ECE 571 – Advanced Microprocessor-Based Design Lecture 28

Vince Weaver

http://web.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

9 November 2020

Announcements

- HW#9 will be posted, read AMD Zen 3 Article
- Remember, no class on Wednesday

When can we scale CPU down?

- System idle
- System memory or I/O bound
- Poor multi-threaded code (spinning in spin locks)
- Thermal emergency
- User preference (want fans to run less)

Non-CPU power saving

- RAM
- GPU
- Ethernet / Wireless
- Disk
- PCI
- USB

GPU power saving

- From Intel lesswatts.org
  - Framebuffer Compression
  - Backlight Control
  - Minimized Vertical Blank Interrupts
  - Auto Display Brightness
- from LWN: http://lwn.net/Articles/318727/
  - Clock gating or reclocking
  - Fewer memory accesses: compression. Simpler background image, lower power

- Moving mouse: 15W. Blinking cursor: 2W
- Powering off unneeded output port, 0.5W
- LVDS (low-voltage digital signaling) interface, lower refresh rate, 0.5W (start getting artifacts)

More LCD

- When LCD not powered, not twisted, light comes through

- Active matrix display, transistor and capacitor at each pixel (which can often have 255 levels of brightness). Needs to be refreshed like memory. One row at a time usually.

Ethernet

- PHY (transmitter) can take several watts
- WOL can draw power when system is turned off
- Gigabit draw 2W-4W more than 100Megabit 10 Gigabit 10-20W more than 100Megabit
- Takes up to 2 seconds to re-negotiate speeds
- Green Ethernet IEEE 802.3az

WLAN

- power-save poll – go to sleep, have server queue up packets. latency
- Auto association – how aggressively it searches for access points
- RFKill switch
- Unnecessary Bluetooth

Disks

- SATA Aggressive Link Power Management – shuts down when no I/O for a while, save up to 1.5W
- Filesystem atime
- Disk power management (spin down) (lifetime of drive)
- VM writeback – less power if queue up, but power failure potentially worse

Soundcards

- Low-power mode

USB

- autosuspend. Can sometimes cause issues
- off by default as some USB you disable don't come back

Results from REU measurement

- ATX measurement
- USB measurement

  How much power does your keyboard use?

  Keyboard latency

A History of Power Management on x86

Halt Instruction

- Oldest power-saving interface on x86
- Tells CPU to stay idle until an interrupt comes in
- 486-DX4 and later enters low-power mode
- Ring 0. The OS does this when idle
- Similar instruction available on 65c816
- ARM has `wfi` in ARMv7 and maybe `hlt` in ARMv8?

APM – Advanced Power Management

- For laptops
- Developed by Intel and Microsoft, 1992
- Made obsolete by ACPI
- Full On / APM Enabled / Standby / Suspend or Hibernate / Off
- Calls to BIOS. BIOS often buggy.

ACPI – Advanced Configuration and Power Interface

- `http://www.acpi.info/presentations/ACPI_Overview.pdf`
- Developed by Intel, Microsoft and Toshiba, 1996 Later HP and Phoenix
- Full ACPI interpreter needed.
- APM was a black box to Operating System. ACPI works with OS
- ACPI code in theory provided by Intel or similar, no need for each manufacturer to implement (like APM)

- OS-directed power management
- Hardware registers for interface
- BIOS provides tables, motherboard initialization

ACPI Sleep States

- Global vs Sleep
- G0/S0 – Working
- G1 Sleeping
  - S1 – Caches flushed, CPU stopped, CPU and RAM power maintained
  - S2 – CPU powered OFF
  - S3 – Standby, Sleep, Suspend to RAM. (RAM still on)
  - S4 – Hibernate/Suspend to Disk – memory to disk
- G2 (S5) – "Soft Off" – power off, but power still supplied

to power switch and wake on lan, etc

- G3 – “Mechanical Off” – all power removed

ACPI C-States (Idle)

- C0 – operating
- C1 – Halt – processor not executing, but can start nearly instantaneously (Intel C1E – lower voltage too)
- C2 – Stop-Clock – all state is stored, but might take some time to get going again (C2E – lower voltage)
- C3 – Sleep – Processor does not keep cache coherent, but otherwise holds state
- Processor specific (Haswell up to C10)

ACPI P-States (Performance/Operational)

- actual values can sometimes be configured via MSR access.
- Some V/F combinations unstable/unsafe so BIOS only exports known good combinations
- P0 – max power and frequency
- P1 – less than P0, DVFS
- P2 – less than P1, DVFS
- Pn – less than P(n-1), DVFS

ACPI T-States

- throttling
- Linear reduction in power, linear reduction in performance
- Does not save Energy! (halve the frequency, double the time)
- Mostly used for passive cooling

ACPI D-States

- for devices such as modems, Cd-ROM, disk drive
- D3 can be hot or cold (hot has aux power and can request being moved back up, cold it is turned off)

CPU Scaling

- Intel SpeedStep

- Enhanced speed step. Change V and F at different points. Slower to change frequency if V not changed first. Bus clock keeps running even as PLL shut down 10ms transition

- AMD PowerNow! (laptop) Cool'n'Quiet (desktop)

- VIA PowerSaver/LongHaul – Fine grained DVFS

- p4-clockmod – mainly for thermal management, skip clocks, hurt performance without saving energy (throttling)
- IBM EnergyScale
- Transmeta LongRun – leakage varies due to process variation Longrun2 monitors performance/leakage and varies Vdd and Vt

DVFS

- Voltage planes – on CMP might share voltage planes so have to scale multiple processors at a time
- DC to DC converter, programmable.
- Phase-Locked Loops. Orders of ms to change. Multiplier of some crystal frequency.
- Senger et al ISCAS 2006 lists some alternatives. Two phase locked loops? High frequency loop and have programmable divider?

- Often takes time, on order of milliseconds, to switch frequency. Switching voltage can be done with less hassle.

Non-x86 Power Saving

IBM EnergyScale

- Thermal reporting
- Static and Dynamic Power Save
- “Power Folding” – reduce the number of CPUs reported to the OS until they are all busy
- Power Capping (like RAPL)
- Fan Control – Avoid “over-cooling”

- Processor Nap – 2ms to wake up
- Processor Winkle (as in Rip Van) – 10-20ms to wake up, 95% of power

ARM Cortex A9 (Pandaboard)

- Cortex-A9 Technical Reference Manual, Chapter 2.4 Power Management

- Energy Efficient Features
  - Accurate branch prediction (reduce number of incorrect fetch)
  - Physically addressed caches (reducing number of cache flushes)
  - Use of micro TLBs

- caches that use sequential access information? reduce accesses to tags
  - small instruction loops can operate without access icache

- Potentially separate power domains for CPU logic, MPE (multi-media NEON), and RAMs

- Full-run mode

- Run with MPE disabled

- Run with MPE powered off

- Standby – entered with `wfi` instruction. Processor mostly shutdown except part waiting for interrupt
- Dormant – caches still powered
- Shutdown

Pandaboard Power Stats

- Wattsuppro: 2.7W idle, seen up to 5W when busy
- http://ssvb.github.com/2012/04/10/cpuburn-arm-cortex-a9.html
- With Neon and CPU burn:

Idle system	550 mA	2.75W
cpuburn-neon	1130 mA	5.65W
cpuburn-1.4a (burnCortexA9.s)	1180 mA	5.90W
ssvb-cpuburn-a9.S	1640 mA	8.2W

Operating System Power Saving Strategies

- We look primarily at Linux, as it is open source and technical debates happen in the open

- Windows and OSX often have measurably better laptop Energy behavior due to tuning and better hardware testing

Governors

- ondemand – dynamically increase frequency if at 95% of CPU load
  introduced in 2.6.9

- performance – run CPU at max frequency

- conservative – increase frequency if at 75% of load

- powersave – run CPU at minimum frequency

- userspace – let the user (or tool) decide

Governors – cont

- Various tunables under /sys/devices/system/cpu
- Can trigger based on ACPI events (power plug in, lid close)
- Laptop tools
- `cpufreq-info` and `cpufreq-set`
  Need to be root

User Governors

- typically can only update once per second
- ondemand people claim it reacts poorly to bursty behavior
- Powernowd – scale based on user and sys time
- cpufreqd
- Obsolete with introduction of “ondemand” governor?

Sources of Info for Governors

- System load
- performance counters
- input from user?

TurboBoost

- Nehalem/Ivy Bridge/Sandy Bridge (AMD has similar Turbo CORE)
- Some Core2 had similar “Intel Dynamic Acceleration”
- Kicks in at highest ACPI Pstate
- “Dynamic Overclocking”

TurboBoost – from HotChips 2011 Slides

- Monitors power, current, thermal limits, overclocks
- 100 uarch events, leakage function of temp and voltage
- P1: guaranteed stable state
  P0: turbo boost, maximum possible
- 12 temp sensors on each core
- PECI – an external microcontroller, used to control fans, package power

TurboBoost example

- From WikiPedia Intel_Turbo_Boost article
- Core i7-920XM
- Normal freq 2.0GHz
- 2/2/8/9 – number of 133MHz steps above with 4/3/2/1 cores active
- 2.26GHz, 3.06GHz, 3.20GHz