State of the art and continuous learning of hardware adaptation possibilities

Georges Da Costa (and Davide Careglio)

WG1 of COST IC0804 about energy efficiency in large scale distributed systems

Plan

1 Introduction

2 Subsystems

3 Current Practices in Large Scale Distributed Systems

4 Conclusion

Cost Action 0804

Energy efficiency in large scale distributed systems

1. **State of the art and continuous learning of hardware adaptation possibilities** (lead by *Georges Da Costa* and *Davide Careglio*).
2. Characterization of energy consumption and energy efficiency.
3. Adaptive actions for distributed systems.
4. Characterization of performance-energy saving trade-off.
5. Scientific coordination and dissemination of the works and definition of a common opened framework.

WG1: State of the art and continuous learning of hardware adaptation possibilities

Which is better ?

- Improve 1% on something vital
- Improve 50% on something unimportant

Provide two type of inputs to other Working Groups:

- Relative power consumption of subsystems
- How to interact with them (captors, actuators)

Relative consumption

Subsystems:

- Processor : 40 to 60%
- Memory : 7 to 17%
- Disk : 4 to 6 %
- Motherboard : 15 to 35%
- Fan : 5 to 20%

Sources: Fan2007, Lim2008

What a Wonderful World

If there was a simple a complete standard for captors and actuators

APM First one, from Intel and Microsoft, 1992 (last version, 1.2, 86 pages)

ACPI Replaced APM, first released in 1996, originally designed by Intel, Microsoft, and Toshiba - later joined by HP and Phoenix. Last version is "Revision 4.0a", on April 5, 2010, **731 pages !**

Vendors Data Sheets with vendor-dependent advertisement included !

Point of view : Power !

For Cost action 0804 *Power, Energy, Heat* are of interest

- *Power+Machine* leads to *Heat*
- *Power+Application* leads to *Energy*

Focus of WG1

- What are the possibilities : actuators
- What are their impact : Instantaneous Power

Heat and Energy are for grown-up : Schedulers, Decision centers,...

Subsystems

- **Processor (and GPU, accelerators,...)**
- **Memory**
- **Disk/Flash**
- **Fan**
- Network (**interfaces**, infrastructure)
- PSU
- *Current practices*

Some subsystems are **dynamic**, some are static.

Plan

1 Introduction

2 Subsystems

3 Current Practices in Large Scale Distributed Systems

4 Conclusion

Processor, King of the kings

At the same time

- Most consuming element
- Largest variation of power consumption
- Remark: Difficult to heat (small area)

$Power = DynamicPower + LeakagePower$
$DynamicPower = \alpha * freq * Voltage^2$
Problem: Impossible to have freq and/or voltage approach 0

Processor, actors

Possible to:

- P-States: change frequency and voltage (following tables)
- C-States: defines how deep it sleeps (changes Leakage power)
  - The deeper it sleeps, the less leakage and the higher latency to come back

High indirect impact (Fan, PSU,...)

Interfaces

- AMD PowerNow!, Cool'N'Quiet
- Intel SpeedStep, Turbo Boost
- ACPI
  - *Thermal control Zone*: prevent processor overheating
  - *C-States*: Control how deep the CPU sleeps when is not active
  - *P-States*: Control frequency/voltage operational points

Other computing elements

Way simpler than processors:

- GPU : possible to change frequency but controls are scarce
- CELL : energy efficient (in watt/flop) but controls are limited
- FPGA : same law as processor but frequency can approach 0

Memory

Heterogeneous technology

RDRAM power states such as active, standby, nap and power-down (like C-States)

MobileDDR deep power down

DRAM Change frequency (reducing refreshing time)

Operating mode	Energy consumption (nJ per cycle)	Resynchronization cycles
Active	3.570	0
Standby	0.830	2
Nap	0.320	30
Power-Down	0.005	9000

Hard Drives

Two main sinks (2/3 of energy consumption)

- Mechanical part
- Communication part

Disk drive technology allows three main control knobs:

- Spindle speed (limited: from 7200 to 5400 RPM when idle, nearly halves consumption)
- Seek speed (Just-In-Time (JIT) seek mode, Automatic Acoustic Management (AAM))
- Disk power mode (ATA8-ACS standard: Active, Idle, Standby, Sleep)

Fan

- Intermediate between cooling and computing
- Often forgotten
- Large impact
  - At full speed can go up to 16-20%
  - $Power = \alpha * Speed^3 + \beta$
  - Can be important during boot (for clusters)
- Can be controlled and probed using ACPI

Network jungle

Network interfaces

- Adaptive Link Rate solutions
- IEEE P802.3az Energy Efficient Ethernet Task Force

Network appliances

- Each provider has its own technology
- Cisco routers can shut down part of themselves using proprietary interface
- Dynamic power is about 3% for routers

PSU, realm of simplicity

A power supply,
it can only be probed,
yeah, ACPI.

Plan

1 Introduction

2 Subsystems

3 Current Practices in Large Scale Distributed Systems

4 Conclusion

Current Practices in Large Scale Distributed Systems

Two main interests for Data Centers:

- Benchmarking
- Cooling/Heat reuse

Motto: The more data are available about power usage, the easier it is to optimize a data center consumption

Benchmarks: Different levels

- Perf/Watt: mainly nodes
- PUE (Power usage effectiveness): ratio (total consumption)/(computing element)
- Green500
- EnergyStar helps evaluating the energy impact of a building
- Tools are available to evaluate energy efficiency of IT infrastructure (IBM for instance)

Niagara falls yahoo center uses circulating exterior air to cool the servers, and is able to achieve a PUE of around 1.1.

Cooling: Innovative datacenters

- Google boats using sea water for water-cooling
- Free cooling buildings
- Heat-reuse (Intel Haifa center in Israel, leads to 200.000$ savings)

Remark: most data centers are cooled at 20°C while they could operate at 26°C

Complexity of managing datacenters

Beware of false intuition

Energy Star study ([Energy-Star2010], slide 23) shows that the external temperature has little impact on the overall electricity consumption of data centers when using air conditioning.

- Processors like to be cold, disks prefer some heat
- Some project are coupling data centers with renewable energy sources

Remark: Servers are not afraid of the dark

Plan

1 Introduction

2 Subsystems

3 Current Practices in Large Scale Distributed Systems

4 Conclusion

Conclusions

Goal of WG1: reduce the burden for researchers

- Extract simple facts and leverages from a complex landscape
- ACPI is too broad and complex
- Some components are not encompassed by ACPI (GPU, memory)
- Some subsystems do not even have standards (GPU, memory,...)

Complete brochure is available: **www.cost804.org**

Any question/remarks/comments ?

IBM Green Data Center Man