Управление энергопотреблением в FreeBSD

Alexander Motin
mav@FreeBSD.org

- Вопрос оптимизации энергопотребления был и всегда будет актуален для мобильных систем. Стационарные системы в последнее время так-же часто сравниваются по критерию энергоэффективности.
- Оптимизация энергопотребления зачастую предполагает достижение баланса между производительностью и функциональностью системы и ее энергопотреблением: отключенный монтор не потребляет энергии, но и не выполняет своих функций.
- Многие из энергосберигающих технологий требуют активного управления со стороны ОС: только ОС по ряду признаков может знать, когда можно отключить монтор без ущерба для функциональности.

- В ряде случаев, энергопотребление и тепловыделение являются лимитирующим фактором для повышения производительсности. Как результат, появляется обратный эффект: чем больше энергии сэкономлено, тем больше дополнительных вычислений можно произвести не нарушая заданных ограничений.
- Это особенно заметно на фоне возростающей избыточности современных систем:
  - 2, 4, 8, ... процессорных ядер -- значительная их часть зачастую простаивает, бесполезно нагревая общий кристалл, ограничивая рабочие напряжение и частоту.
  - применение специализированных акселераторов -- в то время как одни задачи требуют мощного процессора общего назначения, другие выполняются специальным оборудованием (GPU). Баланс загрузки сильно варьируется в зависимости от конкретной задачи, зачастую вызывая простои отдельных компонентов.

- Энергопотребление системы можно разделить на два основных вида: активное, когда система расходует энергию на выполнение задачи, и статическое, когда система простаивает.
- Различные технологии энергосбережения ориентируются на различные виды: можно снизить яркость монитора, а можно его отключить. У каждого метода есть свои преимущества и недостатки.

- ПРОЦЕССОР
- Процессор - один из основных потребителей энергии. Настольные и серверные x86 процессоры потребляют до 130Вт энергии при полной нагрузке. Мобильные - до 60Вт.
- Процессоры простаивают большую часть времени.
- Современные x86 системы предоставляют большой выбор технологий управления энергопотребленем:
  - активное потребление:
    - P-states (Intel SpeedStep, Cool'N'Quiet, PowerNow!),
    - T-states (throttling),
    - TurboBoost,
  - статическое потребление:
    - C-states (idle).

- P-states - состояния процессора, характеризующиеся различными рабочими частотами и напряжениями. Состояние P0 является состоянием максимальной производительности и энергопотребления.
- Источником информации о поддерживаемых P-states является ACPI BIOS, но встречаются и другие варианты. Для каждого состояния сообщается: рабочая частота, максимальная потребляемая мощность, способ входа и задержка входа/выхода.

```
%dmesg | egrep "^est"
est0: <Enhanced SpeedStep Frequency Control> on cpu0
est1: <Enhanced SpeedStep Frequency Control> on cpu1
est2: <Enhanced SpeedStep Frequency Control> on cpu2
est3: <Enhanced SpeedStep Frequency Control> on cpu3
est4: <Enhanced SpeedStep Frequency Control> on cpu4
est5: <Enhanced SpeedStep Frequency Control> on cpu5
est6: <Enhanced SpeedStep Frequency Control> on cpu6
est7: <Enhanced SpeedStep Frequency Control> on cpu7
%sysctl dev.cpu |grep freq
dev.cpu.0.freq: 1200
dev.cpu.0.freq_levels: 2934/106000 2800/82000 2667/70000 2533
/62000 2400/53000 2267/46000 2133/39000 2000/33000 1867/28000
 1733/24000 1600/20000 1467/17000 1333/14000 1200/11000
```

- T-states - в первую очередь служат для защиты процессора от критического перегрева путем пропуска части тактов генератора. Менее эффективны чем P-states, но позволяют снизить потребление для процессоров не имеющих поддержки состояний C1E/C2. По умолчанию управляются аппаратно, но допускают и ручное управление.
- Управляются напрямую, либо через ACPI.

```
% dmesg |grep p4tcc
p4tcc0: <CPU Frequency Thermal Control> on cpu0
p4tcc1: <CPU Frequency Thermal Control> on cpu1
% sysctl dev.cpu |grep freq
dev.cpu.0.freq: 2795
dev.cpu.0.freq_levels: 2795/-1 2445/-1 2096/-1 1746/-1 1397/-
1 1048/-1 698/-1
```

- FreeBSD по умолчанию использует комбинацию из доступных P-states и T-states, компануя их в единый набор частот, где каждая частота образуется из умножения частоты P-state на коэффициент пропуска тактов T-state. При этом преимущество отдается P-states, как более энерго-эффективным.
- Текущая частота задается глобально для всех процессоров через sysctl или /etc/rc.conf. Целесообразно доверить это демону powerd, который будет динамически подстраивать ее под уровень загрузки.
- Для процессоров поддерживающих P-states и C-states, использование T-states можно отключить (установив переменные загрузчика hint.p4tcc.0.disabled=1 и hint.acpi_throttle.0.disabled=1), так как как минимальный выигрыш в энергопотреблении обычно не оправдывает значительного снижения производительности.

- C-state - состояния простоя процессора. Каждое состояние характеризующиеся определенным уровнем статического энергопотребления и временем, необходимым для входа/выхода из состояния:
  - C0 - процессор полностью активен;
  - C1 - исполнение остановлено инструкцией HLT, но все логические блоки функционируют; пробуждение мгновенно;
  - C2 - основные блоки процессора отключены от тактового генератора; пробуждение практически мгновенно;
  - C3/C4/C5 - процессор отключен от генератора и системной шины, напряжение снижено, когерентность кешей может не поддерживаться, LAPIC таймер может не функционировать; пробуждение требует времени для восстановления рабочего напряжения и частичной инициализации состояния.

  - C6 - процессор практически полностью обесточен, состояние регистров переписано во внутреннюю статическую память; пробуждение требует времени для восстановления напряжения, полной инициализации и восстановления состояния.
- Различные процессоры могут поддерживать различные комбинации C-states. Первичным источником информации о поддерживаемых C-states и их параметрах служит ACPI BIOS. Для каждого состояния ACPI BIOS сообщает: способ и семантику входа, ожидаемый относительный уровень энергопотребления, а так-же время выхода из состояния.

- ACPI скрывает от ОС реальные C-states процессора и в некоторых случаях эмулирует дополнительные, используя System Management Mode:
  - C1E (Intel) - использует метод входа как у C1 (HLT), но реально входит в C2; Метод необходим только для старых ОС, не поддерживающих C2.
  - C1E (AMD) - использует метод входа как у C1 (HLT), но когда все ядра процессора вошли в C1 - весь процессор переводится в C4. Это состояние может нарушить работу LAPIC таймера и автоматически блокируется FreeBSD, если LAPIC таймер используется системой.
- ACPI BIOS систем на процессорах AMD, как правило, не сообщают ОС о поддержке C-states кроме обязательного C1, предпочитая обходиться автоматической активацией их через механизм C1E.

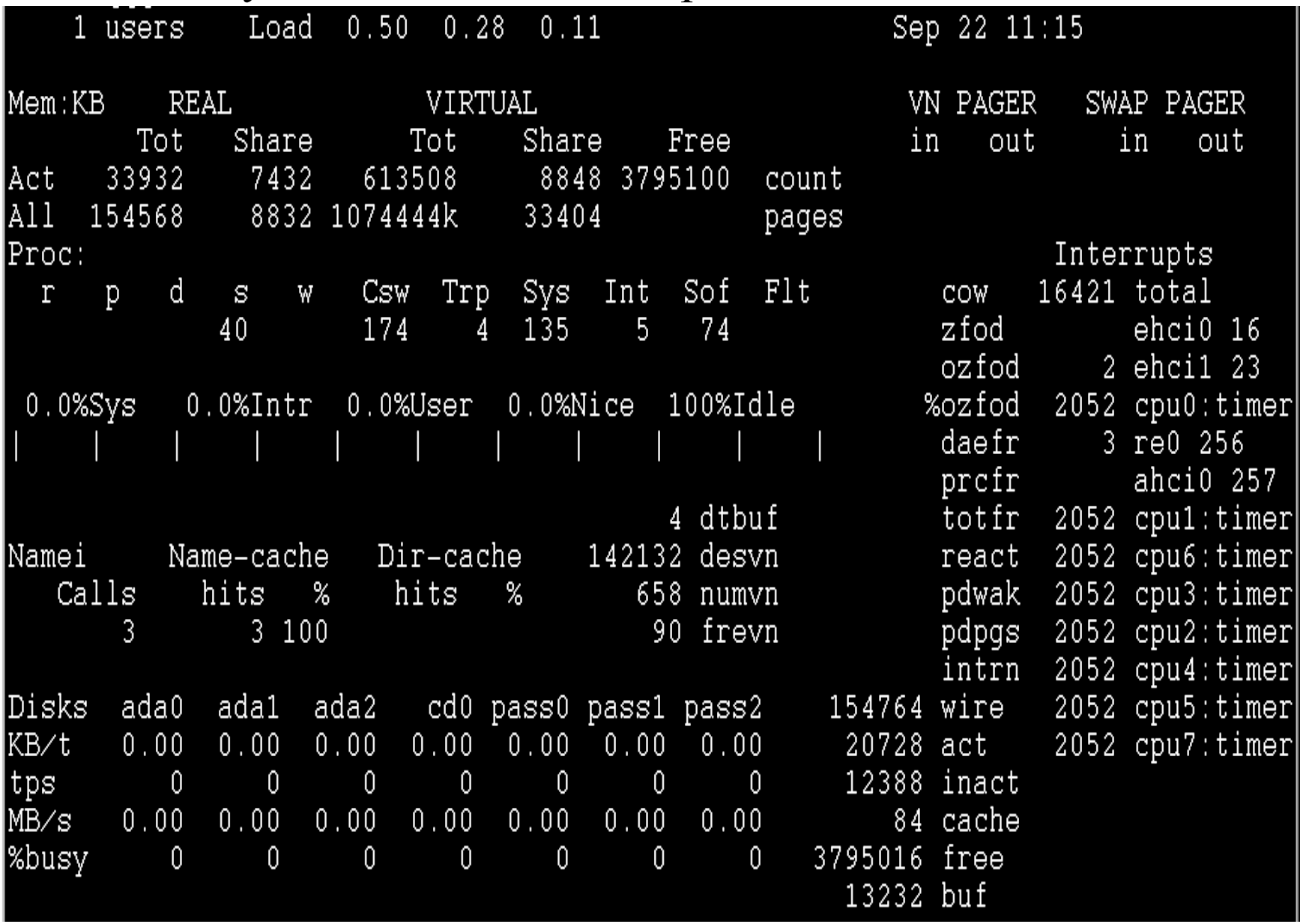
- C-states платы Asus P7H55-M SI с процессором Core i7 870

```
%sysctl dev.cpu | grep cx_
dev.cpu.0.cx_supported: C1/32 C2/96 C3/128
dev.cpu.0.cx_lowest: C3
dev.cpu.0.cx_usage: 0.05% 0.03% 99.91% last 11235us
dev.cpu.1.cx_supported: C1/32 C2/96 C3/128
dev.cpu.1.cx_lowest: C3
dev.cpu.1.cx_usage: 0.02% 0.03% 99.94% last 5000us
dev.cpu.2.cx_supported: C1/32 C2/96 C3/128
dev.cpu.2.cx_lowest: C3
dev.cpu.2.cx_usage: 0.21% 0.10% 99.68% last 29946us
dev.cpu.3.cx_supported: C1/32 C2/96 C3/128
dev.cpu.3.cx_lowest: C3
dev.cpu.3.cx_usage: 0.15% 0.11% 99.72% last 30394us
dev.cpu.4.cx_supported: C1/32 C2/96 C3/128
dev.cpu.4.cx_lowest: C3
dev.cpu.4.cx_usage: 2.24% 1.95% 95.80% last 12189us
dev.cpu.5.cx_supported: C1/32 C2/96 C3/128
dev.cpu.5.cx_lowest: C3
dev.cpu.5.cx_usage: 0.53% 0.86% 98.59% last 16124us
dev.cpu.6.cx_supported: C1/32 C2/96 C3/128
dev.cpu.6.cx_lowest: C3
dev.cpu.6.cx_usage: 0.07% 0.07% 99.85% last 39685us
dev.cpu.7.cx_supported: C1/32 C2/96 C3/128
dev.cpu.7.cx_lowest: C3
dev.cpu.7.cx_usage: 0.10% 0.08% 99.81% last 52874us
```

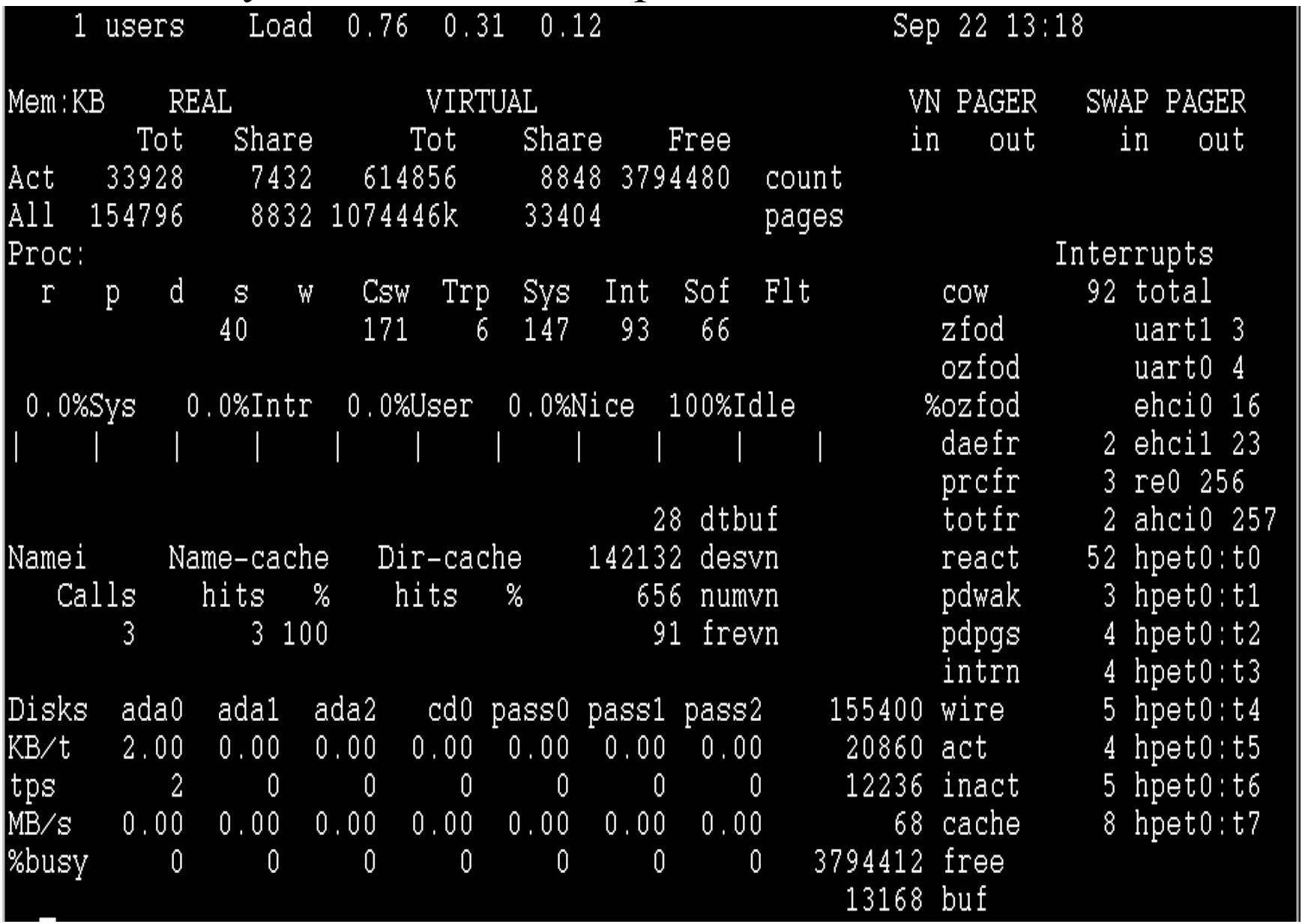
- Для использования C-states в FreeBSD необходимо установить в /etc/rc.conf переменные performance_cx_lowest и/или economy_cx_lowest.
- Использование CPU C-states C3 и глубже может потребовать отключения использования LAPIC таймера, нестабильного в этих состояниях. В случае FreeBSD 8.x для этого нужно установить переменную загрузчика hint.apic.0.clock=0. В случае FreeBSD 9.x можно так-же использовать переменную kern.eventtimer.timer для выбора любого другого доступного таймера.

- Для эффективного использования C-states необходимо обеспечить, чтобы интервал пробуждений процессора из сна был значительно больше времени выхода из соответствующего C-state.
- Как правило, основной причиной пробуждения процессора являются прерывания таймера. Единственным вариантом решения для FreeBSD 8.x сейчас является снижение частоты таймера, через установку переменной загрузчика kern.hz=100 и отключение дополнительного RTC таймера: hint.atrtc.0.clock=0. Результат - 100 прерываний на ядро.
- В FreeBSD 9-CURRENT мной недавно реализована новая подсистема управления таймерами, описанная в eventtimers(7), позволяющая избавиться от большей части прерываний таймера на простаивающих процессорах без уменьшения значения HZ. Результат - 10-20 прерываний на ядро и может быть улучшен.

- Вывод `systat -vm 1` на ненастроенной FreeBSD:



- Вывод `systat -vm 1` на настроенной FreeBSD 9-CURRENT:



- Вывод sysctl на настроенной FreeBSD 9-CURRENT:

```
# sysctl dev.cpu |grep cx_usage
dev.cpu.0.cx_usage: 3.10% 0.58% 96.31% last 10253us
dev.cpu.1.cx_usage: 0.00% 0.00% 100.00% last 18135us
dev.cpu.2.cx_usage: 0.00% 0.00% 100.00% last 33606us
dev.cpu.3.cx_usage: 0.00% 0.00% 100.00% last 33627us
dev.cpu.4.cx_usage: 0.00% 0.00% 100.00% last 78316us
dev.cpu.5.cx_usage: 0.00% 0.00% 100.00% last 80808us
dev.cpu.6.cx_usage: 0.00% 0.00% 100.00% last 136140us
dev.cpu.7.cx_usage: 0.00% 0.00% 100.00% last 136138us
```

- TurboBoost - технология автоматического временного увеличения частоты процессора и интегрированного video ядра при условии соблюдения требований по температуре и потребляемой мощности, реализованная в процессорах Intel серий Core i5/i7.
- TurboBoost использует повышение множителя процессора на несколько ступеней в зависимости от текущих условий.
- TurboBoost задействуется только при условии что процессор находится в состоянии P0 - то-есть уже работает на полной штатной частоте.
- Помимо температуры и потребляемой мощности величина повышения частоты зависит от числа процессорных ядер, активных в данный момент. Чем больше ядер неактивны (находятся в состояниях простоя C3 или глубже), тем выше может быть поднята частота остальных.

- Величина возможного поднятия частоты зависит от конкретной модели процессора:

Модель	Кол-во ядер	Частота	Увеличение множителя *
Core i5-650	2	3200MHz	1/2 (+266)
Core i5-750	4	2667MHz	1/1/4/4 (+533)
Core i7-870	4	2933MHz	2/2/4/5 (+666)
Core i7-620UM	2	1067MHz	5/8 (+1067)
Core i7-840QM	4	1867MHz	2/2/8/10 (+1333)

- * - величина повышения множителя (одна ступень - 133MHz) для 4/3/2/1 активного ядра

- ЭКРАН
- Подсветка 12’’ экрана ноутбука может потреблять до 5Вт энергии.
- Яркость подсветки может управляться либо аппаратными кнопками, либо через набор sysctl acpi_video(4):

```
%kldload acpi_video
%sysctl hw.acpi.video
hw.acpi.video.crt0.active: 1
hw.acpi.video.lcd0.active: 1
hw.acpi.video.lcd0.brightness: 100
hw.acpi.video.lcd0.fullpower: 100
hw.acpi.video.lcd0.economy: 100
hw.acpi.video.lcd0.levels: 100 100 0 10 25 40 55 70 85 100
```

- PCI УСТРОЙСТВА
- PCI предоставляет интерфейс для управления питанием отдельных PCI устройств, определяя набор состояний от D0 - устройство активно до D3 - устройство отключено.
- FreeBSD позволяет отключить устройства для которых не загружены драйвера путем установки переменной загрузчика hw.pci.do_power_nodriver:

```
sdhci1@pci0:10:2:3:     class=0x050100 card=0x011b1025 chip=0x07511524 rev=0x00
hdr=0x00
    vendor     = 'ENE Technology Inc'
    device     = 'PCI Secure Digital / MMC Card Reader Controller'
    class      = memory
    subclass   = flash
    cap 01[80] = powerspec 2  supports D0 D1 D2 D3  current D0
none3@pci0:10:9:0:      class=0x0c0010 card=0x011b1025 chip=0x8024104c rev=0x00
hdr=0x00
    vendor     = 'Texas Instruments (TI)'
    device     = 'TSB43AB23 1394a-2000 OHCI PHY/link-layer Controller'
    class      = serial bus
    subclass   = FireWire
    cap 01[44] = powerspec 2  supports D0 D1 D2 D3  current D3
```

- USB УСТРОЙСТВА
- USB предоставляет интерфейс для управления питанием отдельных устройств.
- FreeBSD позволяет управлять питанием устройств, используя утилиту usbconfig:

```
%usbconfig -u 2 -a 2 power_off
%usbconfig
ugen0.1: <UHCI root HUB Intel> at usbus0, cfg=0 md=HOST spd=FULL (12Mbps) pwr=SA
VE
ugen1.1: <UHCI root HUB Intel> at usbus1, cfg=0 md=HOST spd=FULL (12Mbps) pwr=SA
VE
ugen2.1: <EHCI root HUB Intel> at usbus2, cfg=0 md=HOST spd=HIGH (480Mbps) pwr=S
AVE
ugen3.1: <UHCI root HUB Intel> at usbus3, cfg=0 md=HOST spd=FULL (12Mbps) pwr=SA
VE
ugen4.1: <UHCI root HUB Intel> at usbus4, cfg=0 md=HOST spd=FULL (12Mbps) pwr=SA
VE
ugen5.1: <UHCI root HUB Intel> at usbus5, cfg=0 md=HOST spd=FULL (12Mbps) pwr=SA
VE
ugen6.1: <EHCI root HUB Intel> at usbus6, cfg=0 md=HOST spd=HIGH (480Mbps) pwr=S
AVE
ugen2.2: <Acer CrystalEye webcam SuYin> at usbus2, cfg=255 md=HOST spd=HIGH (480
Mbps) pwr=OFF
ugen6.2: <USB2.0 Hub vendor 0x05e3> at usbus6, cfg=0 md=HOST spd=HIGH (480Mbps)
pwr=SAVE
ugen6.3: <CP2102 USB to UART Bridge Controller Silicon Labs> at usbus6, cfg=0 md
=HOST spd=FULL (12Mbps) pwr=ON
ugen6.4: <USB Multimedia Keyboard BTC> at usbus6, cfg=0 md=HOST spd=LOW (1.5Mbps
) pwr=ON
ugen6.5: <USB RECEIVER Logitech> at usbus6, cfg=0 md=HOST spd=LOW (1.5Mbps) pwr=
ON
```

- РАДИО
- Bluetooth и WiFi адаптеры могут потреблять несколько Ватт энергии.
- Зачастую радио модули в ноутбуке можно отключить используя аппаратные кнопки.
- Для WiFi может быть достаточно ifconfig ... down.
- Большинство Bluetooth адаптеров - USB устройства.

- ЖЕСТКИЕ ДИСКИ
- 2.5’’ Hitachi SATA HDD потребляет 2W в активном состоянии или 1W при остановленном шпинделе.
- В случае классического ATA: atacontrol spindown ...
- В случае CAM ATA: camcontrol idle|standby|sleep ...
- Каждый запуск шпинделя уменьшает ресурс диска.

- SATA ИНТЕРФЕЙС
- Дифференциального метод передачи SATA интерфейса передает служебную последовательность сигналов во время простоя. Это ведет к ненужному расходу энергии. SATA имеет два режима экономии энергии в простое:
  - PARTIAL - передача сигналов прекращается, но питание поддерживается; экономит 0,5Вт, пробуждение - 50-100мкс.
  - SLUMBER - полностью отключает передатчик; экономит 0,8Вт и пробудление - 3-10мс.
- Находясь в этих режимах невозможно определить наличие устройства, что усложняет работу hot-plug.
- Управление через переменные загрузчика:
  - ata(4): hint.ata.X.pm_level,
  - ahci(4): hint.ahcich.X.pm_level,
  - siis(4): hint.siisch.X.pm_level,
  - mvs(4): hint.mvs.X.pm_level.

- РЕЗУЛЬТАТЫ
- Стационарный Core i7-870 с боксовым куленом:
  - полная загрузка: 85C;
  - простой без оптимизации: 55C;
  - простой P-states+C-states: 32C.
- Время компиляции порта net/mpd5 в одну нить:
  - без оптимизации: 12,02c;
  - C-states: 10,79c (-10%).
- Энергопотребление и время автономной работы ноутбука Acer TM6292 (Core2DuoT7700 @ 2.40GHz):
  - без оптимизации: 19,5Вт, 2 часа 24 минуты;
  - предустановленный Windows XP: 3 часа 20 минут;
  - полная оптимизация: 10,1Вт, 4 часа 47 минут.

- Вопросы есть? :)