ACPI a Software Suspend

Lukáš Turek

MFF UK

1.9.2007

O čem to bude

- ACPI
- DSDT tabulka
- Suspend to disk
  - swsusp
  - suspend2
  - µswsusp
- Suspend to RAM

ACPI - úvod

- Advanced Configuration and Power Interface
- Specifikace pro komunikaci OS s hardwarem a naopak (správa napájení je jen podmnožina funkce)
- Abstrakce funkcí hardware
  - Příklad: teplotu CPU je možné číst přes ACPI (**`/proc/acpi/thermal_zone/THRM/temperature`**), nebo přímo komunikací s čipem pomocí lm_sensors *
- Nahrazuje APM (Advanced Power Management)
  - V APM řídil správu napájení BIOS, operační systém měl jen omezené možnosti
  - Linux APM podporuje (**`CONFIG_APM`**, démon **`apmd`**)
  - I na vypnutí počítače je potřeba APM nebo ACPI

ACPI - návrh

- Na specifikaci ACPI se podílely firmy Hewlett-Packard, Intel, Microsoft, Phoenix, a Toshiba
- Specifikace byla navržena multiplatformě pro IA-32 (x86) a IA-64 (Itanium)
- Není omezena jen na přenosné počítače, obsahuje i funkce pro víceprocesorové servery
- Součástí specifikace je i programovací jazyk ASL (ACPI Source Language), ten se překládá do bytekódu
  - Kompilátor Intelu: http://www.intel.com/technology/IAPC/acpi/downloads.htm
  - Možno programovat event handlery (obsluha událostí)
    - Například ACPI timer: probuzení počítače po 15 minutách
    - Potenciální bezpečnostní riziko (rootkit), ale nepřežije restart počítače

ACPI – realita

- „second system syndom“ – první systém nestačil, druhý musí umět všechno a radši něco navíc
- Specifikace ACPI má 631 stran: http://www.acpi.info/DOWNLOADS/ACPIspec30b.pdf
- Výrobci nedodržují ACPI standard (a to ani ti, kteří se na něm přímo podíleli)
- V implementaci ACPI v BIOSu jsou často chyby
  - A ty obchází až ovladač (pro Windows)
- Implementace v Linuxu je velká asi jako TCP/IP stack *
- První doporučovaný krok, pokud počítač nenabootuje je parametr kernelu **`acpi=off`**
- OLPC (One Laptop Per Child) ACPI nepoužívá vůbec

ACPI tabulky

- Informace o systému, v jazyce AML
  - RSDP (Root System Description Pointer)
  - RSDT (Root System Description Table)
  - **DSDT (Differentiated System Description Table)**
  - XSDT (Extended System Description Table)
  - FADT (Fixed ACPI Description Table)
  - FACS (Firmware ACPI Control Structure)
  - SBST (Smart Battery Table)
  - ECDT (Embedded Controller Boot Resources Table)
  - MADT (Multiple APIC Description Table)
  - SRAT (System Resource Affinity Table)
  - SLIT (System Locality Distance Information Table)
  - SSDT (Secondary System Descriptor Table)

DSDT tabulka

- Největší a nejdůležitější z ACPI tabulek
- Dodávaná s hardware (součást BIOSu)
- OS ji načítá při bootu
- Hiearchický formát
- Informace o konfiguraci hardware
- **`/proc/acpi/dsdt`**
- AML bytekód je možné dekompilovat: *
  - **`cat /proc/acpi/dsdt > dsdt.aml`**
  - **`iasl -d dsdt.aml > dsdt.asl`**

DSDT – pokračování

- V DSDT tabulce mohou být chyby (výrobce zajímá jen jestli to funguje ve Windows)
- Naštěstí je možno DSDT tabulku dekompilovat, upravit, zkompilovat a zadat kernelu:

```
Device Drivers  --->
  Generic Driver Options  --->
    [ ] Select only drivers that don't need compile-time external firmware

Power management options (ACPI, APM)  --->
  ACPI (Advanced Configuration and Power Interface) Support  --->
    [*]   Include Custom DSDT
    (dsdt.aml) Custom DSDT Table file to include
```

ACPI States

- Stav počítače
  - G0 – normální běh
  - G1 – spánek
    - S1 – CPU běží, ale nevykonává instrukce
    - S2 – nepoužívá se
    - S3 – stav CPU a periferií uložen v paměti
    - S4 – stav systému včetně paměti uložen na disk
  - G2 – vypnutý počítač, může být probuzen např. ze sítě
  - G3 – při odpojení od napájení
- Vedle toho stav periferií (D0-D3) a procesoru (C0-C3)
- Co váš počítač podporuje zjistíte v **`/sys/power/state`** * (standby = S1, mem = S3, disk = S4)
- S4 (suspend to disk s podporou BIOSu na speciální partition) už v kernelu není, nahrazuje ho swsusp

ACPI - využití

- Různé informace v **`/proc/acpi`**
  - teplota CPU: **`/proc/acpi/thermal_zone/THRM/temperature`**
    - lm_sensors ale řekne víc
  - stav baterie: **`/proc/acpi/battery/BAT0/state`**
  - a další podle hardware...
- LED diody
  - podle výrobce, např. **`/proc/acpi/asus/wled`**
  - chystá se unifikovaný ovladač
- Tlačítka
  - démon acpid
  - možné tlačítkům přiřadit akce (skripty) *
  - skripty se spouští pod rootem, takže spuštění Firefoxu po stisku tlačítka s modrým E je trochu komplikovanější...

Software Suspend

Software Suspend

- Kompletní stav systému se uloží na disk, počítač se vypne
- Při bootu se stav zase obnoví
- Nepotřebuje podporu podporu BIOSu (ani ACPI), pro hardware je to normální vypnutí a zapnutí
- Užitečné pro notebooky, ale i pro desktop: systém naběhne rychleji, nemusíte znovu spouštět programy a přemýšlet kde jste skončili...
- Několik implementací (sdílí dost kódu):
  - swsusp (v kernelu)
  - suspend2 (patch)
  - µswsusp (s podporou userspace)

swsusp

- Již dlouho v kernelu
- Současní správci: Pavel Machek, Rafael Wysocki
- Dokumentace: Documentation/power/swsusp.txt
- Problémy, které musí software suspend řešit
  - Potřebuji uložit kompletní stav paměti
    – ale na to potřebuji paměť (buffery disku,...)
  - Při ukládání paměti nesmí nic do paměti zapisovat, jinak bude image nekonzistentní, zařízení se tedy musí uspat, aby nezapisovaly přes DMA
    – ale pro zápis na disk potřebuji řadič, ten může být na PCI

swsusp - princip

- Paměť se atomicky (při zakázaný přerušeních) kopíruje do volné paměti
  - Image tedy může mít velikost maximálně 50% volné paměti
  - Nadbytečná data v paměti se musí odswapovat, musí se vyprázdnit disková cache...
  - Po probuzení se odswapovaná data načítají pomalu stránku po stránce (náhodné přistupy na disk), zatímco image by se mohlo načíst sekvenčně během několika vteřin
- Zařízení se před kopírování uspí a pak zase probudí
  - probouzí se všechna zařízení, těžko se dá poznat která jsou potřeba (disk připojený na SCSI řadič v PCI za PCI bridge...)

swsusp – upozornění

- Filesystém zůstává připojený
  - je možné nabootovat jiný OS, ale ten nesmí připojovat oddíly, které byly připojeny při suspendu
  - grub musí být schopen načíst kernel, ale připojený oddíl může být v nekonzistentním stavu – doporučuji používat boot partition
- Probouzet se musí s přesně stejným kernelem jako při uspání – je-li jiný, swsusp to pozná a bez varování image smaže
- swsusp automaticky ukládá image na první swap partition, pro obnovu potřebuje partition zadat parametrem **`resume=/dev/hdX`**
- Ovladač řadiče disku musí být zakompilovaný v kernelu

swsusp - použití

- Konfigurace kernelu

```
Power management options (ACPI, APM)  --->
[*]   Software Suspend
(/dev/hda2) Default resume partition
```

  - Default resume partition nahrazuje parametr **`resume=`**
- Spuštění suspend:
  - echo disk > /sys/power/state

swsusp – proces

- Uspání (suspend)
  - Zastavení uživatelských procesů
  - Zastavení vláken kernelu
  - Uvolnění paměti
  - „Zmražení“ zařízení (*devices*): **`suspend(PMSG_FREEZE)`**
  - Atomická kopie paměti
  - Probuzení zařízení: **`resume()`**
  - Zápis image do swapu
  - Uspání zařízení: **`suspend(PMSG_SUSPEND)`**
  - Vypnutí počítače
- Obnova (resume) je opačný proces

swsusp – potřebná podpora

- Software Suspend je transparentní pro uživatelské procesy (ale mohou být problémy, například s náhlou změnou času)
- Musí být upraveno každé kernelové vlákno
  - volání **`try_to_freeze()`** na bezpečném místě, kde nedrží žádný zámek
  - Příklad: **`mm/pdflush.c`**
  - Vlákna potřebná pro uložení image mají flag **`PF_NOFREEZE`**
- Ovladač každého zařízení musí implementovat funkce **`suspend()`** a **`resume()`** pro uložení a načtení stavu
  - Pokud je neimplementuje, musí se modul před uspáním vyhodit, suspend2 má blacklist modulů
  - Někdy ani unload nepomůže (**`sdhci`**)
  - Příklad: **`drivers/net/skge.c`**

Suspend2

- Zatím mimo kernel
- Patch na http://www.suspend2.net/
- Správce: Nigel Cunningham
- Umožňuje vybrat swap partition, navíc možnost ukládat image do souboru
- Volitelně komprese (LZF) a šifrování pomocí cryptoapi v kernelu
- Ukazatel průběhu zápisu image
- Runtime konfigurace v **`/sys/power/suspend2/`**
- Spuštění suspend:
  `echo > /sys/power/suspend2/do_suspend`
  - Ale lepší je použít skript *hibernate*, který vyhodí moduly na blacklistu, přepne do konzole atd., navíc podporuje všechny 3 implementace Software Suspend

Suspend2 – ukládání paměti

- Hlavní výhoda suspend2: image může mít velikost skoro celé paměti
  - Suspend2 dělí stránky do 2 skupin
    - Pageset1 – ty, které je nutno kopírovat atomicky, např. kernel
    - Pageset2 – stránky uživatelských procesů (*pagecache*), které se po uspání procesů nezmění
  - Od verze 2.2.9 jsou v Pageset2 jen read-only stránky, původní chování se nastaví
    `echo 1 > /sys/power/suspend2/full_pageset2`
    - Mě bez tohoto nastavení suspend selže, že se nepodařilo zastavit kswapd0
  - Stačí uvolnit paměť pro kopii stránek v Pageset1
    - řádově 10MB
- Proč tedy není v kernelu?
  - příliš mnoho kódu, autor odmítá rozdělit na nezávislé části

µswsusp

- Userspace Software Suspend
  - Přesun co nejvíce činností do userspace (ukládání image, komprese, šifrování)
- Autoři: Pavel Machek, Rafael Wysocki
- Nutná podpora už je v kernelu
- Documentation/power/userland-swsusp.txt
- Potřebný userspace program: http://suspend.sf.net/
  - databáze potřebných hacků pro jednotlivé počítače *

µswsusp - princip

- Device **`/dev/snapshot`**
  - read – přečtení image
  - write – obnova image
  - ioctl – nastavení parametrů
- Sdílí omezení swsusp
  - image maximálně 50% RAM
  - userspace program nesmí zapisovat na oddíl disku, který byl připojen v okamžiku snapshotu => nelze jednoduše uložit snapshot na root partition
  - userspace program by neměl z připojeného disku ani číst
- Obnova také vyžaduje userspace program => je potřeba initrd/initramfs

Suspend to RAM

- ACPI S3
- Stav periferií a procesoru se uloží do paměti, napájena je jen paměť
- V kernelu musí být vybráno ACPI Sleep States (**`CONFIG_ACPI_SLEEP`**)
- Suspend se spustí echo **`mem > /sys/power/state`**
  - pokud máte štěstí, počítač se uspí
  - pokud máte opravdu hodně štěstí, tak se i probudí
- Pozor na **`acpid`**: probouzíte-li počítač tlačítkem Power, událost se dostane do systému a **`acpid`** vypne počítač

Suspend to RAM - problémy

- Na rozdíl od Software Suspend se neprovede boot, a BIOS nemůže inicializovat periferie
  - Typicky se neinicializuje grafická karta a nerozsvítí se displej – několik triků:
    - suspend z X Window
    - vbetool post
    - acpi_sleep=s3_bios
    - acpi_sleep=s3_mode
- Dnes je spíše štěstí, když S3 funguje, ale blýská se na lepší časy, do věci se vložil přímo Linus Torvalds
  - Debugování Suspend to RAM pomocí hashe ukládané do RTC (hardwarových hodin) – je možné zjistit, kde to vytuhlo