SGI Altix Ultraviolet
Ruediger Wolff
rgw@sgi.com
SGI Altix UV1000

- Highly configurable system
  - SMP SSI with “large memory” datasharing
    - up to 256 Intel Nehalem-EX/Westmere-EX Sockets
    - Up to 16TB main memory
  - MPP w. up to 256 hardware connected partitions
    - Up to 32768 sockets, 16384 blades
    - Proven cross-partition communication
  - Hybrid system design
    - Two internal/external accelerator-cards per node
    - Supported accelerators
      - Nvidia GPU and GPGPU
      - AMD GPU and GPGPU
      - Tilera multicore
      - FPGAs Xilinx, Altera etc.

- High performance NL5 Interconnect
  - communications offload capability
  - 16-port routers
SGI UV Interconnect with Global Addressing

NUMAlink® routers connect nodes to Multi-rack UV systems
HUB snoops Socket QPI and accelerates remote access
HUB Offloads Programming models MPI, UPC, (CoArray not yet)

SGI NUMA: Directory based cache coherence....table nodes that hold copies of local data
UV Interconnect Architecture
UV HUB Based

I/O or FPGA Cards

Intel IOH

Global Memory
Intel QPI Socket or FPGAs

Global Memory
Intel QPI Socket

Global Memory
Intel QPI Socket

Global Memory
Intel QPI Socket

I/O or FPGA Cards

Coherence Directory Memory

Coherence Directory Memory

NUMAlink 5 to Other Nodes

NUMAlink 5 to Other Nodes
Altix UV Compute Blade

- Nehalem-EX
  - 8 cores per socket
  - 2.27 GHz x 4 GFLOPs per clock x 8 cores = 72.64 GFLOPS per socket (peak)
  - 145.28 GFLOPS per blade
- NUMAlink 5
  - 15 GB/s x 4 = 60 GB/s per blade
  - QPI = 25.6 GB/s x 4 = 102.4 GB/s per socket
- Millbrook Memory Buffers
  - 8.53GB/s (1067MHz DDR3 DIMMs) x 4 channels = 34.1 GB/s Read / Socket
- Intel® Scalable Memory Interconnect (SMI) = 30 GB/s/socket

SGI Proprietary
Numalink 5 Interconnect

- Numalink Cables
  - Based on 12x IB-Cables
  - Linkspeed 7.5GB/s per direction, 15GB/s bidirection
  - Numaprotocol

- UV Router
  - 16 Numalink5 ports
  - Packed in groups of 4 routers
    - QCR-Quad-compact Routers

- UV Hub-chip
  - 4 Numalink5 ports
  - 2 QPI ports
  - AMU and GRU
Blade Chassis – Independent Rack Unit

N+1 (5+1) 2.8KW PS

4C or 8C 128GB Compute Node or 128GB Memory-only Node (8GB DIMM)

1+1 48V PS For 220mm Blowers

16 blade IRU for 24” rack

18U

10U

(8) NUMAlink 5 Ports per Router Cabled to Network

(8) NUMAlink 5 Fan-In Ports per Router

Paired Nodes (Dual NUMAlink 5 Cross-Linked)
UV NL5 Barrier and MPI Reduction Performance

Very high message rates, 500k barriers/s and 3x MPI_reduce over IB

<table>
<thead>
<tr>
<th>UV Barrier Latency Performance</th>
<th>Comparison Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency = 0.3 - 1.9us (16-&gt;4096 thread)</td>
<td>50 us on 64 thread (Scali)</td>
</tr>
<tr>
<td>Barriers/s = 490K - 380K “ “</td>
<td>7.5 us on 32 thread (Altix4700)</td>
</tr>
</tbody>
</table>

![Graph showing Single element MPI_reduce and MPI_Reduce]

Collective performance over IB
Integrated I/O Riser Options

- **Base IO**
  - 3 USB 2.0 ports
  - VGA controller, with 1 port
  - SAS controller with 2 mini-SAS 8 ports
  - 2 GbE ports
  - Serial port

- **(2) Hot plug 2.5” Boot HDD**

- **Integrated PCIe Gen2**
  - (1) x16 low-profile
  - (1) x8 low-profile

- **Externalized IO**
  - (2) PCIe Gen2 x16 Cable
  - Connections to IO Expansion Chassis

- **Altix UV Blade**
SGI Altix® Ultraviolet Productfamily
Scalability with flexibility in memory and core density

- **Altix UV 1000**
  - 42U fully integrated cabinet-level solution
  - Scales to 256 sockets & 16TB shared memory (4 racks)
  - Scales to 32k sockets & 2PB global addressable memory
  - Numalink5 interconnect
  - **Up to 18.6 Tflops in shared memory**
  - **Up to 1 Pflop in global addressable memory**

- **Altix UV 100**
  - 3U 19” rackmount
  - Scales to 96 sockets & 6TB shared memory (2 Racks)
  - Numalink5 interconnect
  - **Up to 7.0 Tflops in 2 racks**

- **Altix UV 10**
  - 4U 19” rackmount
  - Scales to 4 sockets, 32 cores & 512 GB shared memory
  - **Up to 290 GF/s**
SPEC OMPL Performance

SPEC OMPL performance

Per core higher absolute performance
Better scaling

Est. SPECmpLinuxbase2001

Number of cores

0 50 100 150 200 250 300 350 400 450

0 200000 400000 600000 800000 1000000 1200000 1400000 1600000

UV 1000
Altix 4700
## Performance

### SPEC JBB2005

<table>
<thead>
<tr>
<th></th>
<th>Blades</th>
<th>JVms</th>
<th>Threads</th>
<th>.Sockets</th>
<th>Bops</th>
<th>bops/JVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altix UV 1000</td>
<td>4</td>
<td>1</td>
<td>48</td>
<td>8</td>
<td>120584</td>
<td>120584</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>8</td>
<td>64</td>
<td>8</td>
<td>342167</td>
<td>427646</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>32</td>
<td>256</td>
<td>32</td>
<td>119194</td>
<td>372481</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>128</td>
<td>256</td>
<td>32</td>
<td>126659</td>
<td>98952</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>64</td>
<td>512</td>
<td>64</td>
<td>263280</td>
<td>411376</td>
</tr>
<tr>
<td>Altix 4700</td>
<td>64</td>
<td>64</td>
<td>256</td>
<td>128</td>
<td>518045</td>
<td>80945</td>
</tr>
</tbody>
</table>

**Faktor 5**

Database Search Engine Benchmark:

expected improvement at the least 2x improvement
UV Software

• Linux OS Community Features to Support UV
  • Key items already submitted to assure adoption by UV launch
  • SLES11 SP1, RHEL6

• Drivers, APIs
  • UV HUB/Node Controller Feature Enablement

• System Management, Integration
  • Console
  • Monitoring, debug
  • Partitioning
  • Integration with storage, data sharing across UV and other systems

• RAS – enable resiliency features of UV HUB + advanced memory RAS

• Unified Parallel C source-to-source translator
  • On Intel or GCC compiler

• Ongoing system management, MPT and other Propack advances
  • Message Passing Toolkit
  • Support for SHMEM, OpenMP
Open: Comprehensive Software Stack

- Tightly integrated SGI software eases use and management
- X86 architecture provides access to huge array of software options

---

Developer Tools and Libraries

System & Resource Management
Altair PBS Professional® & Moab

SGI System Management

SGI ProPack™

SGI Foundation Software

Data Management
XFS®, CXFS®, DMF

Novell SLES or Red Hat RHEL

Altix® UV

SGI InfiniteStorage and Network Assets

---

Develop
Schedule
Manage
Perform
Enhance and support

Industry-Standard Linux OS
Deployed Hardware

---

SGI products
Third party product (available from and/or integrated by SGI)
### Full Range of Fully Supported Software Environments with SGI Systems

<table>
<thead>
<tr>
<th>VMware SW Products</th>
<th>Novell SW Products</th>
<th>SGI SW Products</th>
<th>Other ISVs SW Products</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SUSE Linux Enterprise Server**

<table>
<thead>
<tr>
<th>ISV products certified and Supported with SLES</th>
<th>ISV products certified and Supported with RHEL</th>
<th>ISV products certified and Supported with Windows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Hat SW Products</td>
<td>SGI SW Products</td>
<td>Other ISVs SW Products</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------</td>
<td>------------------------</td>
</tr>
</tbody>
</table>

**Red Hat Enterprise Linux**

<table>
<thead>
<tr>
<th>Microsoft SW Products</th>
<th>SGI SW Products</th>
<th>Other ISVs SW Products</th>
</tr>
</thead>
</table>

**Microsoft Windows Server**

<table>
<thead>
<tr>
<th>VMware ESX</th>
<th>MS Hyper-V</th>
<th>RHEL KVM</th>
<th>SLES KVM</th>
<th>No Virtualization Hypervisor (Bare Metal)</th>
</tr>
</thead>
</table>

**Full SGI Product family**

- VMware
- SUSE
- Novell
- Other ISVs
- SGI
- Red Hat
- Microsoft
- Other products

*SGI Proprietary*
# German UV Customers

<table>
<thead>
<tr>
<th>Institution</th>
<th>System</th>
<th>Cores</th>
<th>CPU Typ</th>
<th>GB memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Konrad Zuse Zentrum fuer Informationstechnik</td>
<td>UV1000</td>
<td>2432</td>
<td>cores(X7560)</td>
<td>9728</td>
</tr>
<tr>
<td>RRZN Universitaet Hannover</td>
<td>UV1000</td>
<td>2432</td>
<td>cores(X7560)</td>
<td>9728</td>
</tr>
<tr>
<td>Leibniz-Institut f. Atmosphaerenphysik e.V.</td>
<td>UV1000</td>
<td>600</td>
<td>cores(X7542)</td>
<td>3200</td>
</tr>
<tr>
<td>CTBTO Preparatory Commission</td>
<td>UV100</td>
<td>432</td>
<td>cores(X7542)</td>
<td>4608</td>
</tr>
<tr>
<td>Leibniz Rechenzentrum</td>
<td>UV1000</td>
<td>256</td>
<td>cores(X7550)</td>
<td>512</td>
</tr>
<tr>
<td>Merck KGaA</td>
<td>UV100</td>
<td>192</td>
<td>cores(X7542)</td>
<td>512</td>
</tr>
<tr>
<td>Universitaet der Bundeswehr Muenchen</td>
<td>UV1000</td>
<td>168</td>
<td>cores(X7542)</td>
<td>896</td>
</tr>
<tr>
<td>CvO Universitaet Oldenburg</td>
<td>UV100</td>
<td>120</td>
<td>cores(X7542)</td>
<td>640</td>
</tr>
<tr>
<td>Fraunhofer Einrichtung ENAS</td>
<td>UV1000</td>
<td>120</td>
<td>cores(X7542)</td>
<td>640</td>
</tr>
<tr>
<td>Heinrich-Heine-Universitaet Duesseldorf</td>
<td>UV1000</td>
<td>96</td>
<td>cores(X7560)</td>
<td>768</td>
</tr>
<tr>
<td>Access e.V</td>
<td>UV100</td>
<td>72</td>
<td>cores(X7542)</td>
<td>192</td>
</tr>
<tr>
<td>DESY</td>
<td>UV100</td>
<td>72</td>
<td>cores(X7542)</td>
<td>354</td>
</tr>
<tr>
<td>Humboldt-Universitaet zu Berlin</td>
<td>UV1000</td>
<td>48</td>
<td>cores(X7542)</td>
<td>256</td>
</tr>
<tr>
<td>SL-Rasch GmbH</td>
<td>UV1000</td>
<td>48</td>
<td>cores(X7542)</td>
<td>256</td>
</tr>
<tr>
<td>Saarländischer Rundfunk</td>
<td>UV10</td>
<td>24</td>
<td>cores(X7542)</td>
<td>128</td>
</tr>
<tr>
<td>Leibniz-Institut f. Atmosphaerenphysik e.V.</td>
<td>UV10</td>
<td>24</td>
<td>cores(X7542)</td>
<td>128</td>
</tr>
</tbody>
</table>

**Recently won:**
- Uni. Zuerich 256 cores
- CSCS 256 cores
- Uni. Linz 2048 cores

...