EcoGrappe - Energy Consumption Management in Clusters

Eugen Feller

Ph.D. student
Advisor: Dr. Christine Morin
INRIA MYRIADS research team

July 8, 2010

Why do we need energy management?

Overview

- Always been critical for mobile devices (e.g. laptops)
- Last two decades of distributed computing
  - Performance at any cost (i.e. FLOPS)
- Consequently: Immense cooling, power and backup power costs
  - Japanese Earth Simulator (2000 - 2004): 18 MW for 35.86 Tflops $\Rightarrow$ 10 million dollar/year for power and cooling
  - Data centers: 61 billion kWh of U.S. energy in 2006 $\Rightarrow$ Enough energy to power 5,8 million average U.S. households
- Additionally: Increased carbon footprint
- Energy conservation efforts
  - Green Destiny, BlueGene/L, The Green500 List, The Green Grid, INRIA GREEN-NET, COST Action IC0804, etc.

Oak Ridge National Laboratory, Jaguar - Cray XT5-HE - TOP 500

TOP500 List - June 2010 (1-100)

R_{max} and **R_{peak}** values are in TFlops. For more details about other fields, check the TOP500 description.

Power data in KW for entire system

next

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron Six Core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60
2	National Supercomputing Centre in Shenzhen (NSCS) China	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU / 2010 Dawning	120640	1271.00	2984.30	
3	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2009 IBM	122400	1042.00	1375.78	2345.50

Energy conservation in laptops and servers

How to save energy

- Save energy on mobile devices
  - Utilize idle times
  - Standby or shutdown of resources (e.g. CPU, Disk, Memory and Display)
- Servers have different workloads [7]
  - Less or no idle times
  - Low-power (e.g. shutdown) modes often not feasible
  - Limited use of techniques from mobile devices
- One possible solution
  - Create or extend idle times
  - Move load and shutdown unused nodes [6]

Measuring the power consumption

Measuring the power consumption

- Node level
  - Wattmeter (e.g. Watt's Up Pro)
  - Power Distribution Unit (PDU) with power per outlet monitoring
  - ACPI enabled PSU
  - Intelligent Platform Management Interface (IPMI)
- Component level
  - Non-trivial task ⇒ No internal power measurement equipment available
  - Take information from data sheet ⇒ Only peak power consumption
  - Derive implicitly from the Performance Monitoring Unit (PMU) registers information
  - Make a custom solution (e.g. separate the DC lines)

Managing the power consumption

Managing the power consumption

- Several approaches to reduce power [7]
  - Power off idle resources
  - Slow down resources
  - Move work to others
  - Less work with less quality
- Ideal: Power management without performance degradation
  - Energy proportional to time: $E \propto t$
- Many low-power modes enhance the execution time
  - Negative effect on energy consumption

Techniques for power management

CPU

- Methods to lower the power consumption [7]
  - CPU with lowest power-to-clock ratio (Not a good idea)
  - CPU with highest IPC-to-power ratio (Not always a good idea)
  - Power down CPU when idle (Interrupt)
  - Use less power when demand permits it (DVS)
- Dynamic Voltage Scaling (DVS)
  - Power consumption = ACV^2f [2]
  - Voltage decrease $\Rightarrow\sim$ Proportional frequency reduction ($f \propto V$)

Disk

- Methods to lower the power consumption [7]
  - Spin down while idle
  - Shutdown parts of the device logic
  - Do work more efficiently (multi-speed disks)
- Transition costs ($Spun - down \Rightarrow Idle$ and $Idle \Rightarrow Spun - down$)
  - Depend on the device (up to 40 joules)
  - Can have bad impact on total energy savings [4]

Memory

- Power savings at the costs of performance
  - Multiple power saving modes
  - Different transition overheads (Time and Energy)
- Power modes influence the logic
  - Turning of row and column decoders and clock signals [7]
  - Keeping refresh signal
- Same trade-offs as for all devices
  - Lower power: Delay access ⇒ Increase energy consumption
  - Extensive transitions ⇒ Reduce savings

Networking

- Node level
  - Slowdown link $\Rightarrow$ Adaptive Link Rate (ALR)
- Infrastructure level
  - Turn off unused links
- Improve the hardware
  - Optimize the layout of the router components (e.g. buffers, links, etc.)
- Similar energy vs. performance trade-offs to other server components (e.g. CPU)

Power usage by components

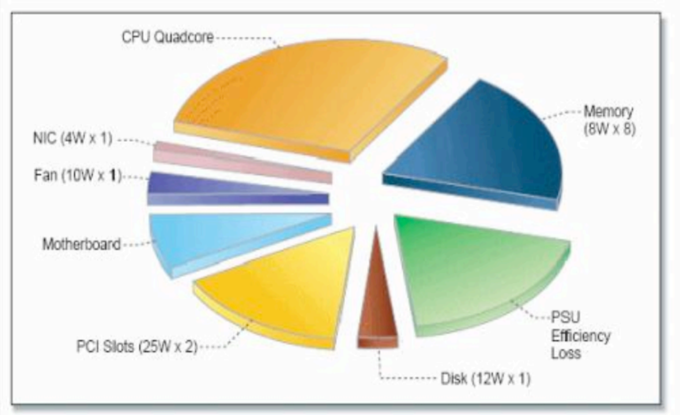


Figure: Intel Labs [5], 2008

Overview

Overview

- Work can be classified into two areas: Low-power computing and Energy-aware frameworks [3]
- Low-power computing
  - Use low-power hardware (e.g. PowerPC 440 CPU and SoC technology)
  - Performance through high-density and parallel applications
- Energy-aware frameworks
  - Use traditional components
  - Adapt the system performance to match the workload
  - Utilize various low-power modes provided by the hardware

Low-power computing and Energy-aware frameworks

Examples

- Low-power computing
  - Green Destiny, BlueGene/L, etc.
- Energy-aware frameworks [1]
  - Node level ⇒ DVS, Core On/Off, Request Batching, Transcoding, Code optimization
  - Cluster level ⇒ Node On/Off, Virtualization, Moab Workload Manager
  - Grid level ⇒ INRIA GREEN-NET

Overview

EcoGrappe

- New energy conservation initiative
  - Funded by the French ANR research agency
  - Started in December 2009
- Objective
  - Lower total energy consumption of clusters $\Rightarrow$ Decrease energy costs
  - Generate less heat $\Rightarrow$ Increase reliability
- Three partners involved
  - INRIA Rennes (MYRIADS research team)
  - Kerlabs
  - EDF R&D

Kerrighed operating system

Kerrighed operating system

- Kerlabs: Spin-off from the INRIA PARIS (now MYRIADS) research team
  - Developer and maintainer of the Kerrighed operating system
- Kerrighed: Single System Image (SSI) operating system for clusters
  - Started as a research project in 1999 at the INRIA PARIS research team
  - Extension to the Linux operating system
- Provides basic functionality for our work
  - Customizable global scheduler
  - Process migration
  - Node addition and Node removal

Overview

Energy-aware framework for clusters

- Exploit present technology
  - Use the Kerrighed operating system and its global scheduler
  - Complement it with a resource manager
- Resource manager
  - Provides job specific information (e.g. job size, duration, etc)
  - Holds the job history information
  - Takes global energy conservation decisions
- Kerrighed operating system and its global scheduler
  - Provides resource information (e.g. resource availability, resource utilization, power consumption, etc)
  - Takes local energy conservation decisions

Objectives

Energy-aware job scheduling

- Main contribution
  - Design of energy conservation policies and algorithms
- Take energy conservation decisions such as
  - Concentrate jobs on a subset of nodes and turn off the unused ones
  - Scheduling of jobs according to the energy costs (day/night)
  - Scheduling of jobs to the most energy-efficient systems
  - Move jobs away from *hot servers* and avoid thermal emergencies
  - Learn application specific energy-consumption $\Rightarrow$ place jobs on servers with the best energy vs performance trade-off

Current status

Hardware and Software

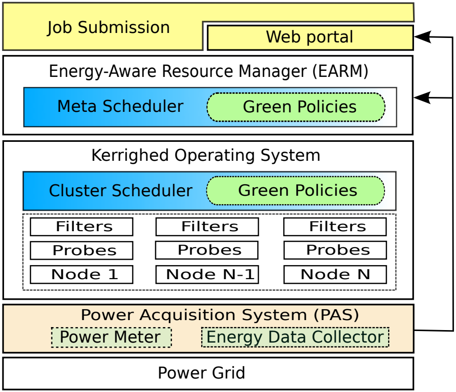
- Studied the state of the art in energy management
  - First deliverable released at the end of May 2010
  - Website: *http://ecograppe.inria.fr*
- Hardware to measure the power consumption
  - PDU with per outlet power monitoring (16 outlets)
  - Supports the SNMP protocol
- Experimental Kerrighed cluster
  - Four nodes: Dell PowerEdge 1950
  - Together 16 Intel Xeon 2.33GHz CPUs
- Software installed
  - Resource monitoring: Ganglia, Munin and MRTG
  - Resource manager/scheduler: Torque and Maui

Current status

Measurements

- Benchmark used to measure the performance of the cluster
  - HPCC (HPC Challenge Benchmark)
- Initial power consumption measurements
  - Idle power: 175 Watt
  - Busy power: 219 Watt (running *cpuburn*)

Energy-aware framework for clusters - Architecture



Future work

Future work

- Propose final architecture of our energy conservation framework for clusters
  - Based on Kerrighed and its global scheduler
  - Combined with a resource manager
- Study the energy consumption under different workloads
  - CPU, memory and I/O intensive
- Extend and design new energy-aware task placement policies/algorithms
- Build a prototype of the framework and verify the algorithms

Conclusion

Conclusion

- EcoGrappe (12/2009): New initiative for energy conservation in clusters
  - Complements other energy conservation efforts
- First steps towards the objective already taken
  - Studied the state of the art
  - Initial work on proposing an architecture
  - First experiments with a real system
- Next steps
  - Design of energy-aware task placement policies and algorithms
  - Verification within a prototype

Thank you for your attention!

References

- Ricardo Bianchini and Ram Rajamony.
  Power and energy management for server systems.
  IEEE Computer, 37:2004, 2003.
- E.N. (Mootaz) Elnozahy, Michael Kistler, and Ramakrishnan Rajamony.
  Energy-efficient server clusters.
  In *In Proceedings of the 2nd Workshop on Power-Aware Computing Systems*, pages 179–196, 2002.
- Wu-chun Feng, Xizhou Feng, and Rong Ge.
  Green supercomputing comes of age.
  IT Professional, 10(1):17–23, 2008.
- Dennis Lee.
  Energy management issues for computer systems.
- Lauri Minas and Brad Ellison.
  The problem of power consumption in servers.
- Eduardo Pinheiro, Ricardo Bianchini, Enrique V. Carrera, and Taliver Heath.
  Dynamic cluster reconfiguration for power and performance, 2002.
- Eduardo Souza De Albuquerque Pinheiro.
  Energy conservation for server systems.
  PhD thesis, New Brunswick, NJ, USA, 2005.
  Director-Bianchini, Ricardo.