Altix® UV also provides the following functionality designed by SGI:

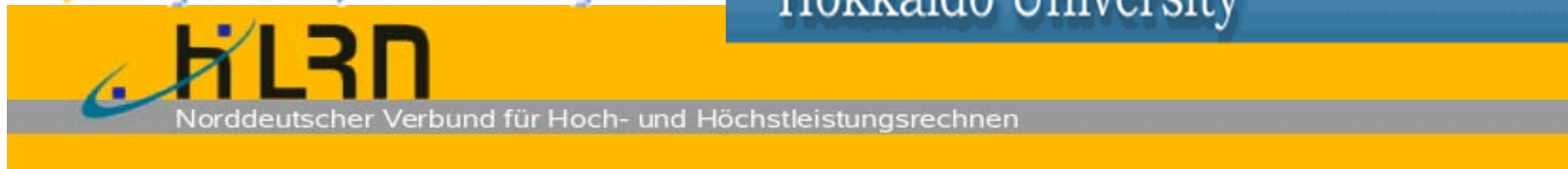
<b>System</b>	<ul style="list-style-type: none"><li>▪ Data path checking (including single bit correct)</li><li>▪ Firmware provisioning</li><li>▪ Redundant chassis controllers</li><li>▪ FRU failure analysis</li><li>▪ Online diagnostics</li><li>▪ Uptime management</li></ul>
<b>Blade Interconnect</b>	<ul style="list-style-type: none"><li>▪ LLP, CRC and retry support</li><li>▪ Hot connect / disconnect</li><li>▪ Lane failover and redundant routing</li><li>▪ Dynamic reconfigurations</li><li>▪ Alpha immune latches</li></ul>

<b>Processors</b>	<ul style="list-style-type: none"><li>▪ Dynamic and boot time isolation</li></ul>
<b>Memory</b>	<ul style="list-style-type: none"><li>▪ DRAM failure analysis</li><li>▪ Page migration</li><li>▪ Boot time disable</li><li>▪ Tiered failure containment</li></ul>
<b>Power and Cooling</b>	<ul style="list-style-type: none"><li>▪ Redundant, hot-swappable power supplies and cooling fans.</li><li>▪ Redundant line cords</li><li>▪ Online fault detection and ACPI support</li></ul>

# Sample of UV Customers



**PITTSBURGH SUPERCOMPUTING CENTER**



# Altix UV Graphics and GP-GPU Packaging

## NVIDIA® Tesla™ or Quadro® Plex Enclosures Up to 8 GPUs per System Partition



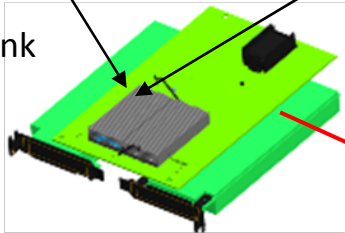
3U

NVidia Tesla unit =  
4 GPU, Two x16 links

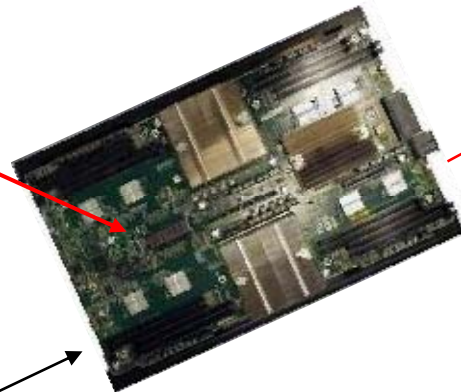


1U

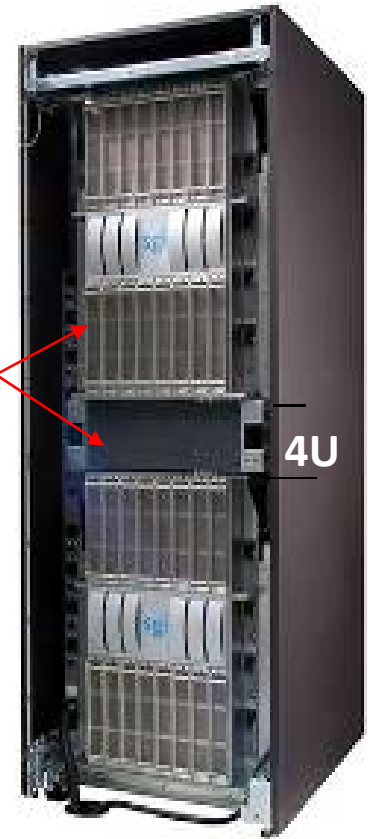
NVidia Quadro Plex  
unit = 2 GPU +  
Gsync, one X16 link  
(2 units shown)



Altix UV 100/1000  
PCIe X16 slot



Each UV 100 or UV 1000 blade can connect up to one  
NVIDIA Tesla or Quadro Plex enclosures (Altix UV 10  
uses NVIDIA host cards to achieve similar connectivity)



# Different Types of Data Demand Different Storage Solutions

## Transactional or

## Persistent Data

## Vaulted Data

### Dynamic Data

- I/O intensive
- Small files
- Modest storage growth
- Steady growth rates

### Data Protection and Archive Data

- Large files
- Very large storage
- Throughput
- Sequential
- Explosive growth

- Offsite
- Copy of copy
- Sequential
- Compliance






**E-mail** **Document** **Database**





8KB      80KB      8MB

**Backup**      **Replication**      **Maps**      **Video**      **Imaging**

10MB      20MB      60MB      300MB      48GB



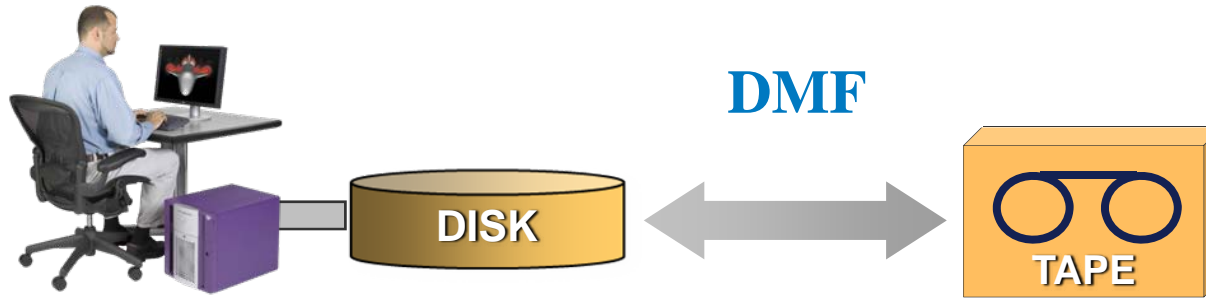
Relative proportions of data in the typical enterprise



Within 30 days the majority of transactional data becomes persistent data



# Storage Software : DMF



## Transactional Storage

Performance —

FC RAID : **DISK** —



**2 - 3 mins**



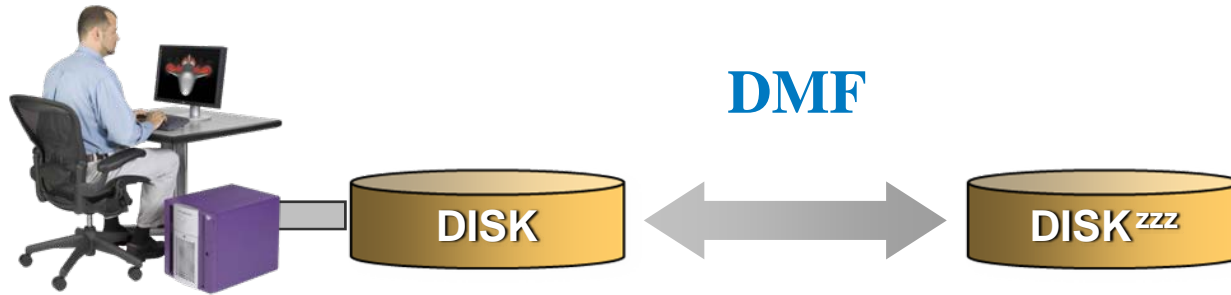
## Persistent Storage

— Density, Cost

— **TAPE**



# Storage : DMF with COPAN MAID



## Transactional Storage

Performance —

FC RAID : **DISK** —



**20 - 40 secs**



## Persistent Storage

— Density, Cost

— **MAID COPAN**

o Smart Sleep

o SATA 3TB

o Aerobics (P)

o Vibro-cancel (P)



# Storage : COPAN MAID 400 Canister



# SGI COPAN 400 Platform Details

## ■ Disk-Based Core Platform

- Enterprise 1TB or 2TB SATA Drives
- 1 to 8 MAID Shelves
- Up to 1,792TB raw storage per cabinet with 2TB drives

## ■ Performance

- Up to 6,400 MB/s (native MAID) or
- Up to 3,200 MB/s (VTL)

## ■ Multiple Solutions

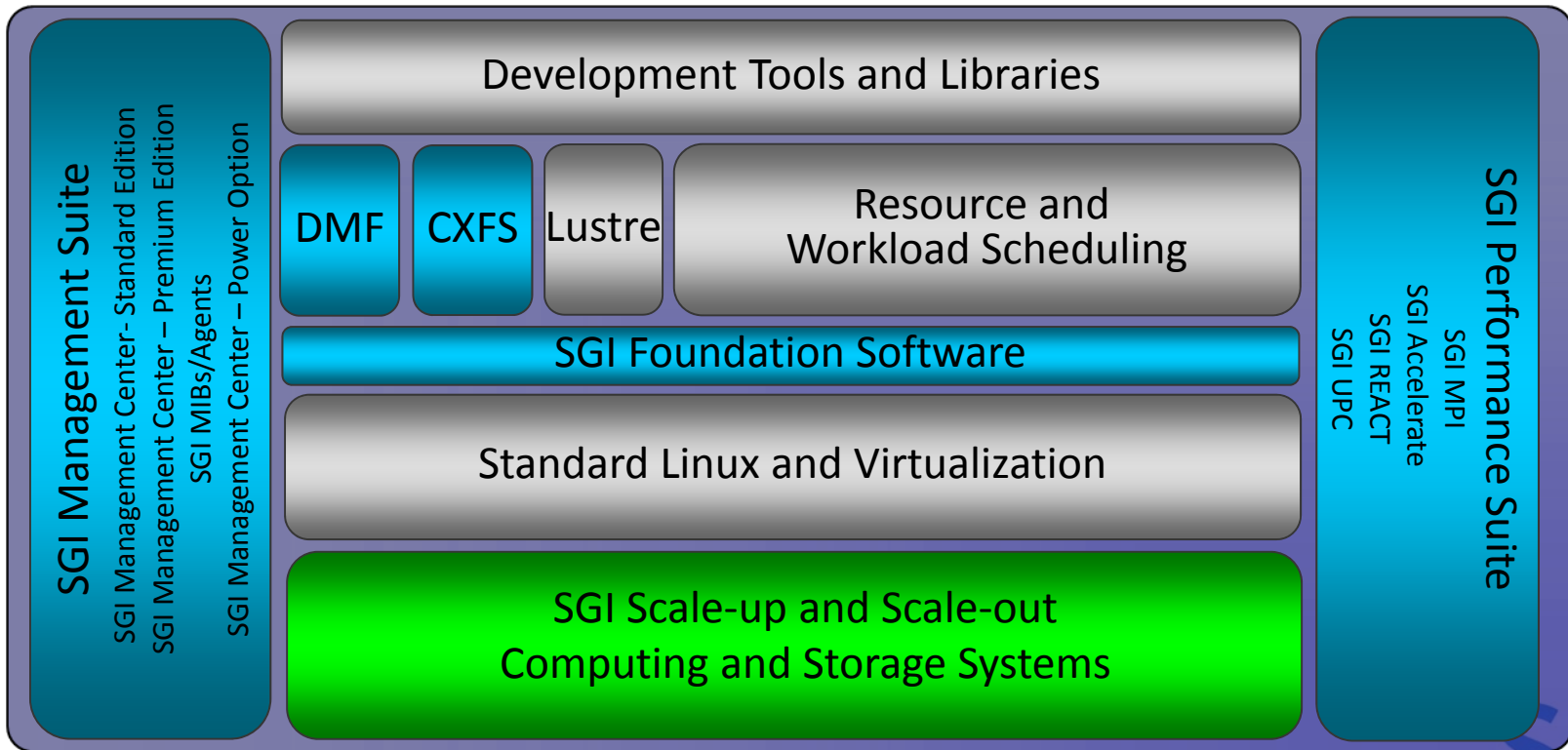
- **Native MAID:** ideal for HSM and D2D applications
- **VTL:** reliable, high performance target for backup applications.





# SGI Technical Computing Software Stack

CSM	CFD	CEM	CCM	BIO	RES	SPI	CWF	SRE	DBA
ANSYS Nastran Abaqus LS-Dyna VPS	Fluent, CFX StarCCM+ OpenFOAM CFD++	FEKO, FMSLIB,	Gaussian, VASP, NAMD Jaguar, Amber CASTEP	BLAST, FASTA, HMMER, ClustalW	Eclipse, Intersect, VIP, Nexus,	ProMAX, EPOS, Geoclusr	WRF, MM5, Aladin,CCSM Hirlam,POP NEMO	MATLAB, R, Mathematica Maple	Oracle, SQL TimesTen, VoltDB, DataRush



SGI HW and SW products
 
 Third party product (available from and/or integrated by SGI)



Lots has happened since 2008 workshop...

sggi<sup>®</sup>

Thank You