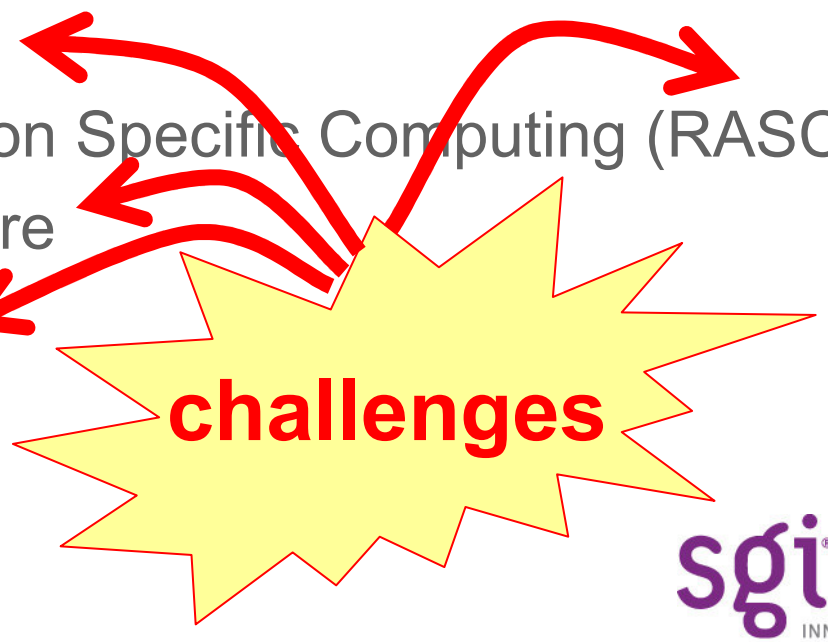
Challenges of Sustained Petascale Computing
SOS11 – Key West – June 12, 2007

Jean-Pierre Panziera
Chief engineer, Applications group

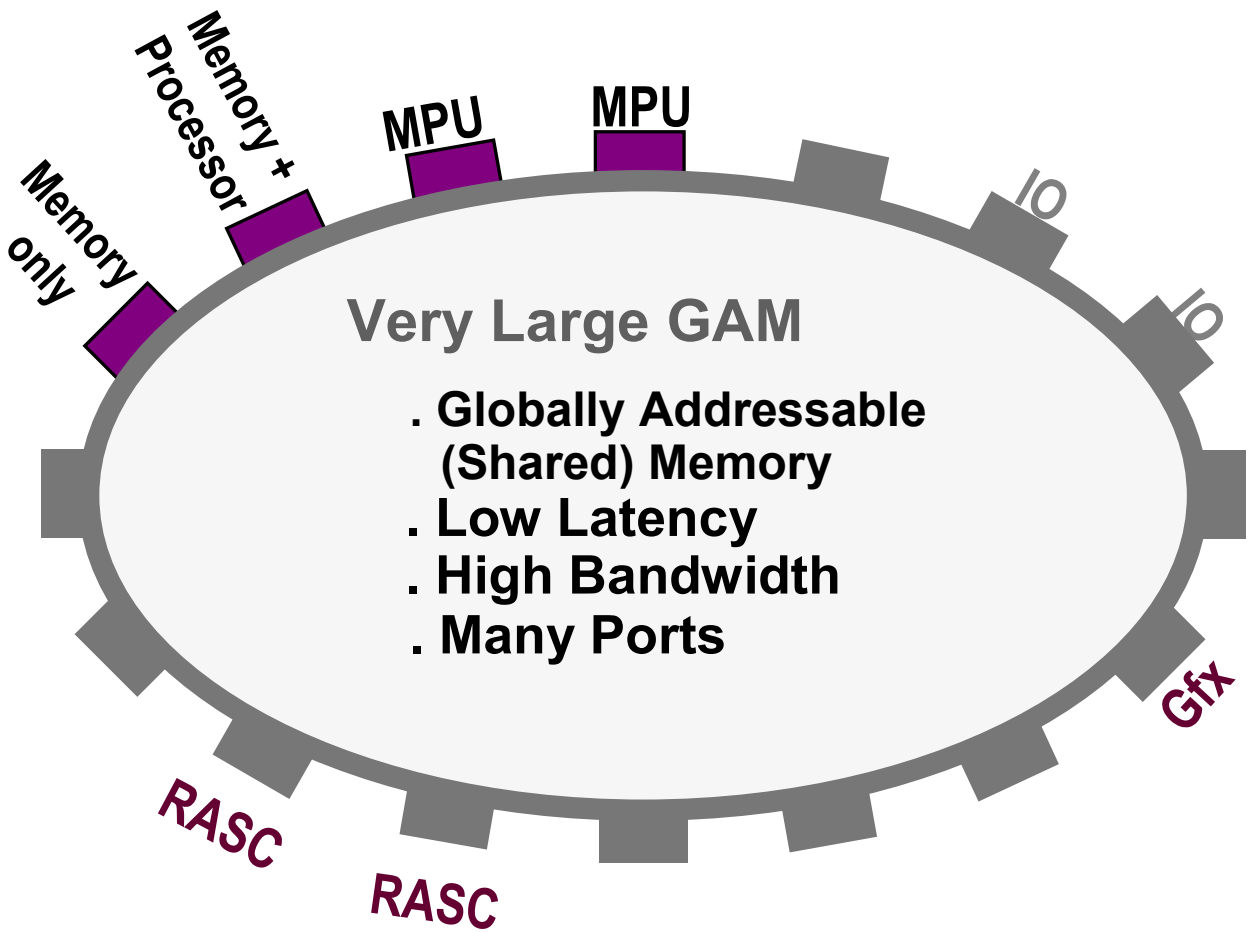
Challenges of Petascale Computing

- the Globally Addressable Memory (GAM) concept
- NUMAflex architecture evolution
- a current configuration: LRZ
- Scaling HW to Petascale
- Reconfigurable Application Specific Computing (RASC)
- Petascale system software
- Petascale applications



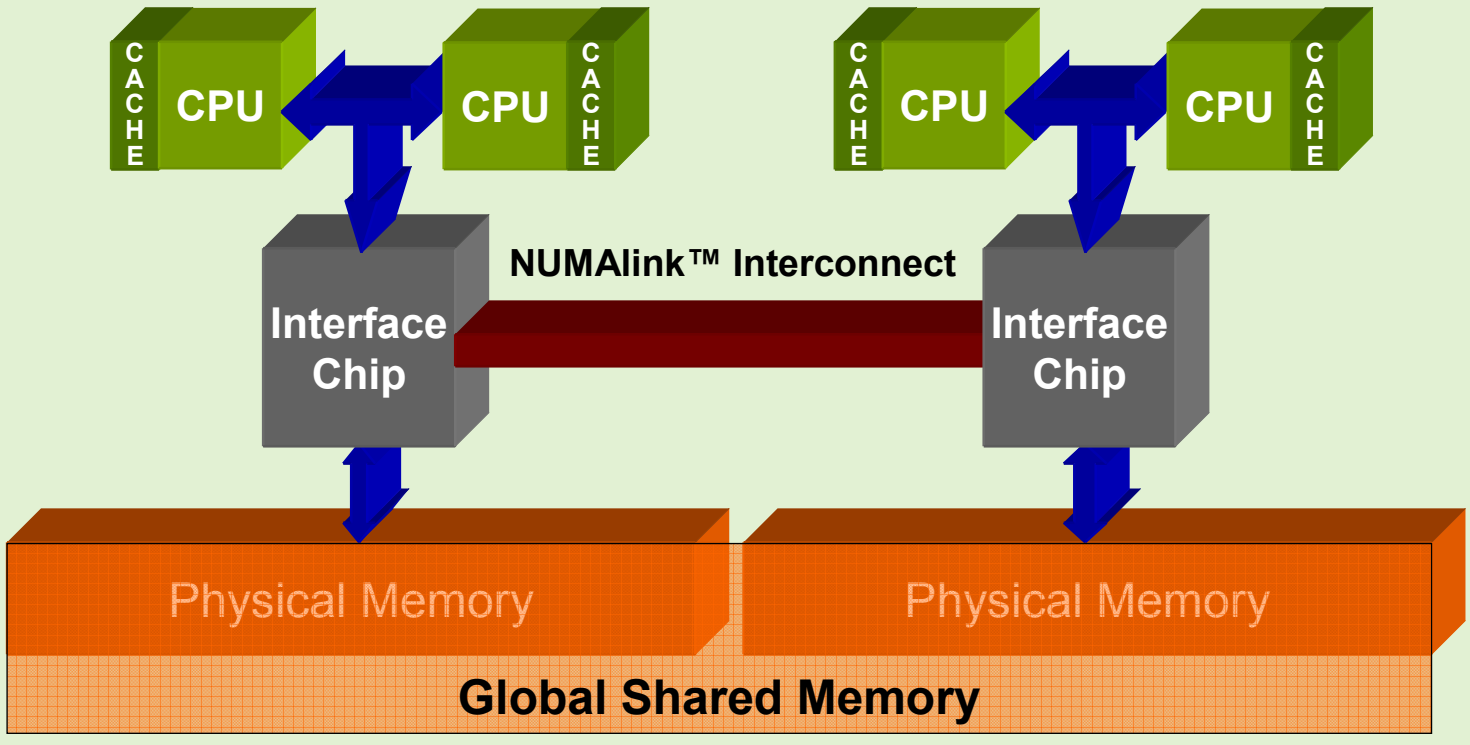
sgi INNOVATION FOR RESULTS

Globally Addressable Memory (GAM) Data Centric Architecture



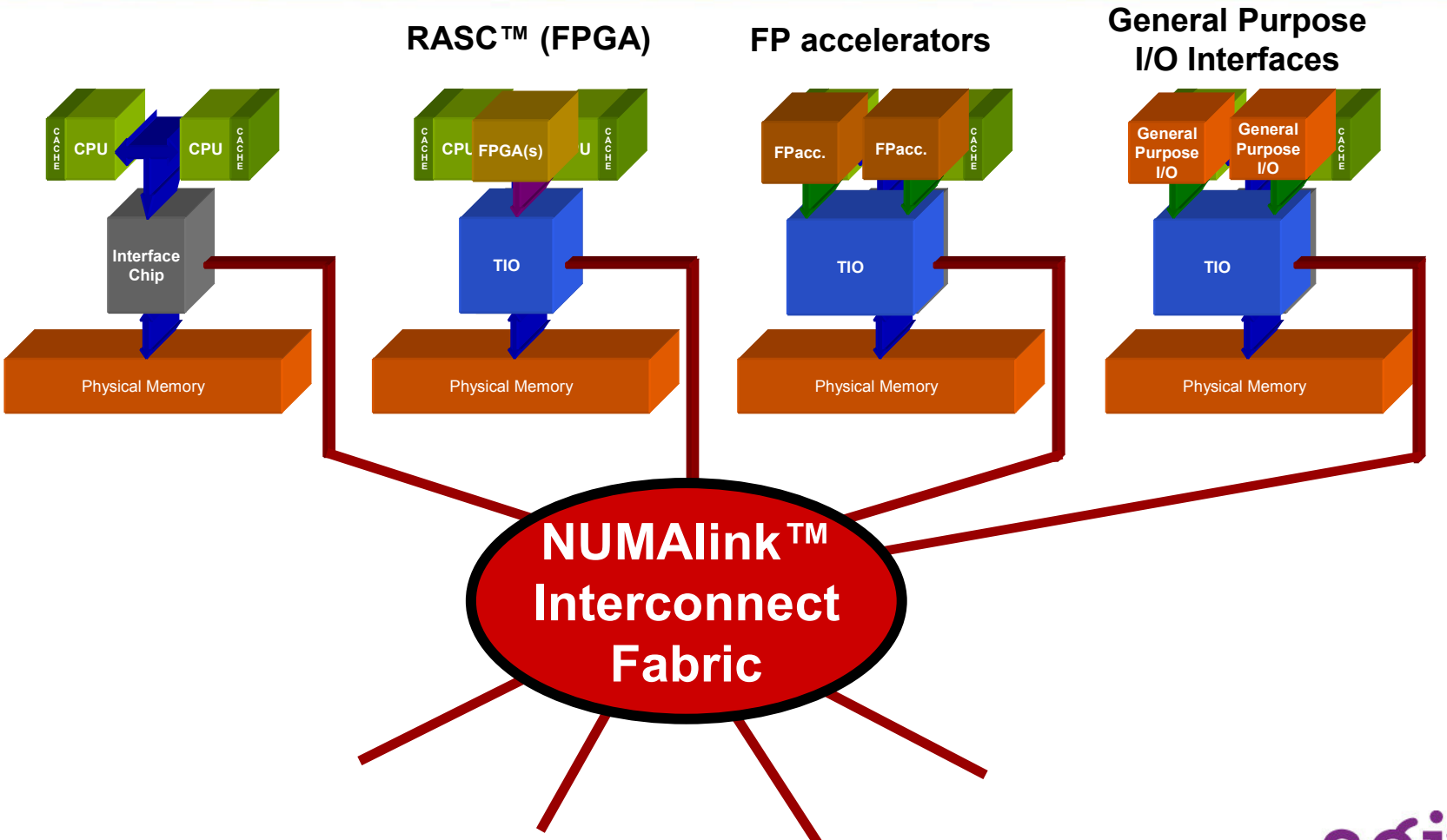
sgi INNOVATION FOR RESULTS

ALTIX scalable system infrastructure



sgi INNOVATION FOR RESULTS

ALTIX scalable system infrastructure
a global shared memory space + IO and accelerator nodes



sgi INNOVATION FOR RESULTS

NUMAflex architecture evolution

year	system	max SSI procs/cores	max GAM procs/cores	Memory capacity
1996	Origin2000	256 (512)	-	512 GB
2000	Origin3000	512 (1024)	-	1 TB
2003	Altix3000	512	2 048	32 TB
2006	Altix4000	4 096 (1024)	16 384	196 TB
-	UV	petascale	petascale	petascale

MIPS

IA64

sgi INNOVATION FOR RESULTS

LRZ_2

128 racks, 4096 nodes, 9728 cores = 19 partitions x 512
38 TB memory, 4GB/core
300 TB disks - CXFS

2 816 routers (2048 in backplane + 766 routers)
14 082 links (9216 in backplane + 4866 cables)

62 TFlops (peak)
43 TB/s aggregate local memory bandwidth
409 GB/s bisection bandwidth

petascale? ➔ x20 !!! (current generation)
x2- 4 with next generation

May 2007

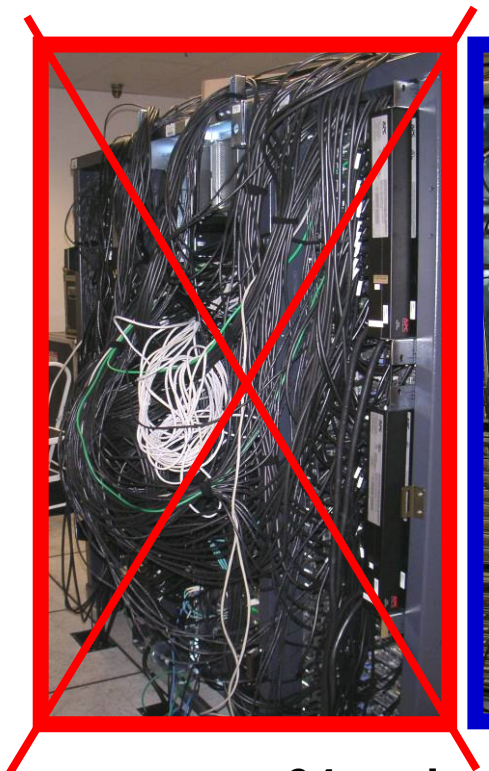
sgi INNOVATION FOR RESULTS

Hardware Challenges of Sustained Petascale Computing

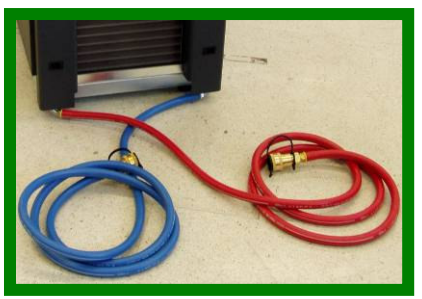
- Scale to more nodes/cores/memory → NUMAflex architecture
- Scale memory bandwidth → Bytes/Flop ratio
- Scale interconnection bandwidth → NUMAlink enhancements

- System packaging, denser, tighter, easier to deploy
- Power management, efficient (water) Cooling,
- Integrate Accelerators (FPGA, FP accelerators, …)
- Resilient systems, improved RAS features

Packaging:
rack optimized cabling and water cooling



64 node cluster

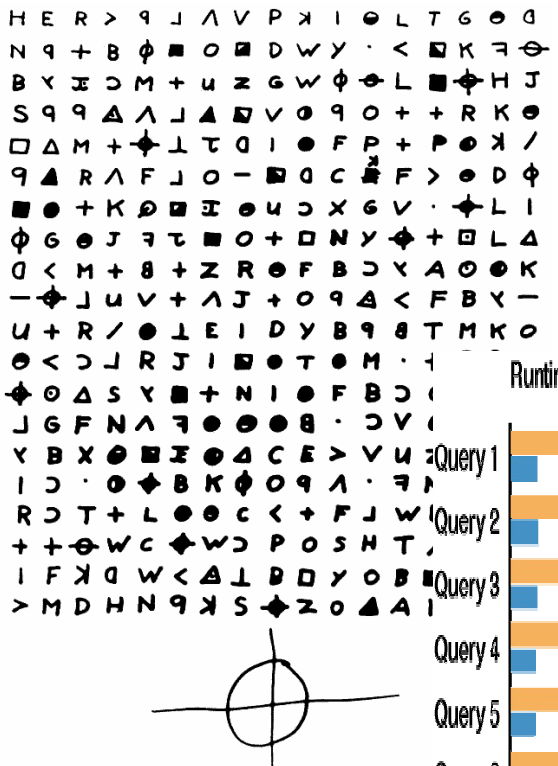


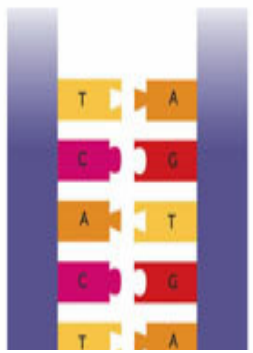
Water cooling

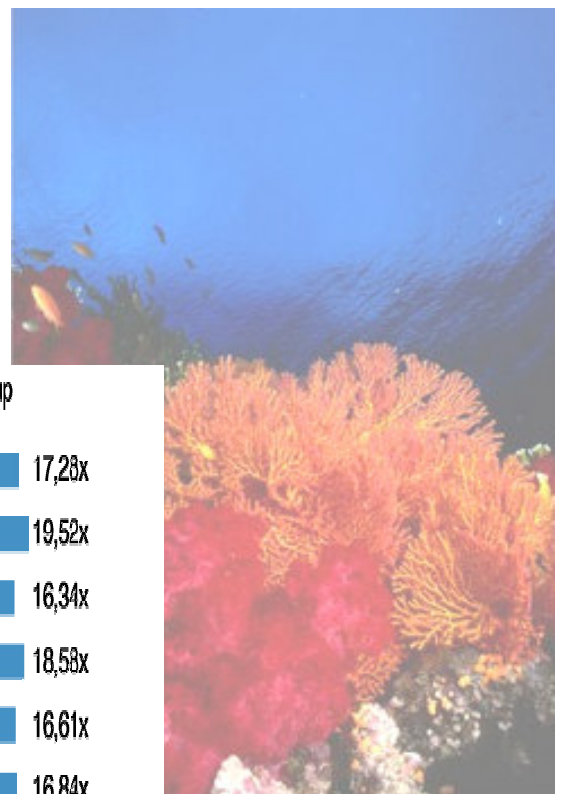
Heat Rejection
95% water / 05% air

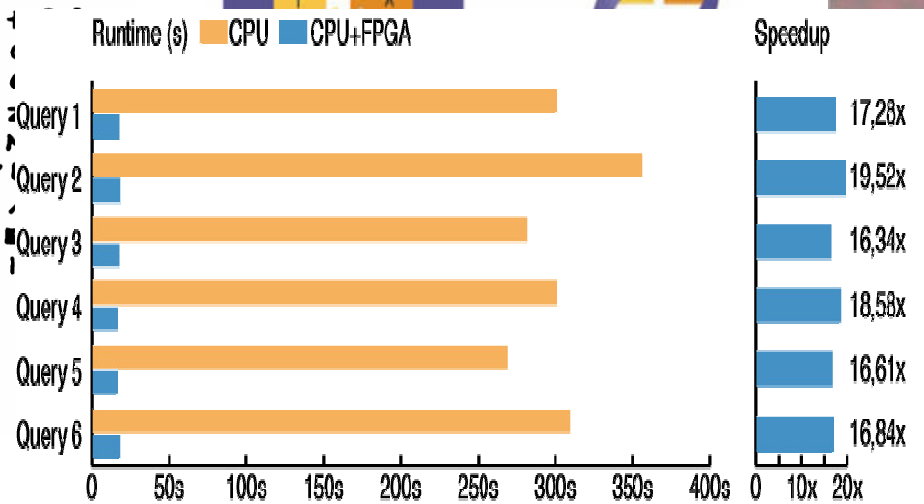
sgi INNOVATION FOR RESULTS

RASC/FPGA accelerators
encryption, BLASTn, jpeg, …







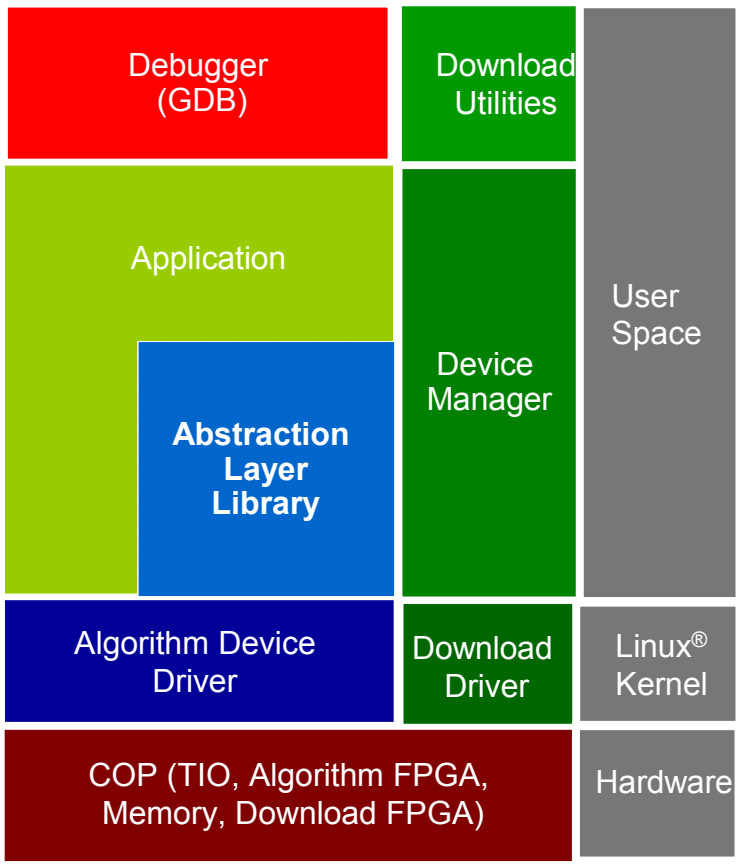


BLASTn: RASC RC100 FPGA acceleration

sgi INNOVATION FOR RESULTS

SGI® RASC™ Integrated Solution Stack Simplifies Development

Tool	Function
FPGA Aware Version of the Gnu Debugger (GDB)	Simultaneously debugs both the CPU based app and the FPGA accelerated app
RASC Abstraction Layer (RASCAL)	Enables serial or parallel FPGA scaling
RASC API and Core Services Library	Provides tools to develop reconfigurable computing elements in a multi-user, multiprocessing environment
3rd Party HLL Development Tool Support	Fully integrated tools including Celoxica Handel-C and DK Design Suite, Mitrionics Mitrion C, Impulse Impulse-C (in process)



sgi INNOVATION FOR RESULTS

RASC challenges
FPGAs, FP accelerators, …

- No standard HW interface
- No standard SW interface (IO library)
- Managing Accelerator private memory
- Low level programming (VHDL/ assembler language)
- No standard high level compiler
- Limited code size (kernel)
- FP format, FP double precision
- Data Error Protection

Software Challenges of Sustained Petascale Computing

- Scale to more nodes/cores/memory → NUMAflex architecture
- Scale memory bandwidth → Bytes/Flop ratio
- Scale interconnection bandwidth → NUMAlink enhancements

- Integrate all types of nodes: CPU+Memory, Memory-only, IO, RASC, gfx, …within an SSI partition
- Integrate nodes/services across partitions
- HW Optimized Middleware
- Resilient systems, improved RAS features
- Peta-scaling applications

System software environment

- Linux SLES10 & RHELv5 out of the box
  - Linux scales up to 1024p (SLES10) … “constellation” configurations (large partitions)
  - XFS file system
- SGI ProPack 5
  - Linux enhancements (numatools, FFIO, cpuset, XVM)
  - Cross-partitions cluster-wide services (PCP, Array services, MPI, shmem)
- CXFS Clustered file system
- DMF, TMF, openVault

RAS features

- Extensive components testing (DIMMs screening)
- Full System staging
- Error Prevention: memory protection, redundant data and IO paths, power supplies, cooling fans
- Error Detection: enhanced monitoring, system notification
- Error Containment: proc isolation, at worst within 1 partition only
- Maximizing serviceability: hot swappable components
- Fast checkpoint/restart at application level
- Automatic system reconfiguration

Peta-scale Application Challenges

- Requiring new levels of parallelism 1,000s → 100,000s threads (MPI?)
- Taking advantage of multi-cores CPUs (SMP+MPI?)
- Explicit Parallel languages: UPC, co-array Fortran ?
- Integrating FPGA / FP accelerators (compilers?)
- Peta-Debugging
- Peta-Performance tools
- Minimal development effort

Conclusion

- Sustained peta-scale computing presents formidable challenges for HW , system SW and Applications.
- SGI NUMAflex architecture provides a base for peta-scale systems
- SGI next generation UV systems will be peta-scale
- UV software environment will enable sustained peta-scale computing Applications

sgi®

INNOVATION
FOR RESULTS™

sgi®
INNOVATION
FOR RESULTS™