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
Globally Addressable Memory (GAM)

Data Centric Architecture

Very Large GAM
- Globally Addressable (Shared) Memory
- Low Latency
- High Bandwidth
- Many Ports
ALTIX scalable system infrastructure

- CPU
- Cache
- Interface Chip
- Physical Memory
- Global Shared Memory
- NUMAlink™ Interconnect
ALTIX scalable system infrastructure
a global shared memory space + IO and accelerator nodes

NUMAlink™ Interconnect Fabric

RASC™ (FPGA)  FP accelerators  General Purpose I/O Interfaces
### NUMAflex architecture evolution

<table>
<thead>
<tr>
<th>year</th>
<th>system</th>
<th>max SSI procs/cores</th>
<th>max GAM procs/cores</th>
<th>Memory capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>Origin2000</td>
<td>256 (512)</td>
<td>-</td>
<td>512 GB</td>
</tr>
<tr>
<td>2000</td>
<td>Origin3000</td>
<td>512 (1024)</td>
<td>-</td>
<td>1 TB</td>
</tr>
<tr>
<td>2003</td>
<td>Altix3000</td>
<td>512</td>
<td>2 048</td>
<td>32 TB</td>
</tr>
<tr>
<td>2006</td>
<td>Altix4000</td>
<td>4 096 (1024)</td>
<td>16 384</td>
<td>196 TB</td>
</tr>
<tr>
<td>-</td>
<td>UV</td>
<td>petascale</td>
<td>petascale</td>
<td>petascale</td>
</tr>
</tbody>
</table>
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
Hardware Challenges of Sustained Petascale Computing

• Scale to more nodes/cores/memory $\rightarrow$ NUMAflex architecture
• Scale memory bandwidth $\rightarrow$ Bytes/Flop ratio
• Scale interconnection bandwidth $\rightarrow$ 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

Heat Rejection
95% water / 05% air
RASC/FPGA accelerators
encryption, BLASTn, jpeg, …

BLASTn: RASC RC100 FPGA acceleration
**SGI® RASC™ Integrated Solution Stack**

Simplifies Development

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

**Diagram:**

- **Debugger (GDB)**
- **Download Utilities**
- **Application**
- **Device Manager**
- **Abstraction Layer Library**
- **Algorithm Device Driver**
- **Download Driver**
- **COP (TIO, Algorithm FPGA, Memory, Download FPGA)**
- **User Space**
- **Linux® Kernel**
- **Hardware**
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 $\rightarrow$ NUMAflex architecture
- Scale memory bandwidth $\rightarrow$ Bytes/Flop ratio
- Scale interconnection bandwidth $\rightarrow$ 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