



CSC Test Cluster Project First Results and Experiences

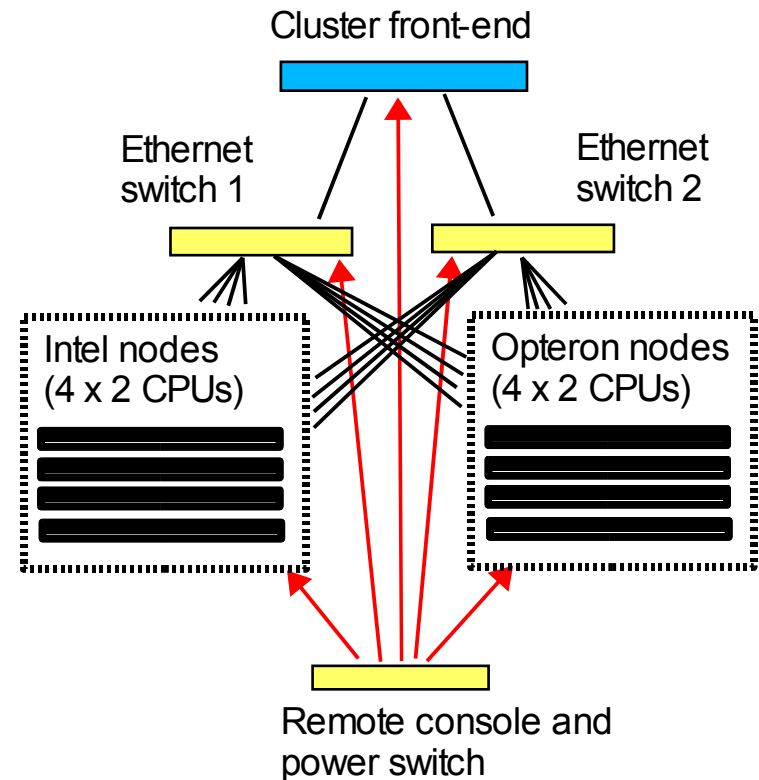
Material Science Grid Meeting
March 26, 2004

Arto Teräs
arto.teras@csc.fi



Test Cluster Hardware

- Front end 2 x 1.6GHz AMD Opteron, 4 GB RAM, 2 TB disk
- AMD Opteron and Intel Xeon nodes, 4 units each
 - Dual Opteron 1.6 GHz / Xeon 2.8 GHz, 2 GB RAM, 80 GB local disk
 - Additional AMD 4 CPU unit
- Two HP 2824 Gigabit Ethernet switches
- Remote administration hw.





Installation Experiences

- Unexpected problems with new hardware always come up
 - Important to test with exactly same hardware (same motherboard etc.) before deployment
- Rocks Linux cluster distribution version 3.1.0 used as the OS
 - Relatively easy to customize to new needs and environments
 - Managing nodes well thought out, but needs improvement in remote administration and installation of the front end
 - Does not go below OS level, so monitoring hardware failures, managing alarms, remote power switches etc. needs supporting hardware + software from the vendor
 - Overall we've been quite happy with Rocks



Disk performance tests

- Hardware: 3ware Escalade 8506-12 12-channel SATA RAID, Maxtor 7Y250M0 Maxline+II 250 GB 7200 rpm (8 disks)
- Some figures of read/write performance of big continuous files (measured with iозone)

| Configuration | Read | Write |
|---------------------------------|-------------|-------------|
| Hardware RAID 5, default | 45 MB/s | 25 MB/s |
| Hardware RAID 5, tuned | 100 MB/s | 25 MB/s |
| Software RAID 5, default | 55 MB/s (*) | 25 MB/s (*) |
| Software RAID 5, tuned | 105 MB/s | 75 MB/s |
| NFS 4 clients, HW RAID, default | | 4 MB/s |
| NFS 4 clients, SW RAID, tuned | | 25 MB/s |

(*) Slightly different test system



Disk benchmarks, continued

- Performance varies significantly depending on the configuration and tuning parameters
 - Linux software raid gives much better performance than hardware raid on the 3ware card
 - Software raid is more sensitive to the tuning parameters
- NFS performance can be a bottleneck even if the raw performance of the disk system is good
 - Increasing the block size and using asynchronous writing helps, but still not very fast
 - NFS protocol inefficient for high speed network traffic
 - Using jumbo frames would probably improve the performance, but the current switch doesn't support them => not tested



Basic benchmarks

- Xeon beats Opteron in matrix multiply (dgemm benchmark, 200 MB data)
- Opteron beats Xeon in memory bandwidth
 - Relatively close to each other with 1 CPU, Xeon drops to half speed per CPU with 2 CPUs with AMD maintaining almost the same performance
- AMD hypertransport performance (1 process, accessing memory of up to 4 CPUs) good, only about 20% drop compared to accessing the memory of the host CPU
 - allows to run jobs requiring large memory on 2 or 4 CPU boxes
- Analysis still in progress, more detailed figures later



Application benchmarks, CPMD (Juha Lento)

- Large set of results available at <http://www.theochem.ruhr-uni-bochum.de/~axel.kohlmeyer/cpmd-bench.html>
- Results vary largely even between systems with the same processor (compiler / configuration differences?)
- Rough estimate: Xeon slightly faster than similarly priced Opteron with 1 CPU, but Opteron scales better to 2 CPUs
- Encountered compilation problems on the Intel platform, so our own measurements not complete
- Also compilation problems on Opteron with more than 2 GB of memory, looking into that problem



Application benchmarks, others

- Selected applications: CPMD, Gromacs, VASP, Gaussian
- Gromacs tests in good progress
- VASP, Gaussian not yet being actively tested, starting probably next week
- Perhaps should test Siesta instead of VASP?



General remarks about compiling the benchmarks

- Compilation problems frequent, even when errors are simple solving them takes time
- Have to compile all components (MPI library, mathematical libraries and the application) with the same compiler, cross-linking can produce weird error messages
 - Supporting a large number of compilers / libraries a lot of work
- Basic compiler optimizations such as -O3 and using CPU specific functionality (flags such as -fastsse in Intel compiler) have not caused any problems
- More aggressive optimizations (e.g. -fast in Intel compiler, -Mipa=fast in PGI compiler) lead to errors in some applications