The PDC PRACE WP8 Prototype

by
Daniel Ahlin dah@pdc.kth.se PDC
Kungliga Tekniska Högskolan

Parties involved

- PDC (KTH)
  - Project leadership by Prof. Lennart Johnsson
  - Evaluating and hosting the prototype.
- Southpole
  - Acting as system integrator and vendor.
  - Coordinating assembly, delivery and physical installation of the system.
- AMD
  - Providing technical knowledge.
  - Providing CPUs.
- Supermicro
  - Providing the system platform which is also the main customization point.
- SNIC and PRACE (EU)
  - Funding the system

Primary goals

- Achieve competitive power efficiency with commodity parts.
- Ability to run existing code with no or minimal porting efforts.
- Explore possibilities of system level customization while still using commodity products (and paying commodity prices).
- Explore power/performance characteristics of running cores at lower than specified frequency.
- Utilize cooperation with system vendor in order to control features not usually available to the end customer e.g.
  - Fan speeds

The case for Power Efficiency

- You have heard this before...
- Dominant infrastructure cost
- Critical factor for most site upgrades
- Of course - Environmental impact

Minimizing porting waste

- Porting may prove to be necessary to utilize the highest end systems
  - Scaling issues unneccessary to handle at low process-counts may become critical when running wider jobs.
- Effort spent to increase scalability is likely to yield fairly long-lasting advantages.
  - Looking back – increasing general scalability has been an advantage for the last 25 years.
- However - porting to specific paradigms and systems is an uncertain investment.
  - What is the longevity of the particular paradigm?
  - What becomes of code complexity when supporting several paradigms in the same application?

Customization in a commodity setting

- HPC may not a niche-market anymore
  - Hardware for virtualized hosts share design criterias with hardware for HPC.
    - Both need memory and CPU and preferably external storage but little else.
    - At least one main difference – interconnect bandwidth and latency requirements.
- Commodity hardware
  - The case for Beowulf rehashed
- Possibilities for customer driven design within the mass-market segment.
  - Allways present on some levels but other levels are integrated – notably integration of functions on the main-board.
- Goals deemed realistic for this project
  - Influence or create a main board design either specifically for this project or one that can also be made into a more generic product.

Design Challenges

- The curse of commodity – you have to pay to get rid of things other people want
  - Do we use it?
    - Yes – fine!
    - No – is it cost efficient to get rid of it?
      - Yes – fine!
      - No – can we at least turn it off?
  - Examples:
    - Ethernet – about 2-3W / node.
    - Graphics and KVM – unknown wattage.
- Current experience – it is easier and cheaper to disable or turn components off than to remove them.
  - Does this reflect actual costs or is it mostly a matter of convenience?

Actual design – CPU and RAM

- Six 7U chassis to a standard 42U rack.
- 10 blades/systems to a chassi.
- 4 CPU-sockets to a blade.
- 6 cores to a CPU socket (AMD Istanbul 2.1GHz HE)
- A total of 1440 cores to a standard 42U rack
  - Theoretical peak performance above 12.1TF per rack.
- Projected power draw is about 30.6kW/rack or 395MFlop/W.
- 4 DIMM slots per socket.
- We have choosen 1.5Gb RAM per core and are currently evaluating both 2- and 4-GB DIMMs.
- Of course the density of this type of solution will increase significantly with the 8- and 12-core CPUs expected to be released early next year.

Actual design - Interconnect

- One Infinihost IV 36-port QDR switch per chassis
  - Passive pass-through would have been preferred but was not feasible.
  - Provides 10 internal and 10 external ports.
  - 16 ports not used and consequently disabled – obvious room for improvement.
- Chassis connecte with a set of external Infinihost IV 36-port QDR switches into a fat tree theoretical full bisection network.
- Each node has a theoretical 4Gbyte external bandwidth but each core has, at most, about 170Mbyte external bandwidth.
  - This situation will become worse. Things to do:
    - Ever higher link bandwidths.
    - Multi-rail configurations. Combining increased aggregate bandwidth with increasing the number of near neighbours in switched networks.

Actual design – other things

- Diskless solution running a minimal RAM-filesystem and most traditional root-disk contents from AFS (distributed file-system).
- Lustre as high-performance parallell filesystem.
- Aim to disable ethernet on the nodes and use only Infiniband for connectivity to the node.
- Management through traditional chassis/blade management setup i.e.:
  - I²C between Baseboard Management Controller (BMC) of each blade and the Chassis Management Controller (CMC)
  - 100Mb Ethernet between a set of controlling servers and the CMCs
  - This provides IPMI-2 to each blade.
    - Not necessary (the BMC being a potential candidate for power saving) but very convenient.

Actual design – power usage

Component	Power (W)	Perc. (%)
CPUs	2880	56.8
Memory	800	15.8
PS	355	7.0
Fans	350	6.9
Motherboards	300	5.9
HT3 Links	120	2.4
IB HCAs	100	2.0
IB Switch	100	2.0
GigE Switch	40	0.8
CMM	20	0.4
Total	5056	100.0