# SGI® Update

Michael Woodacre
Chief Engineer
woodacre@sgi.com

# Server Products


**Scale-Out** **Scale-Up**

**Rackable™**
1U, 2U, 3U, 4U & XE

**CloudRack™**
Tray Architecture

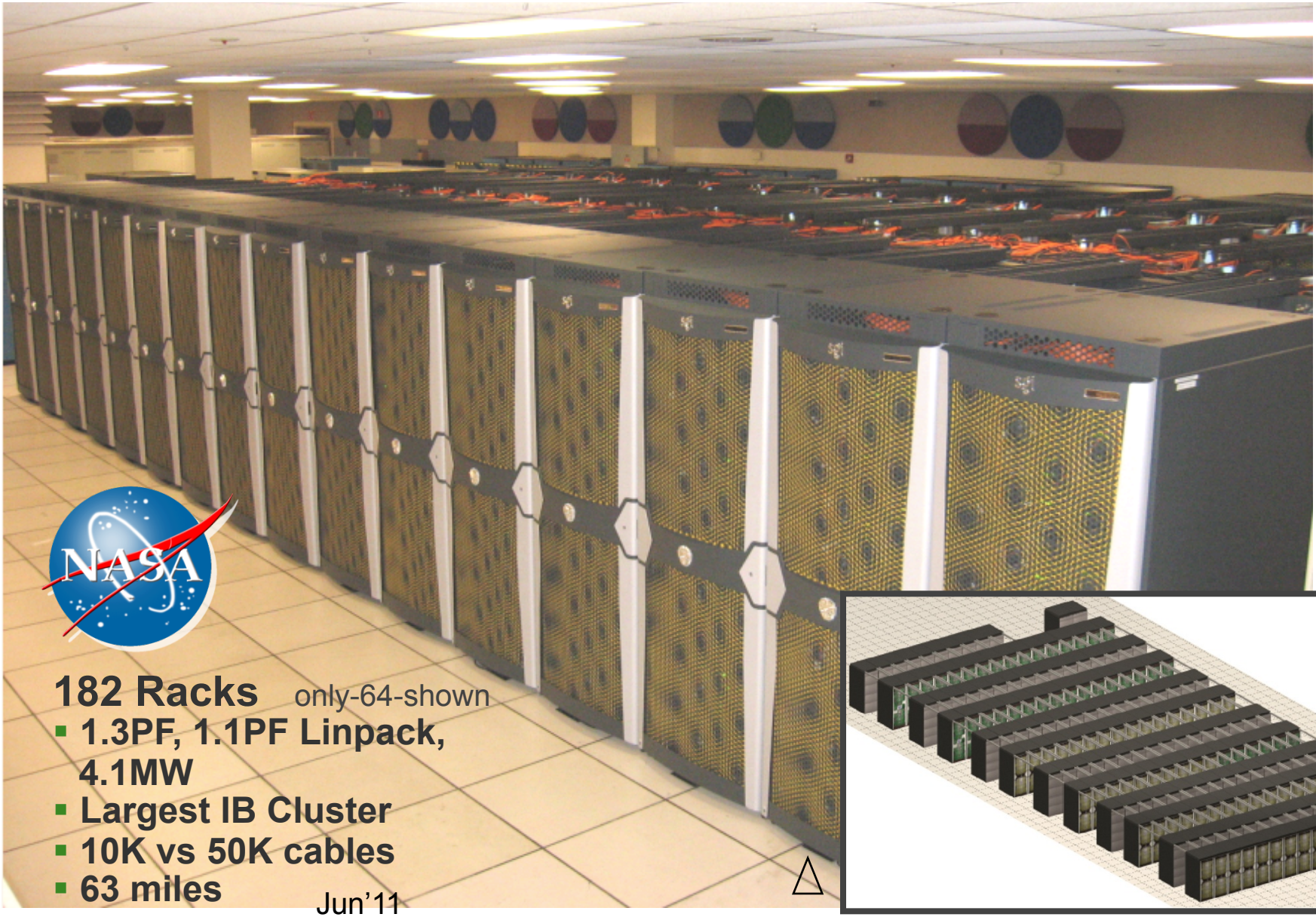
**SGI ICE-X**
Blade Architecture

**SGI UV**
Shared-Memory Systems

# SGI ICE-X

# SGI® ICE : NASA Pleiades

184 Racks = 92 HPT DDR + 20 NHM QDR + 70 WSM QDR + SDB FDR



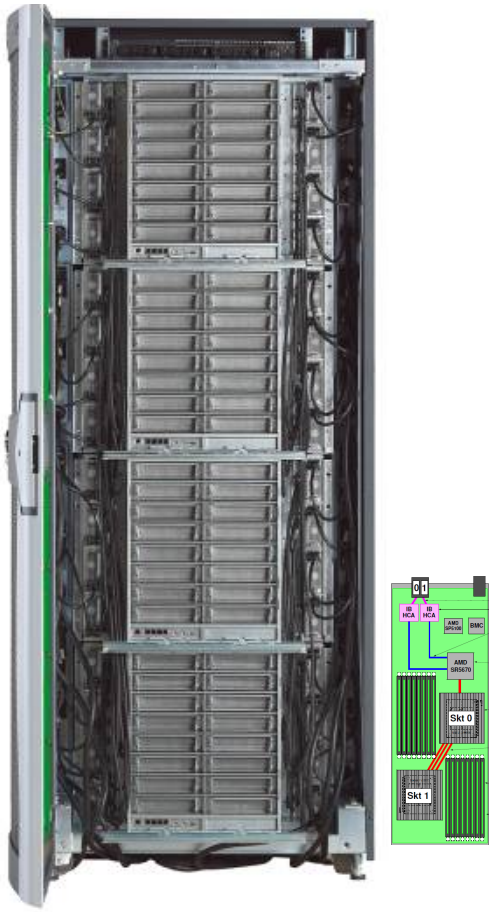
**182 Racks** only-64-shown

- **1.3PF, 1.1PF Linpack, 4.1MW**
- **Largest IB Cluster**
- **10K vs 50K cables**
- **63 miles**

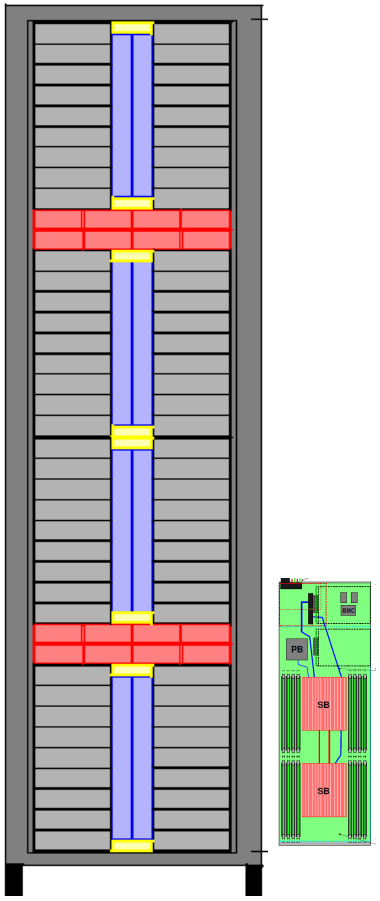
Jun'11

# SGI ICE 8400 → SGI ICE-X

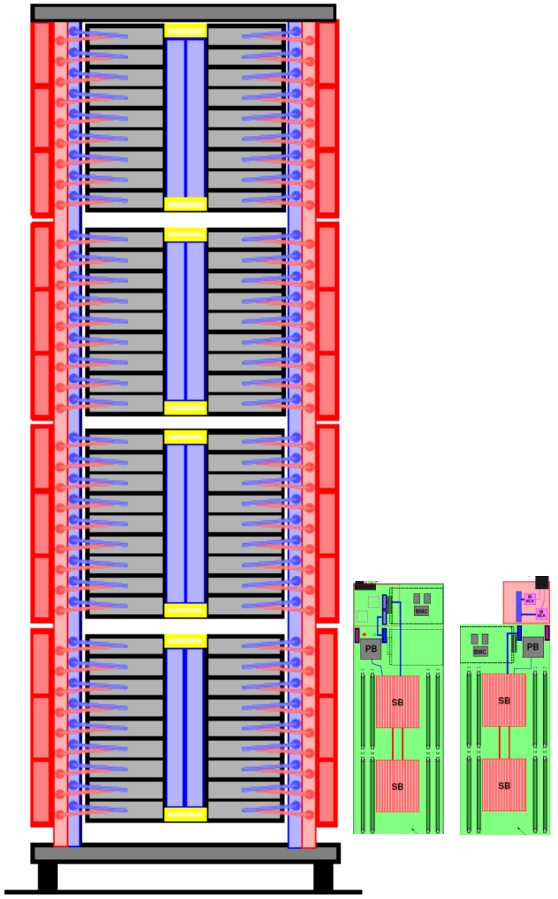
**SGI ICE 8400 = 64N**
(128 Sockets)




**D-Rack = 72N**
(144 Sockets)



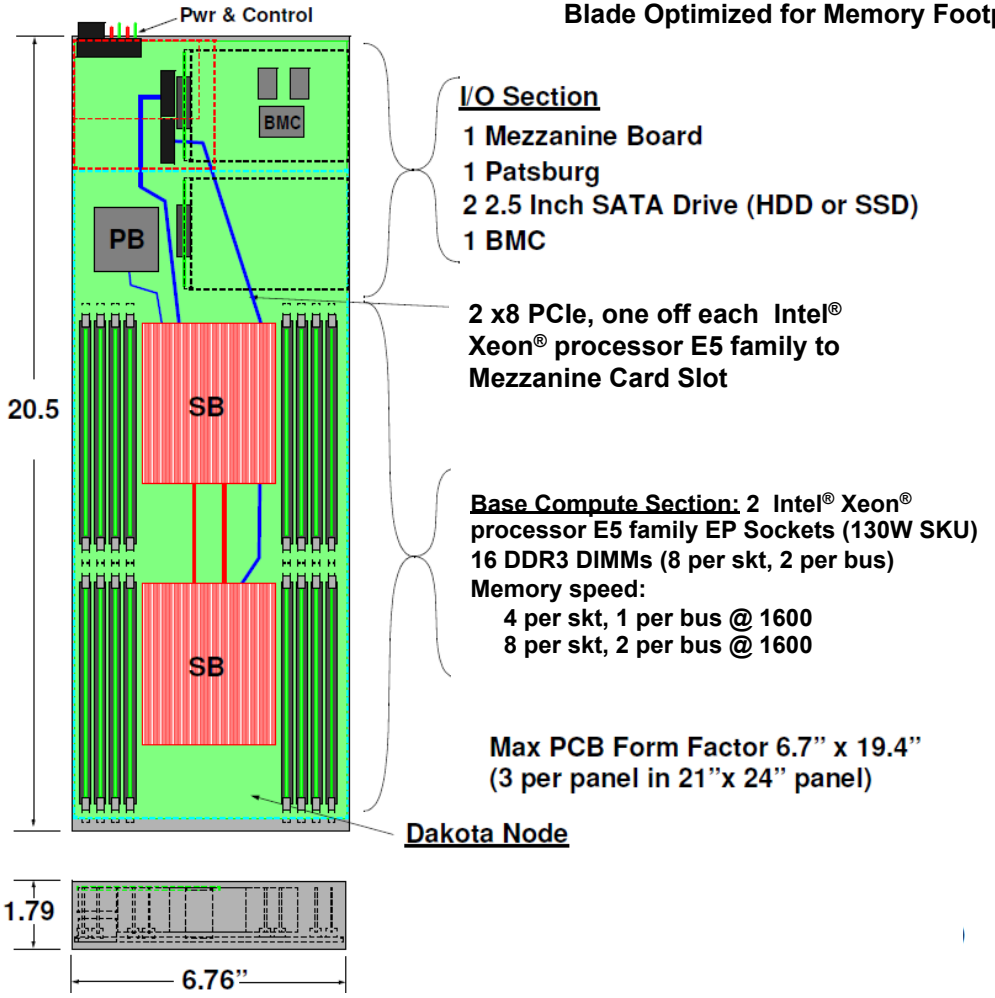

**M-Rack** 72 x 2 = **144N**
(288 Sockets)



# SGI ICE X Compute Blade
## IP-113 (Dakota) for “D-Rack” or “M-Rack”

**Dakota Blade**

**Intel® Xeon® processor E5 family-based 2-Socket Node Blade Optimized for Memory Footprint & Power**




**I/O Section**
1 Mezzanine Board
1 Patsburg
2 2.5 Inch SATA Drive (HDD or SSD)
1 BMC

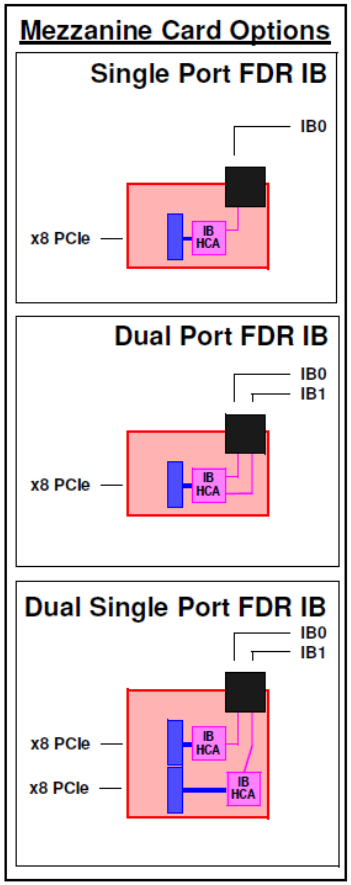
2 x8 PCIe, one off each Intel® Xeon® processor E5 family to Mezzanine Card Slot

**Base Compute Section:** 2 Intel® Xeon® processor E5 family EP Sockets (130W SKU)
16 DDR3 DIMMs (8 per skt, 2 per bus)
Memory speed:
4 per skt, 1 per bus @ 1600
8 per skt, 2 per bus @ 1600

Max PCB Form Factor 6.7” x 19.4” (3 per panel in 21”x 24” panel)

**Dakota Node**

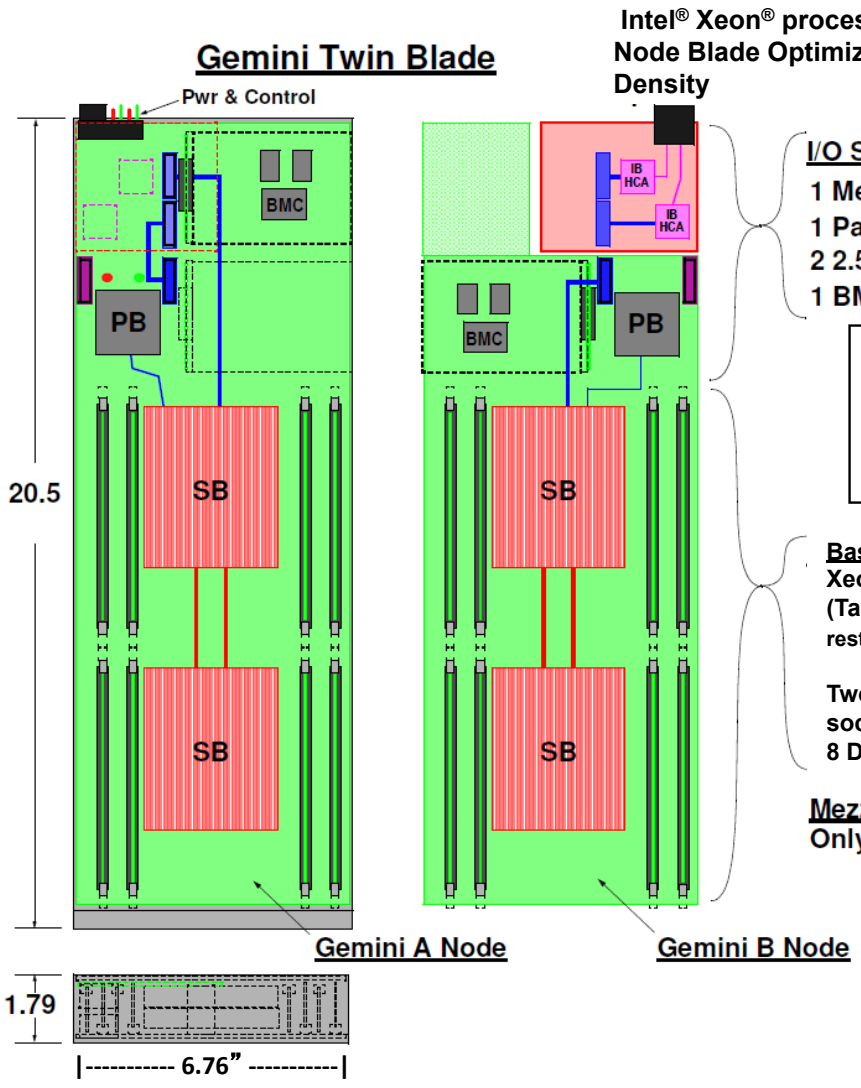
**Mezzanine Card Options**




**Main Features:**

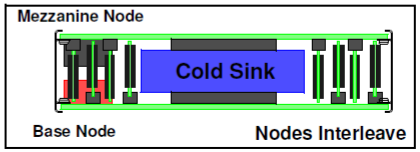
- Supports single or dual plane FDR InfiniBand
- Supports two Intel® Xeon® processor E5 family CPUs
- Supports up to eight DDR3 DIMMs per socket @ 1600 MT/s
- Houses up to two 2.5” SATA drives for local swap/scratch usage
- Utilizes traditional heat sinks

# SGI ICE X Compute Blade
## IP-115 (Gemini Twin) for "M-Rack"

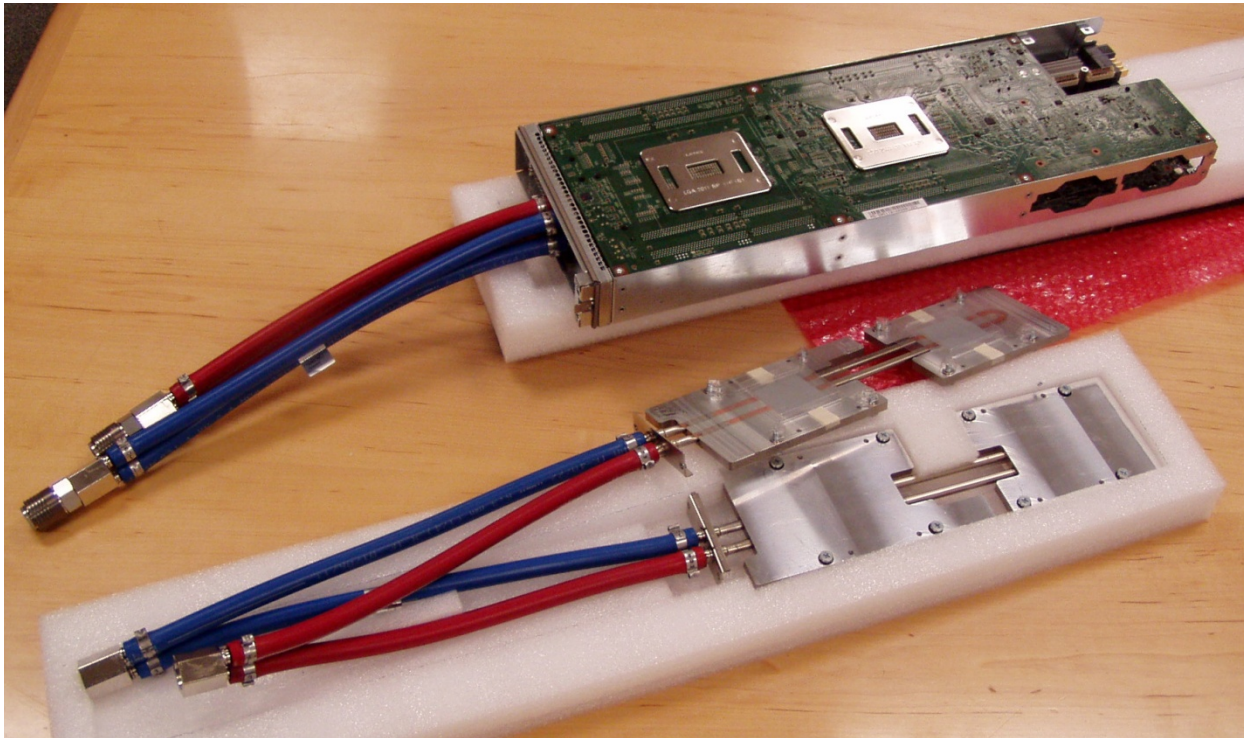


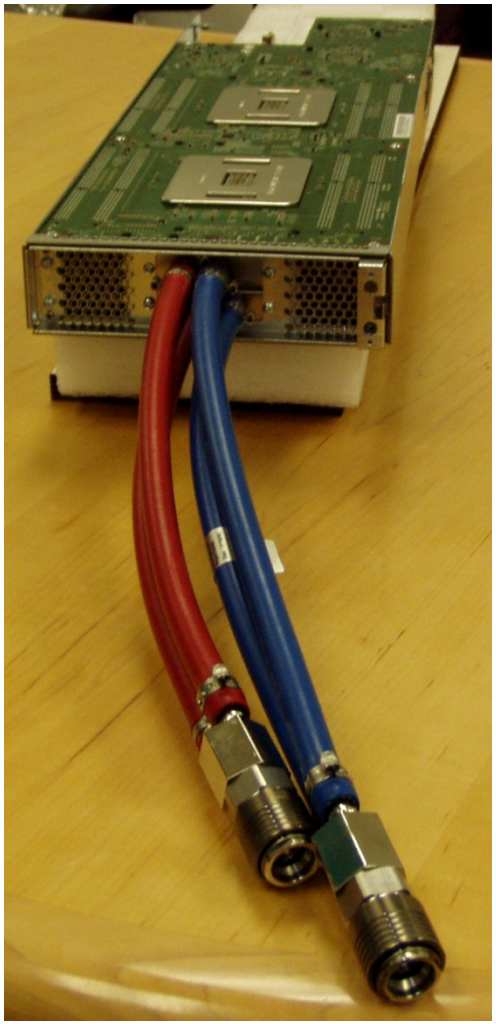

**Main Features:**

- Supports single plane FDR InfiniBand
- Supports four Intel® Xeon® processor E5 family CPUs
  - Two dual socket nodes
- Supports four DDR3 DIMMs per socket @ 1600 MT/s
- Houses up to two 2.5" SATA drives for local swap/scratch usage
  - One per node
- Utilizes traditional heat sinks and cold sinks (liquid)

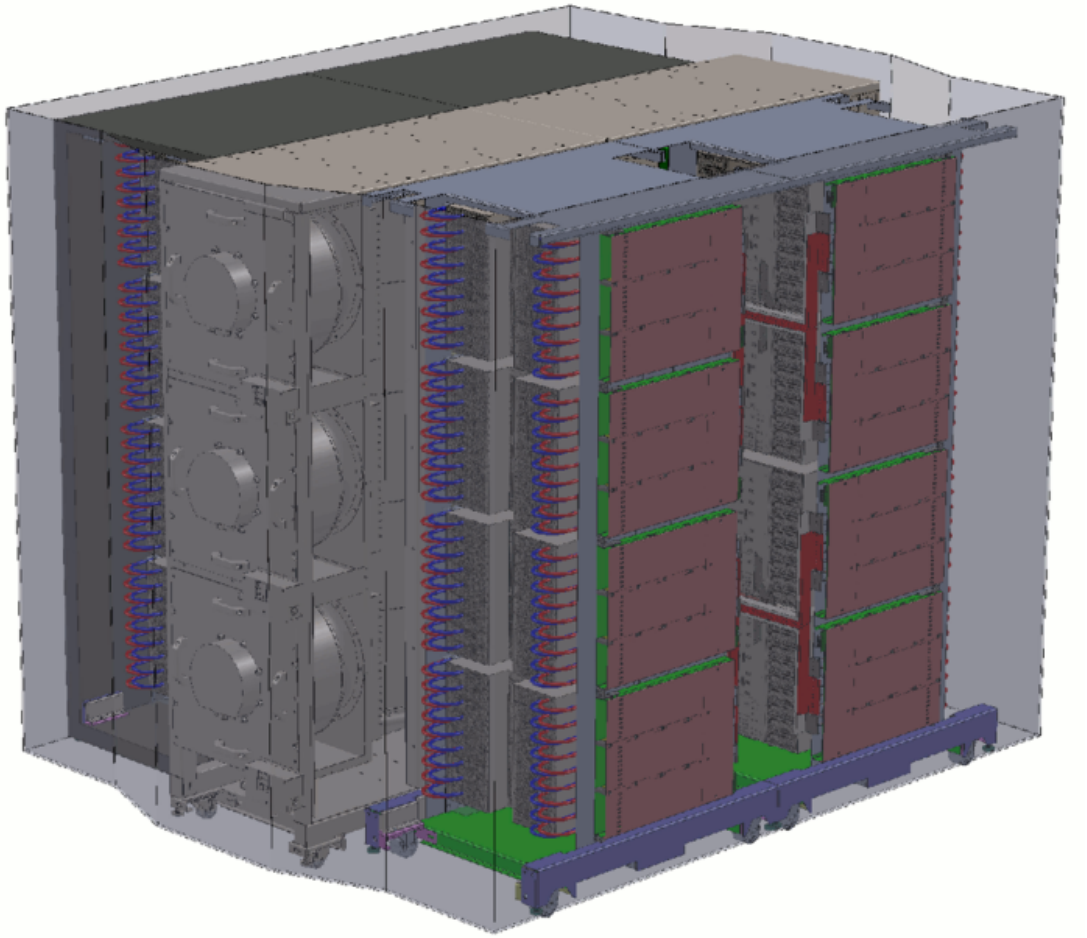


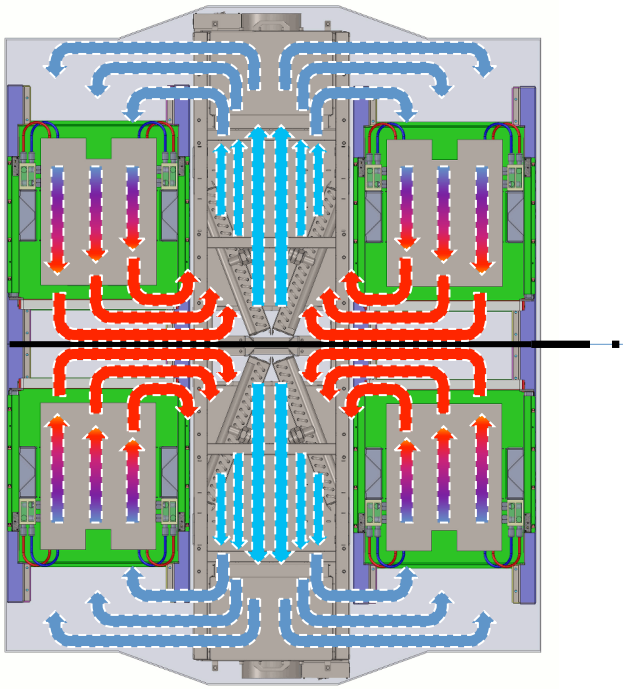
# SGI Cold Sink Technology – IP-115 Gemini Twin



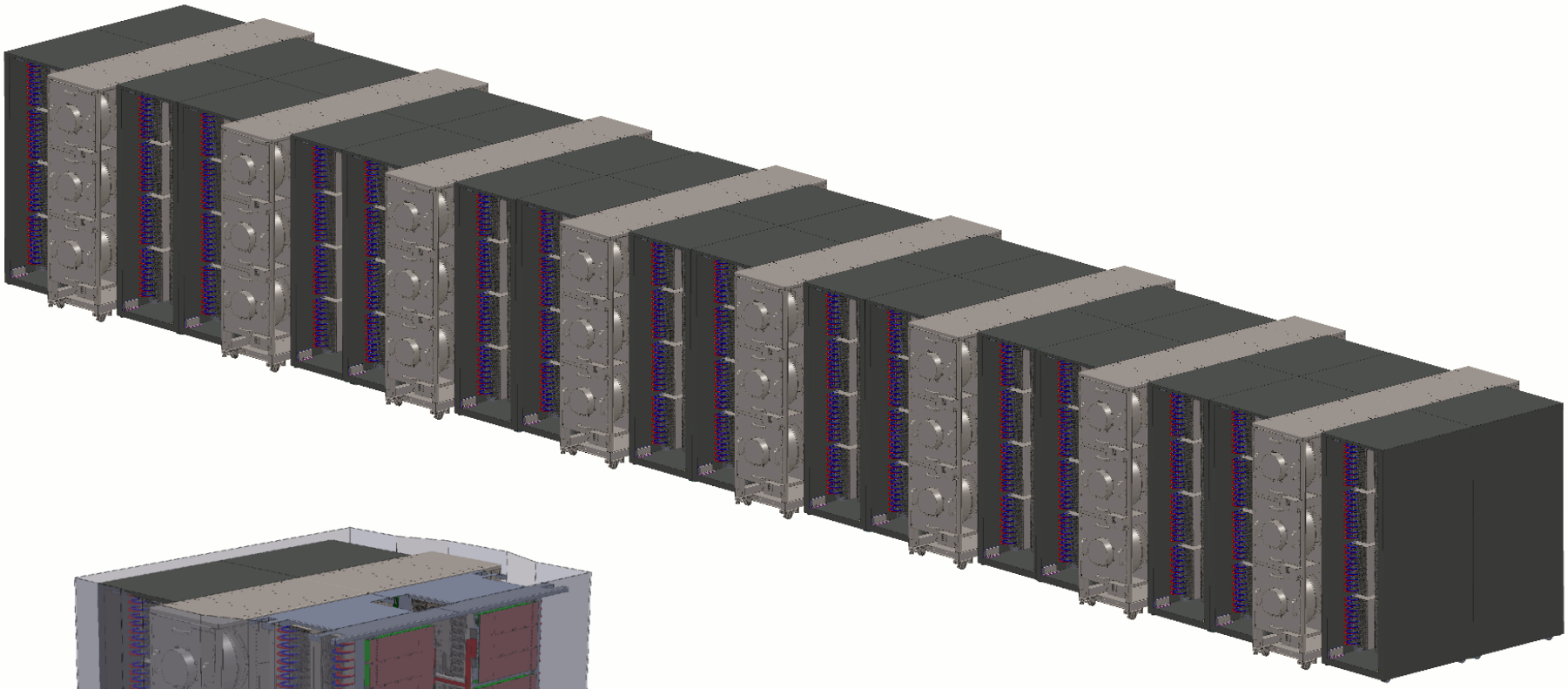


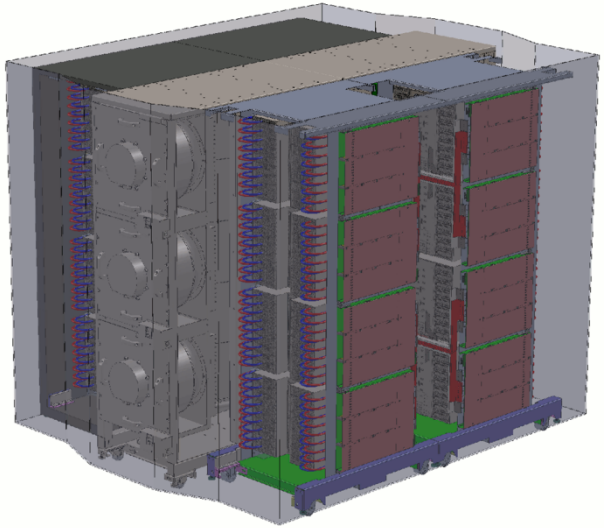
# ICE X M-Rack Cell





# ICE X Row : 8 x Cells



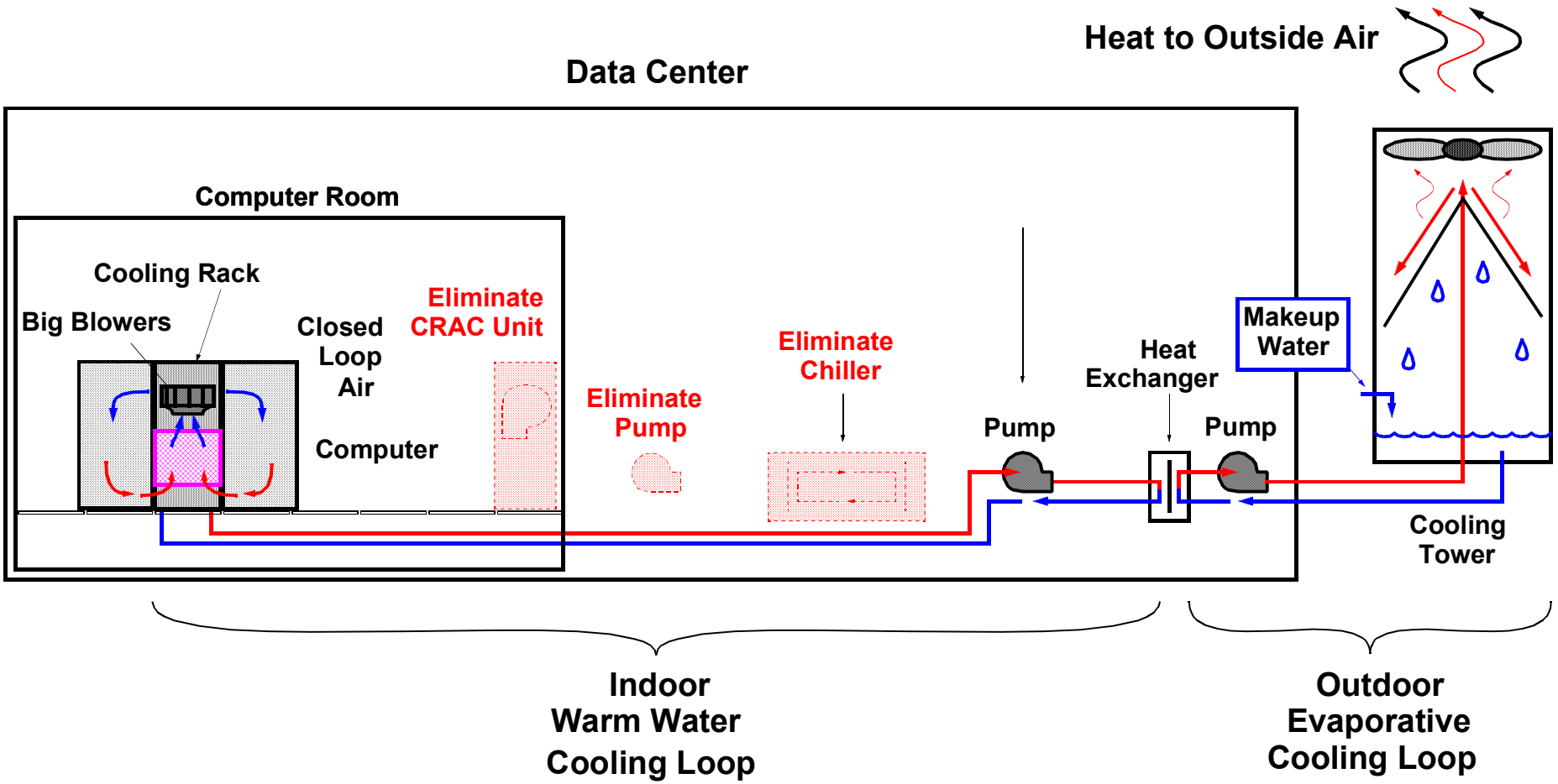


**(2304) IP-115 Gemini Twin**
**Blades (9216) Sockets**
**73K cores with SandyBridge**



sgi

# Data Center Heat Flow- "Cell" Technology



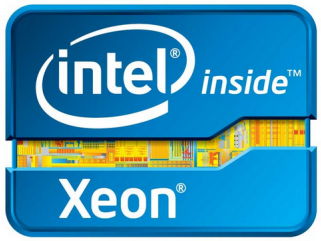
# Introducing the **SGI** UV 2

# SGI® UV™ 2000 : COSMOS Cambridge University

# SGI® UV™ 2000 : 512 sockets, 4096 cores, 64TB

# SGI UV 2

## Open platform -- Intel® Xeon® processor E5-4600 product family and off-the-shelf Linux®.






- Intel® Xeon® Processor E5-4600 product family
- SGI is major Linux contributor
  - Red Hat® Enterprise Linux 6
  - SUSE® Linux® Enterprise Server 11
- Runs Linux SW off-the-shelf, unmodified
- Standard Management, Storage Interfaces
  - SGI Infinite Storage Solutions
  - Interface with common management software schemes





sgi

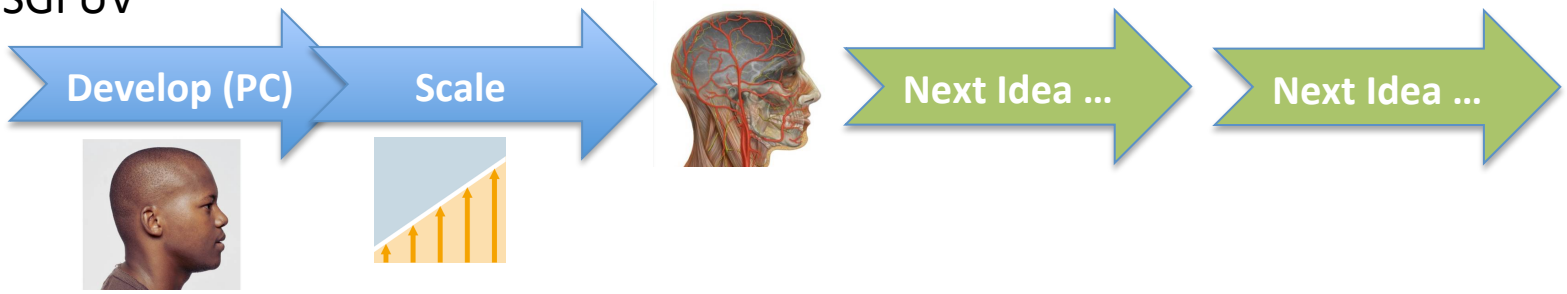
# SGI UV 2

Rapid innovation: Invent on your laptop, scale on SGI UV, no re-write required.

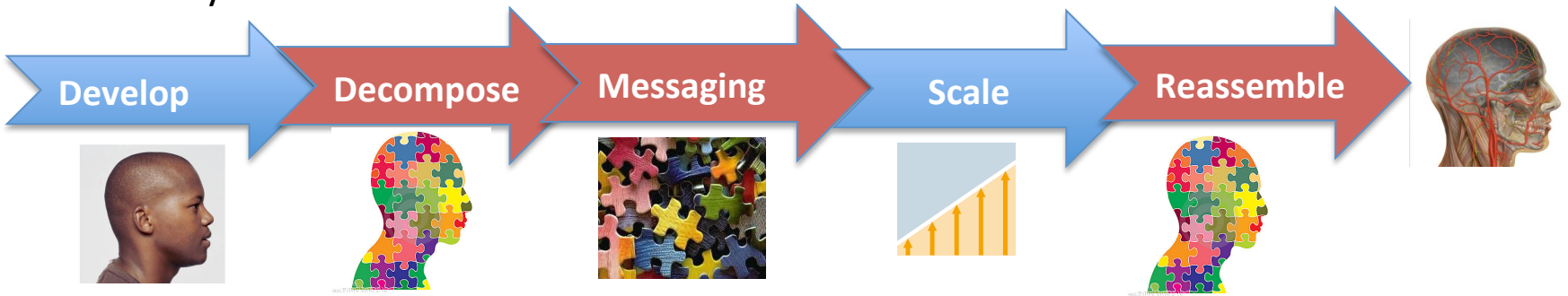
*"…unparalleled ease of use for rapidly testing new ideas … dramatically increasing users' productivity."*

*Pittsburgh Supercomputing Center*

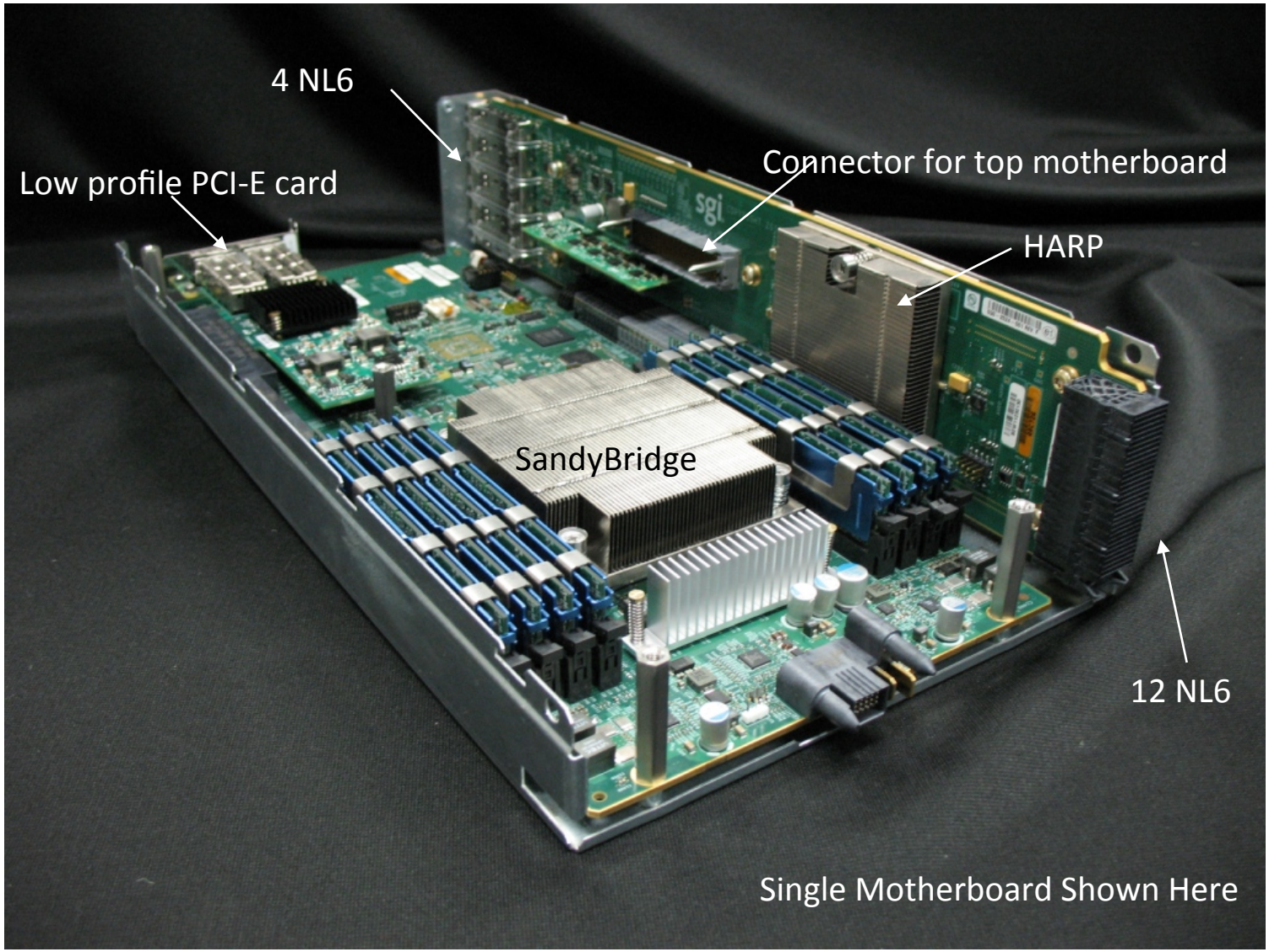
SGI UV



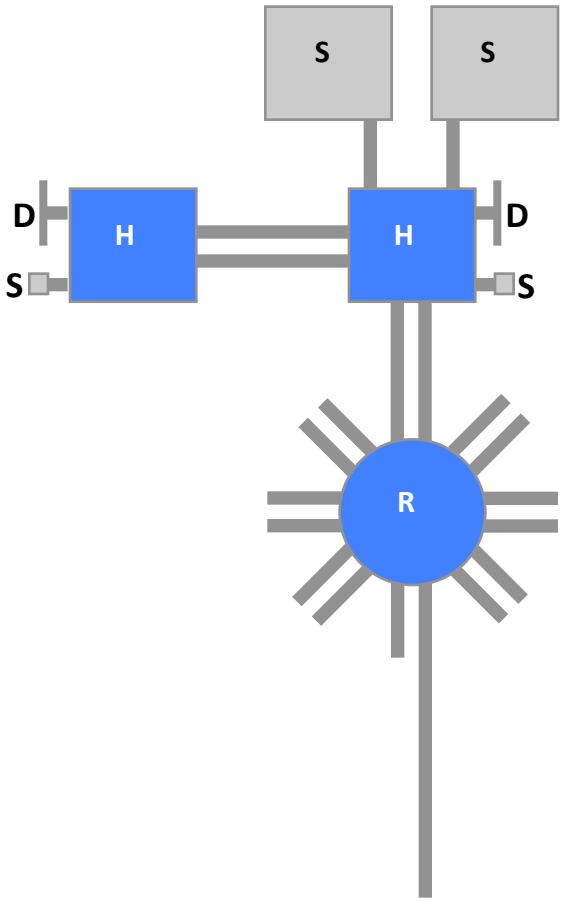

Scale-out Systems



# SGI UV 2000 Blade



# UV1 vs. UV2




**Socket**
- NHM-EX
- WSM-EX

- QPI 1.0

**Glue**
- H + H + R
- 3 separate Chips
- 90nm
- (D) Directory DIMM
- (S) Snoop DRAM

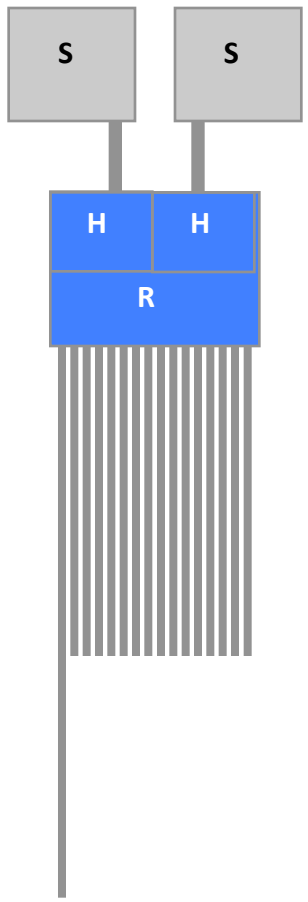
**Interconnect : NL5**
- 6.25 GT/s
- 8B/10B encoding
- 4 x 12 lanes
- Cu only
- 7m max

**Socket**
SNB-EP –
IVB-EP -

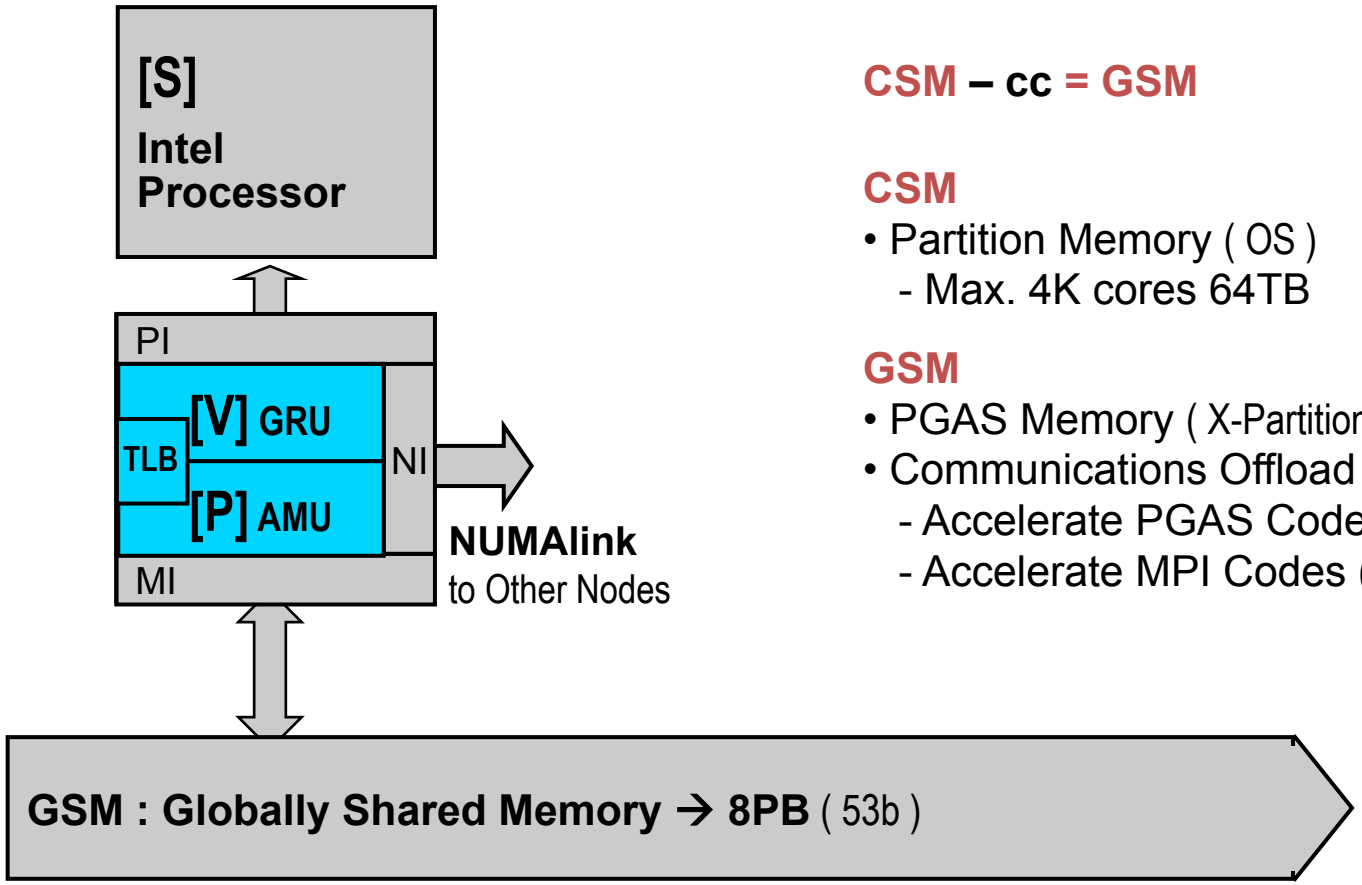
QPI 1.1 -

**Glue**
H + H +R –
into 1 Chip –
40 nm –
No Directory DIMM –
No Snoop DRAM –
Better AMOs -

**NL6 : Interconnect**
Higher Payload –
16 x 4 lanes –
Cu & Optical –
20m max –



# UV Foundation : Coherent Shared Memory + Communications Offload




**CSM – cc = GSM**

**CSM**

- Partition Memory ( OS )
  - Max. 4K cores 64TB

**GSM**

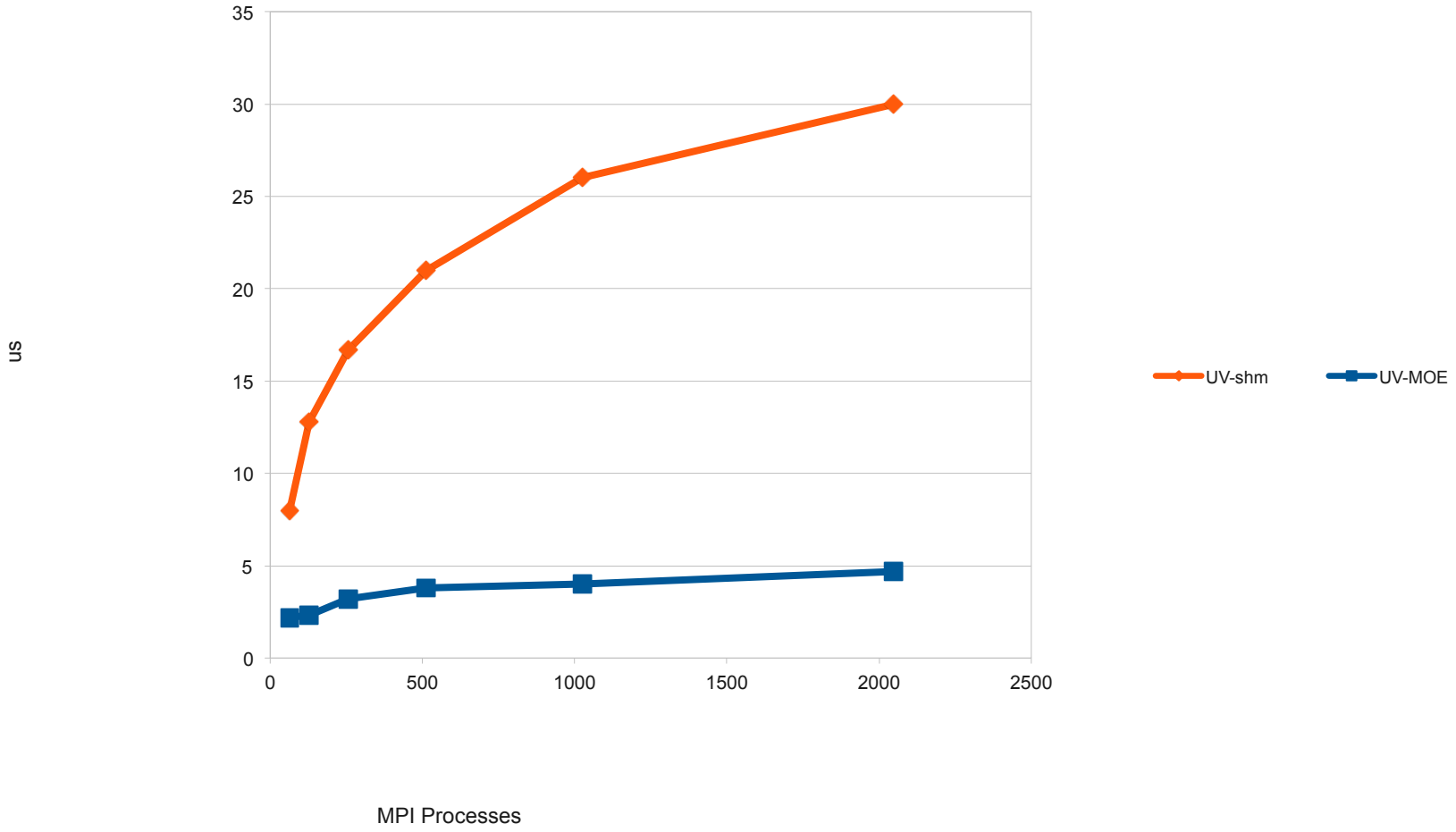
- PGAS Memory ( X-Partition )
- Communications Offload ( GRU + AMU )
  - Accelerate PGAS Codes
  - Accelerate MPI Codes ( MOE v.v. TOE )

# GRU Programming

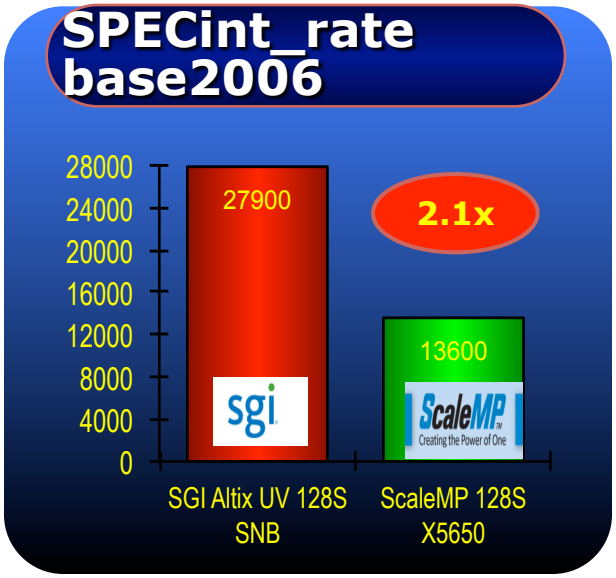
```
#include <string.h>
#include <uv/gru/gru.h>
#include <uv/gru/gru_instructions.h>
#include <uv/gru/gru_alloc.h>
void *egru_memcpy(void *dst, const void *src, size_tnbytes)
{
        int tri, lines;
        gru_alloc_thdata_t thd;
        if (gru_temp_reserve_try(&thd) != 0)
                return memcpy(dst, src, nbytes);
        lines = thd.dseg_size / GRU_CACHE_LINE_BYTES;
        lines &= ~1;
        tri = gru_get_tri (thd.dsegp);
        gru_bcopy(thd.cbp, src, dst, tri, XTYPE_B, nbytes, lines, 0);
        gru_wait_abort(thd.cbp);
        gru_temp_release();
        return dst;
}
```

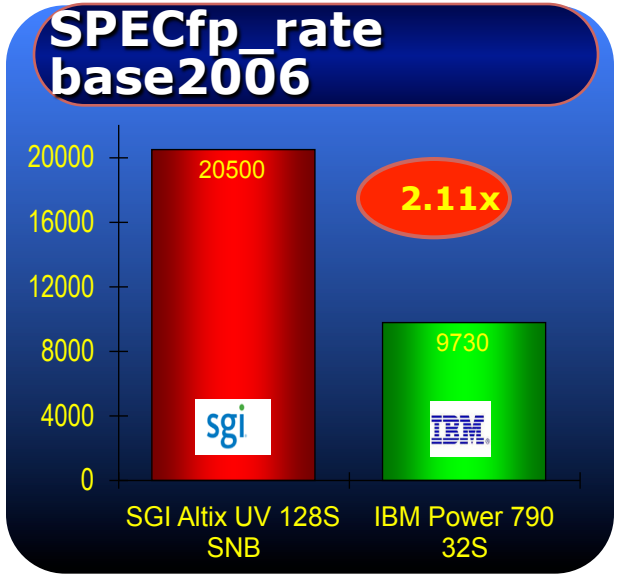
| | |
|---|---|
| Memory and Message Operations | gru_bcopy,gru_bstore, gru_ivload, gru_ivset, gru_ivstore, gru_vload, gru_vset, gru_vstore gru_gamir, gru_gamirr, gru_gamer, gru_gamerr gru_send_message |
| Completion Operations | gru_wait, gru_wait_abort gru_check_status |

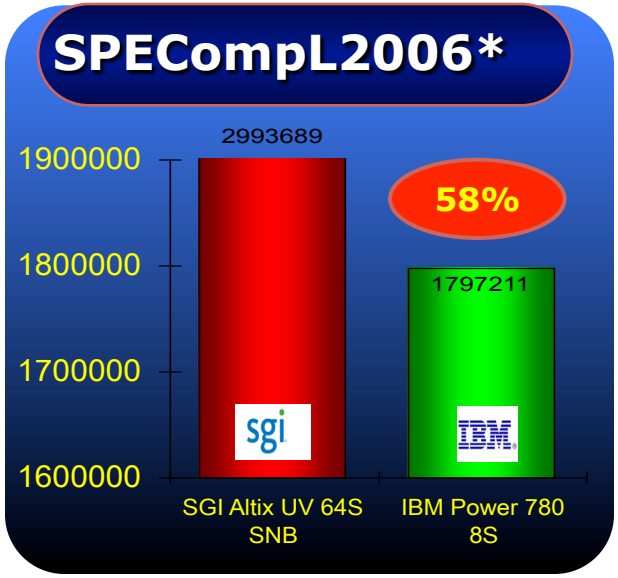
# UV MPI Barrier

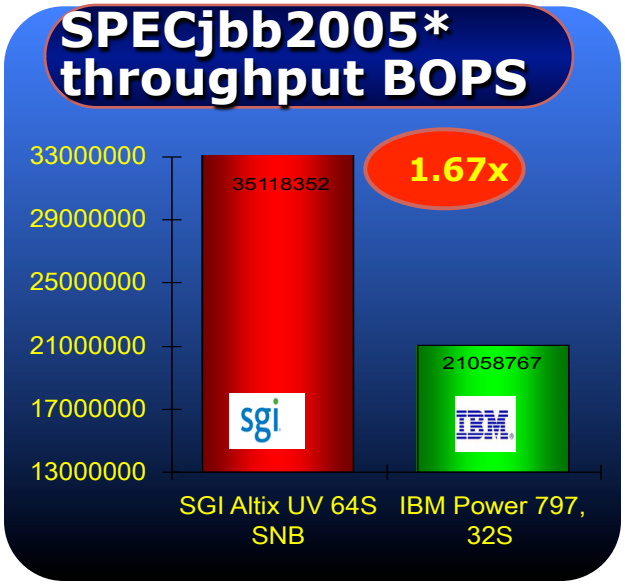


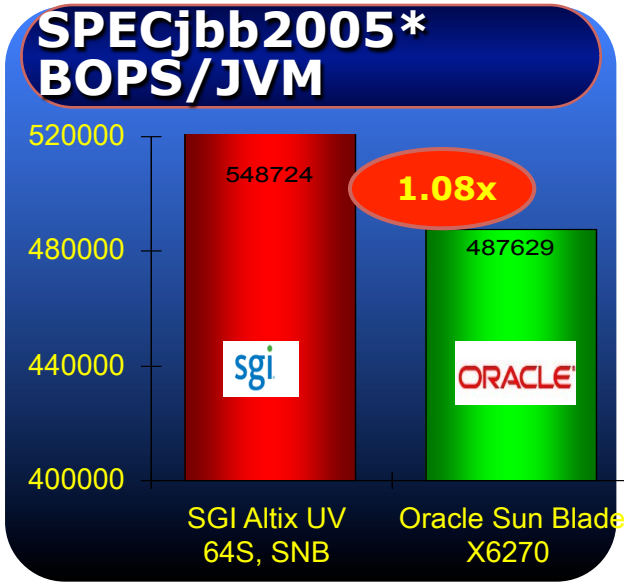
# New World Records for Altix UV2000 SNB

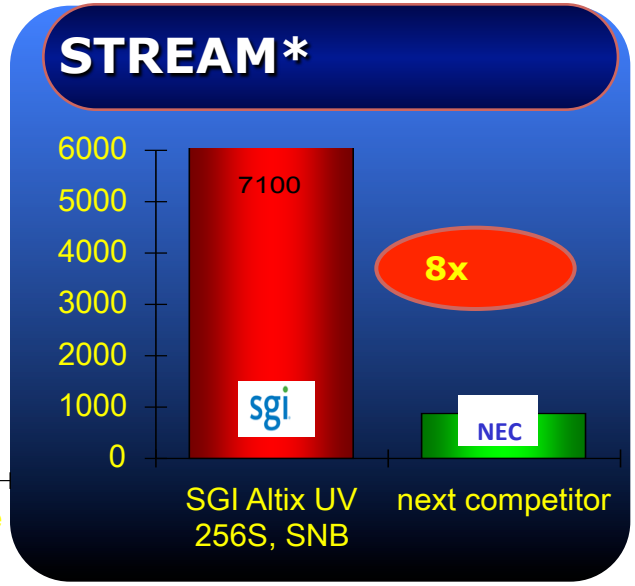


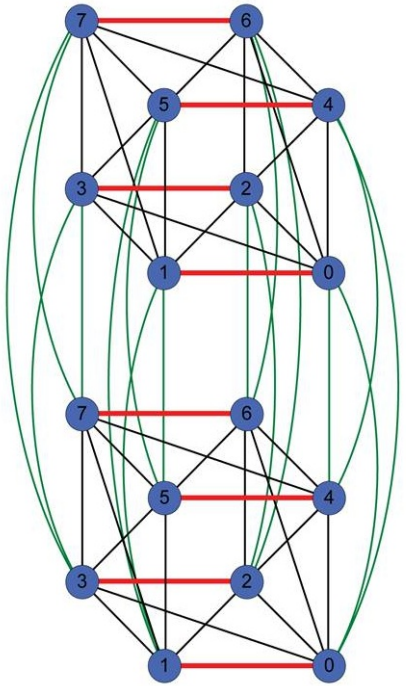



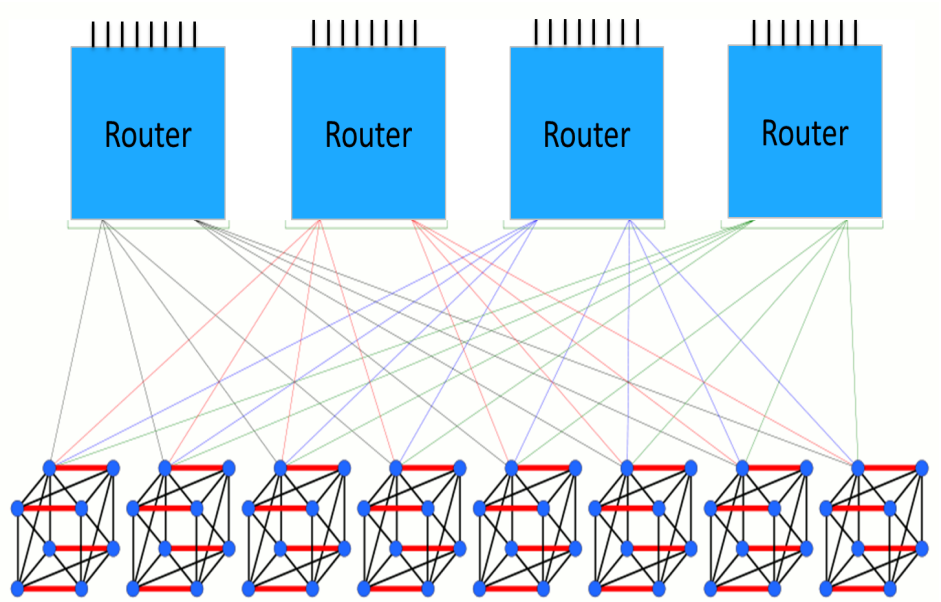



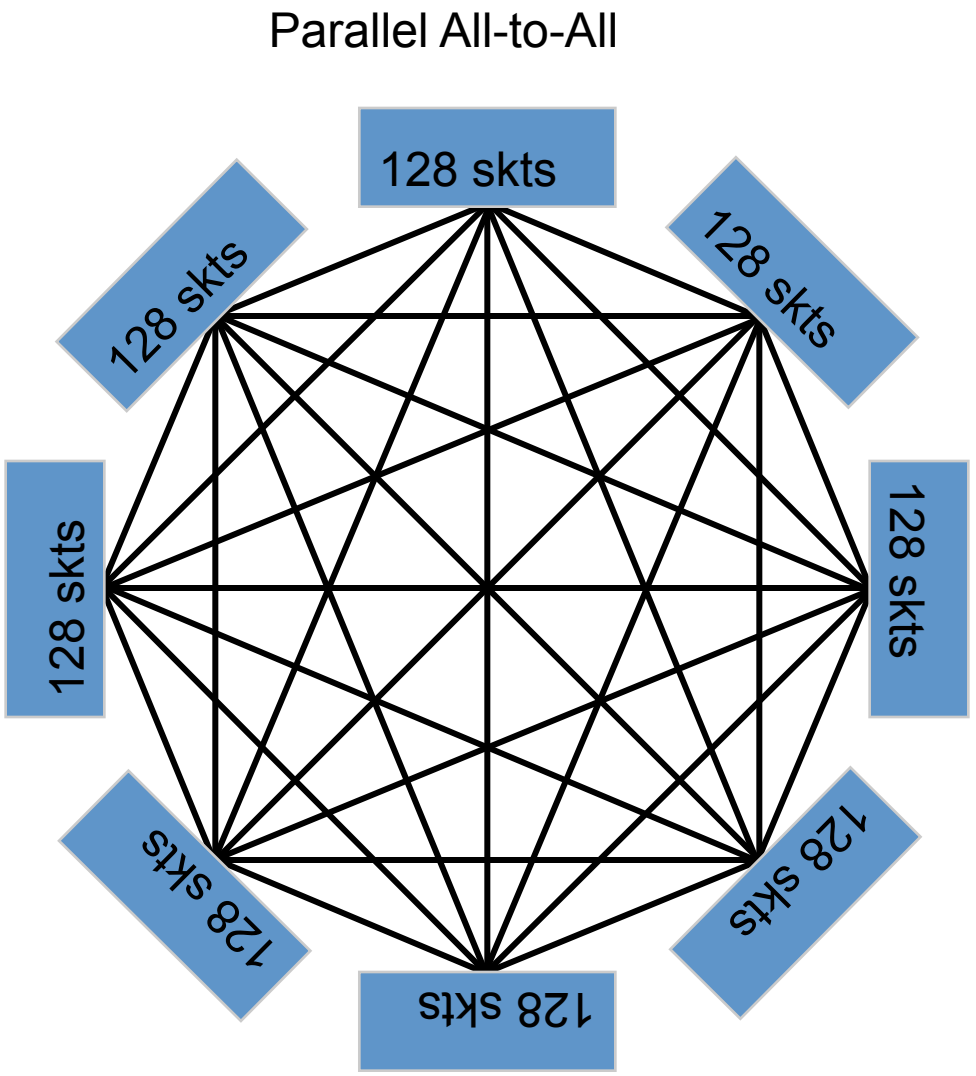



# SGI® NUMAlink™ 6 Topology

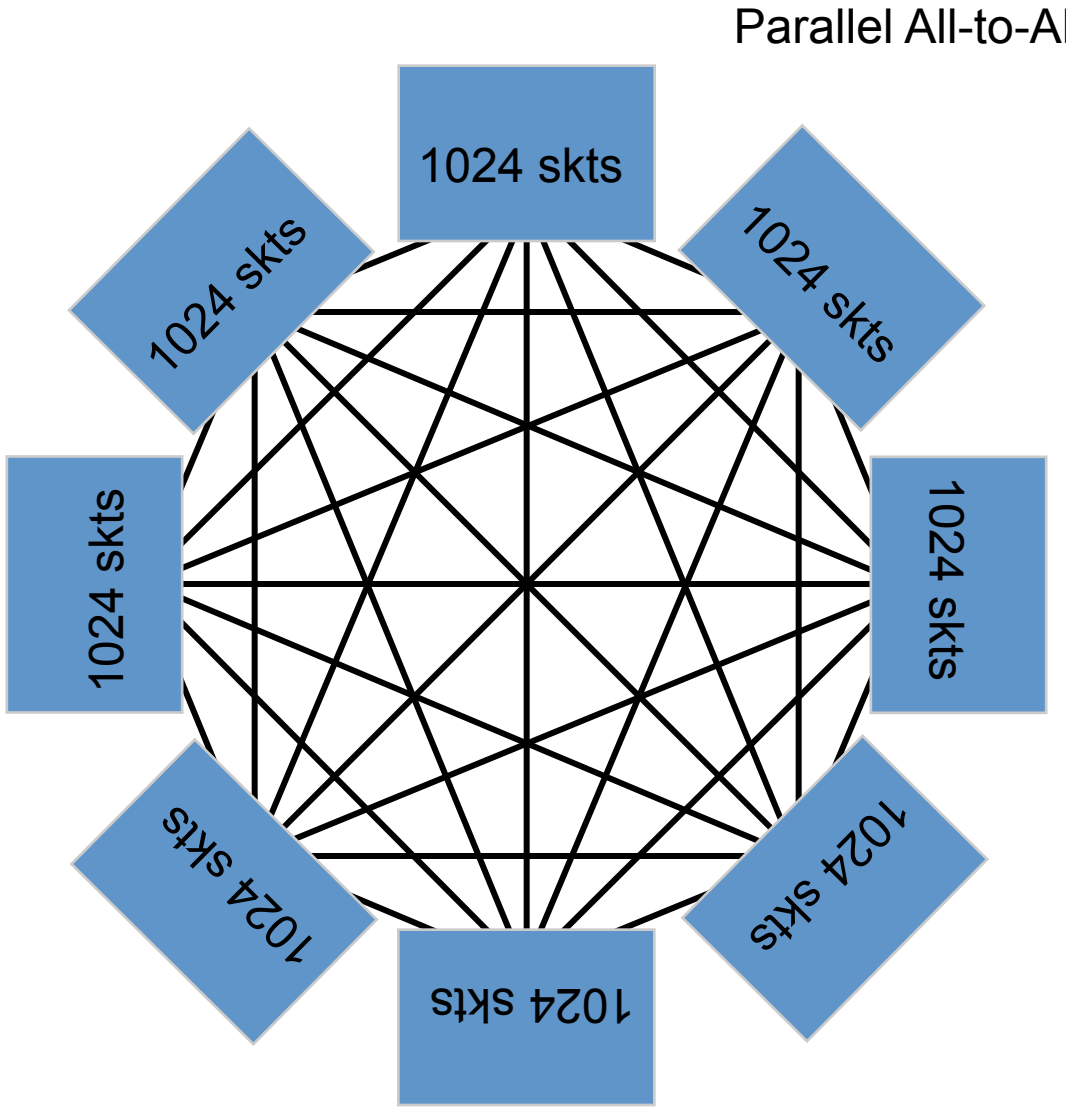


# UV2: 1024-Socket Topology (16 Racks)



# UV2: 8096-Socket Topology



# UV 2 Feature Advances

| Feature | UV1/NUMAlink 5 | UV2/NUMAlink 6 |
|---|---|---|
| System SSI scale | 2560c/4096t | 4096c/4096t |
| Memory/SSI | 16TB | 64TB |
| Local Read Latency | 130 ns | 80 ns |
| Full SSI Read Latency | <1 us | <1 us |
| NL MPI/PGAS Scale | 32K sockets | 16K sockets |
| Processor | Nehalem EX | Sandybridge EP4S |
| Sockets/rack | 64 (large 24”) | 64 (19”) |
| Memory/rack | 8TB | 16TB |
| 64-bit update rate | .017 GUPS/Socket | .100 GUPS/Socket |
| IO | PCIe Gen 2 | PCIe Gen 3 w/ 2X remote, unbuffered IO performance |
| Flops/rack | 6Tflop | 11Tflop |