PDC Power Efficiency Projects

PDC Center for
High Performance Computing

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

Presentation Overview

- Part One – Describes a project done within PRACE to find power efficiency with commodity hardware.
- Part Two – Pragmatic heat reuse

PDC Center for
High Performance Computing

PDC Center for
High Performance Computing

Part One

Exploring HPC Power Efficiency on Commodity Hardware

Project goals overview

PDC Center for
High Performance Computing

- 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.

Parties involved in the project

- PDC (KTH)
  - Project leadership by Prof. Lennart Johnsson
  - Evaluating and hosting the prototype.
- South Pole AB
  - 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 (Swedish national HPC funding body) and PRACE (EU)
  - Funding the system

PDC Center for
High Performance Computing

Efficient porting

PDC Center for
High Performance Computing

- Porting may prove to be necessary to utilize the highest end systems
  - Scaling issues unnecessary 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 is not a niche-market
  - Hardware for virtualized hosts share design criteria with hardware for HPC.
    - Both need memory and CPU and external storage but little else.
    - At least one main difference – interconnect topology, bandwidth and latency requirements.
- Commodity hardware
  - Cost efficient (mostly)
  - Known (staffing, longevity of knowledge, etc)
- Possibilities for customer driven design within the mass-market segment.
  - Always 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.

KTH VETENSKAP OCH KONST

PDC Center for High Performance Computing

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 design convenience?

PDC Center for
High Performance Computing

Actual design – CPU and RAM

PDC Center for
High Performance Computing

- 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 4-GB DIMMs.
- Of course the density of this type of solution have increased significantly with the 8- and 12-core CPUs released 2010

Actual design - Interconnect

PDC Center for
High Performance Computing

- One Infinihost IV 36-port QDR switch per chassis
  - Passive pass-through would have been preferred but was not feasible.
  - Provides 10 internal and 16 external ports.
  - 16 ports not used and consequently disabled – obvious room for improvement.
- Chassis connected with 5 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 file-system and most traditional root-disk contents from AFS (distributed file-system).
- Lustre as high-performance parallel file-system.
- System Ethernet on the nodes is disabled - only Infiniband for connectivity to the systems.
- Management through traditional chassis/blade management setup i.e.:
  - I^2C and 100Mb Ethernet 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.

PDC Center for
High Performance Computing

Designed power usage

PDC Center for
High Performance Computing

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

Linux embedded prevalence

- Quite a lot of BMCs and CMCs are implemented as a SOC Linux design.
  - Sometimes SDKs are publicly available. I.e:
    - http://sourceforge.net/projects/raritan-oss/)
    - http://www.supermicro.com/products/nfo/IPMI.cfm
  - Risk free (almost) implementation of added management/instrumentation options.
  - Possibility of sharing development effort with vendors.
  - AMD and Intel have now - and are increasing the possibilities of - side-band CPU control.

KTH VETENSKAP OCH KONST

PDC Center for High Performance Computing

Project Process Summary

- We have to understand that board designers are not likely to be HPC system architects.
  - Great ideas you have can actually be easily and rather cheaply implemented. The board designer might just not have understood it is something you wanted.
- Which items can an HPC - customer or interest group – effect:
  - Infrastructure investments – moderate impact – quick results
  - CPU design – uncertain impact – long view
  - Memory design – likely low impact – very long view
  - Main board design – potentially large impact – quick results
  - Packaging – moderate impact - quick results
  - Firmware/BIOS – high impact – quick results
  - OS – high impact – quick results
  - Application layer – we own the problem

PDC Center for High Performance Computing

Part Two

Pragmatic Heat Reuse

PDC Center for
High Performance Computing

Project goals overview

- Add about 800kW cooling capacity for a new 306TF TPP Cray XE6.
- Retain savings from heat reuse within university.
- Conserve floorspace.

PDC Center for
High Performance Computing

Parties involved in the project

- PDC (KTH)
  - Project leadership by Gert Svensson
  - Hosting site
- Akademiska Hus
  - Owner of premises (owns practically all real estate used by Swedish universities and research institutes)
- Cray – system vendor
- Sweco Energiguide AB– energy consultants
  - Technical design of the heat reuse

PDC Center for High Performance Computing

Starting points

- System is a 16 rack Cray XE6 with TPP about 306 TF and 48.5 TB RAM
- 44kW per rack or 704 kW in total in current configuration
- Existing district heating/cooling system
  - City wide system with local circuits within university
  - Stockholm district cooling (largest in the world) uses free cooling (sources are the Baltic Sea and lake Mälaren).
- Heat reuse already in place but:
  - Uses heat pumps and is only run some parts of the year
  - At utility company's discretion to make use of
  - Low return temperatures (about 18°C)
  - Low incentives for exploiting small scale possibilities
- PDCs neighbour the School of Chemical Science and Engineering at KTH is:
  - Extensively renovating their premises
  - Large users of fresh air (fume hoods and ventilated chemical storage)

KTH VETENSKAP OCH KONST

PDC Center for
High Performance Computing

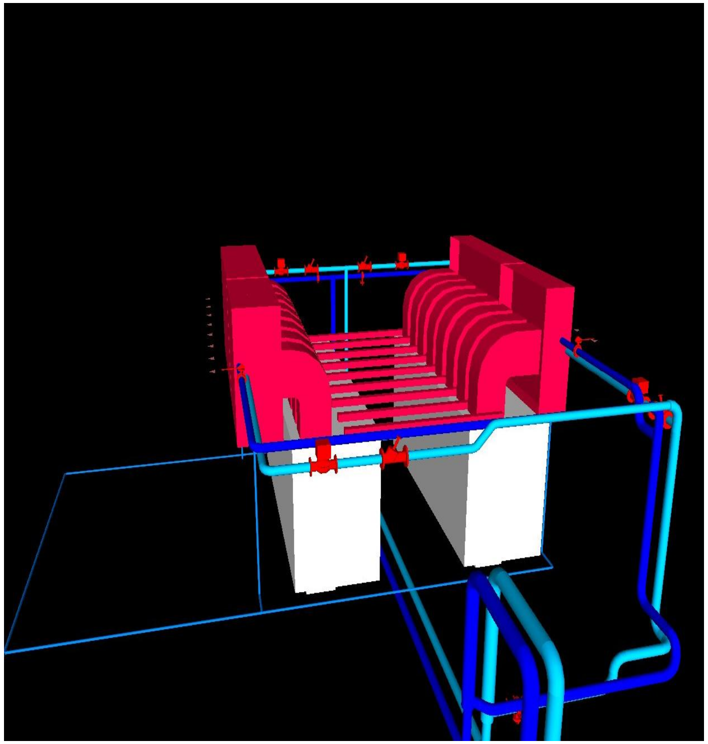
Situation summary

I. Pay for district cooling – about €43/MWh
II. Pay for district heating – about €54/MWh
III. Pay for power
    I. PDCs price is about €108/MWh
    - For 2010 average spot price at Nordpool is €88/MWh

- Project aims to decrease the costs of the first two items by heat reclamation within the campus.

PDC Center for
High Performance Computing

Solution



PDC Center for
High Performance Computing

- Fitted hoods on top of the racks (Cray XE6 is air-cooled bottom-up) guides the air to heat-exchangers which heats the water passing through them to about 24°C
- This 24°C degree water while not ordinarily considered useful for heating purposes in traditional radiators can be used to heat large air-volumes which is in demand by the School of Chemistry

Projections

- This system is projected run a surplus of €113000 (annuities included) per year from a capital expenditure of €700000 for a repayment time of about 4 years.
  - Life time of computer system is likely about 4 years (disregarding updates).
  - Life time of infrastructure is projected to be 15 to 20 years.
- After paying for the investment the reuse can defray about 24% of the power costs (and 30% if PDC were paying spot prices which can be seen as a lowest possible power cost for PDC)

PDC Center for
High Performance Computing

Which Efficiency Matters

PDC Center for
High Performance Computing

- Is the computer system power efficient?
  - Well – let us say this – the Cray XE6 is designed primarily for performance.
- Is the the setup energy efficient?
  - Electrical power is a poor heating source (at least in Sweden were various forms of district heating is usually available).
  - But if
    - – the computer system is considered necessary

  then the setup can be considered making a virtue of necessity and also considered efficient from a total system point of view.
- Is the setup cost-effective – yes – as seen above.
- Is the setup environmentally sound – is it green?

PDC Center for
High Performance Computing

DISCUSSION

(else I assume you agree with everything I've said)