Linux Idle Power Checkup

Len Brown
Principal Engineer
Intel Open Source Technology Center
len.brown@intel.com

Aug 10, 2010

Linuxcon - Boston, MA

Topics for Today

- CPU idle power states (C-states)
- Linux cpuidle
- intel_idle
- Idle power measurements

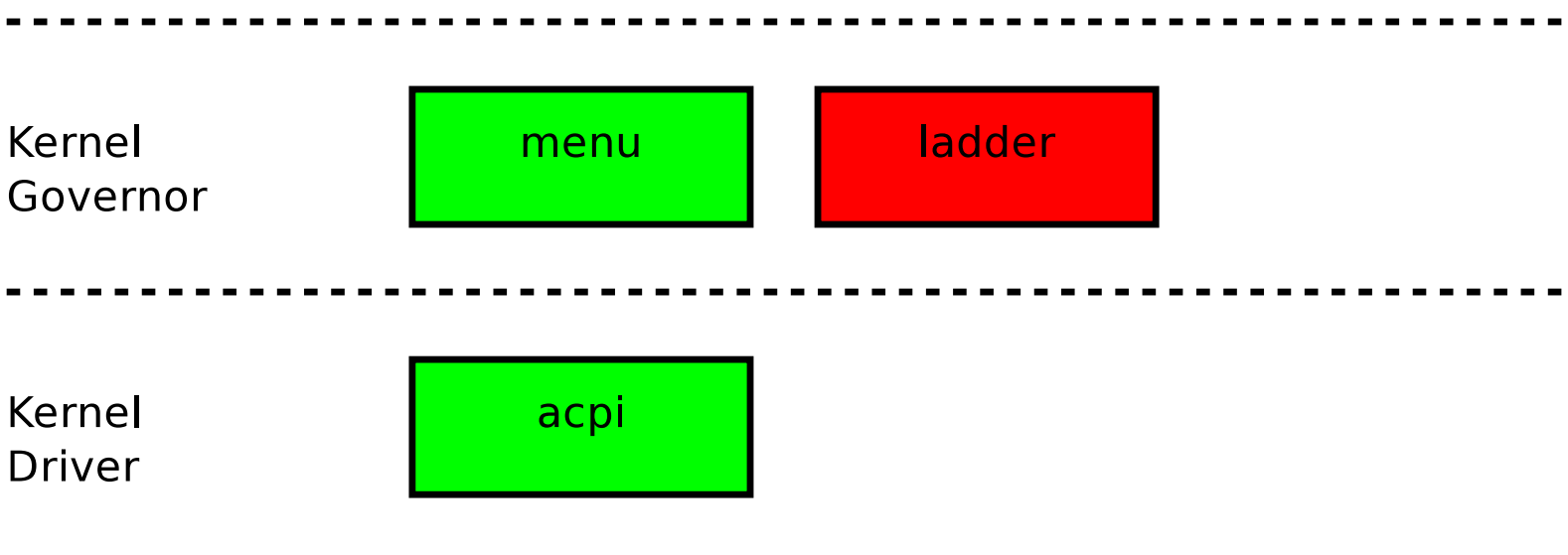
Viewing C-states with turbostat

```
# ./turbostat cat /dev/zero > /dev/null
 cor CPU   %c0   GHz  TSC   %c1    %c3    %c6   %pc3   %pc6
           8.49 3.64 3.38  12.09   5.57  73.85   0.00   0.00
   0   0   0.08 2.96 3.38  99.92   0.00   0.00   0.00   0.00
   0   6  99.98 3.65 3.38   0.02   0.00   0.00   0.00   0.00
   1   2   0.01 2.89 3.38   0.08   0.00  99.91   0.00   0.00
   1   8   0.02 3.44 3.38   0.07   0.00  99.91   0.00   0.00
   2   4   0.01 3.03 3.38   0.04   0.00  99.96   0.00   0.00
   2  10   0.01 2.42 3.38   0.03   0.00  99.96   0.00   0.00
   8   1   1.23 3.64 3.38  21.87  14.85  62.05   0.00   0.00
   8   7   0.38 3.64 3.38  22.73  14.85  62.05   0.00   0.00
   9   3   0.03 3.47 3.38   0.10  18.58  81.30   0.00   0.00
   9   9   0.03 3.42 3.38   0.09  18.58  81.30   0.00   0.00
  10   5   0.01 3.15 3.38   0.11   0.00  99.89   0.00   0.00
  10  11   0.04 3.35 3.38   0.07   0.00  99.89   0.00   0.00
10.701063 sec
```

Viewing C-states with turbostat

```
# ./turbostat
 cor CPU    %c0   GHz  TSC    %c1    %c3    %c6   %pc3   %pc6
             0.04 1.62 3.38   0.12   0.00  99.84   0.00  71.22
   0   0    0.05 1.62 3.38   0.07   0.00  99.88   0.00  71.22
   0   6    0.02 1.62 3.38   0.10   0.00  99.88   0.00  71.22
   1   2    0.01 1.63 3.38   0.07   0.00  99.92   0.00  71.22
   1   8    0.01 1.61 3.38   0.06   0.00  99.92   0.00  71.22
   2   4    0.03 1.62 3.38   0.06   0.00  99.91   0.00  71.22
   2  10    0.03 1.62 3.38   0.06   0.00  99.91   0.00  71.22
   8   1    0.01 1.62 3.38   0.04   0.00  99.95   0.00  71.22
   8   7    0.02 1.62 3.38   0.03   0.00  99.95   0.00  71.22
   9   3    0.09 1.62 3.38   0.33   0.00  99.58   0.00  71.22
   9   9    0.03 1.62 3.38   0.39   0.00  99.58   0.00  71.22
  10   5    0.04 1.62 3.38   0.15   0.00  99.81   0.00  71.22
  10  11    0.06 1.62 3.38   0.13   0.00  99.81   0.00  71.22
```

Linux cpuidle sub-system



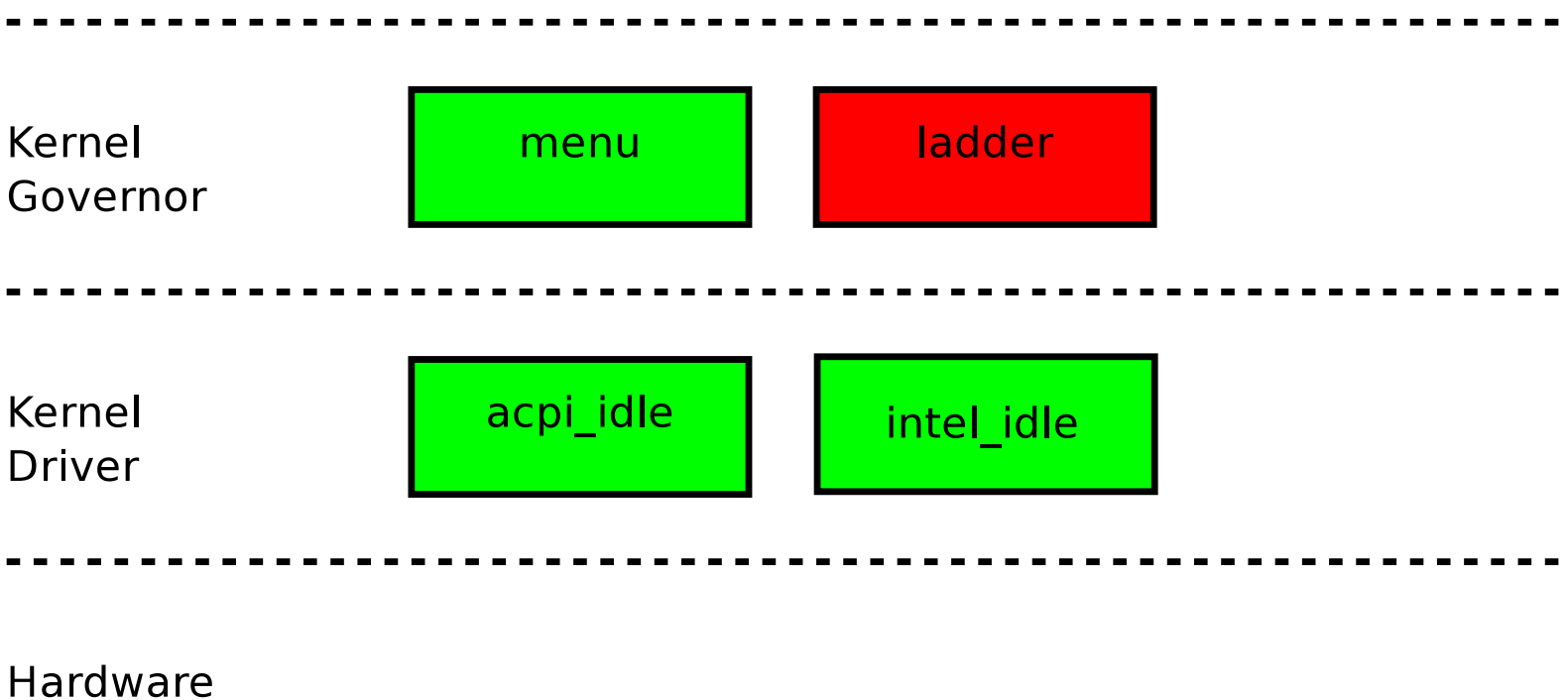
Issues with acpi_idle

- BIOS blunders
- Inaccurate C-state latencies
- No concept of C-state energy break-even
  - acpi.latency_factor=2 (2.6.25) - was 6
- No concept of core vs package C-states

Catalyst for developing intel_idle

- At Intel Core i7 Processor Introduction
  - OEM ACPI BIOS bugs increased idle power to 100W from 85W
  - OEM ACPI BIOS bugs decreased max to clock 2.9 from 3.2 GHz

Linux cpuidle sub-system w/ intel_idle



intel_idle upstream in Linux-2.6.35

```
intel_idle: native hardware cpuidle driver for Intel processors

This EXPERIMENTAL driver supersedes acpi_idle on
Intel Atom Processors, Intel Core i3/i5/i7 Processors
and associated Intel Xeon processors.

It does not support the Intel Core2 processor or earlier.

For kernels configured with ACPI, CONFIG_INTEL_IDLE=y
allows intel_idle to probe before the ACPI processor driver.
Booting with "intel_idle.max_cstate=0" disables intel_idle
and the system will fall back on ACPI's "acpi_idle".

Typical Linux distributions load ACPI processor module early,
making CONFIG_INTEL_IDLE=m not easily useful on ACPI platforms.

intel_idle probes all processors at module_init time.
Processors that are hot-added later will be limited
to using C1 in idle.

Signed-off-by: Len Brown <len.brown@intel.com>
```

intel_idle results

On offending system...

- 100 Watts reduced to 85 Watts
- Max frequency increased to 3.2GHz from 2.9 GHz

- Though after ~6 months, OEM did ship fixed BIOS...

Issues with intel_idle

- OS must manage platform/device latency limitations
  - No ACPI BM_STS bit...
  - Exactly what PM_QOS is for...
- OS must manage AC/DC C-state policy (if any)
  - No ACPI _CST re-evaluation upon AC/DC transition
  - PM_QOS useful here?
- OS must handle any hardware/platform bugs
  - No bug workarounds via ACPI BIOS update

More Issues with intel_idle

- intel_idle must be taught about new HW
  - if intel_idle does not load, acpi_idle loads
- Table proliferation
  - non-issue b/c small and much sharing

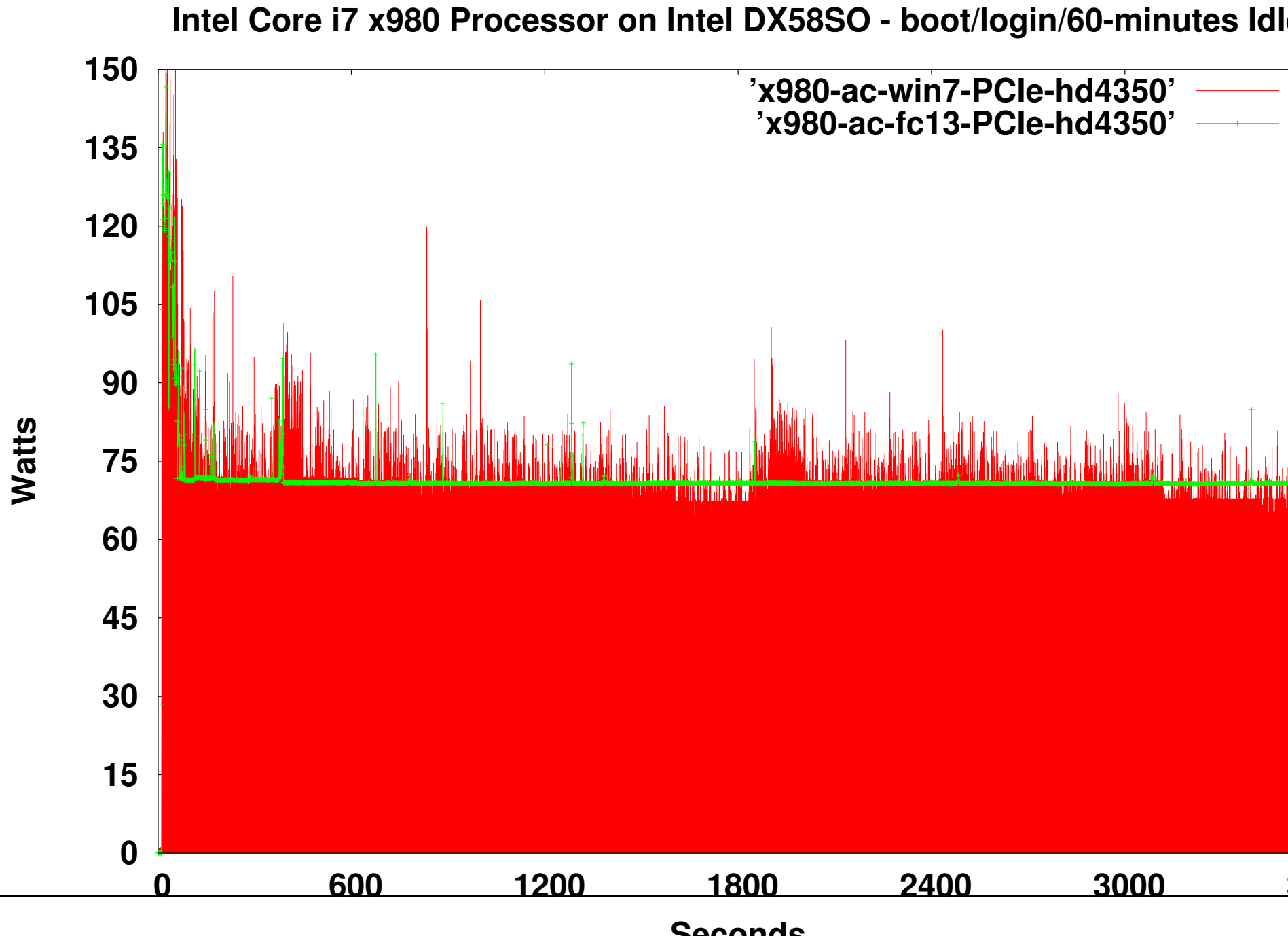
intel_idle plans

- CONFIG_EXPERIMENTAL=n in 2.6.36
- tuning
- make hot-plug friendly
- expose previously unavailable states
  - CPU-centric
  - platform-wide

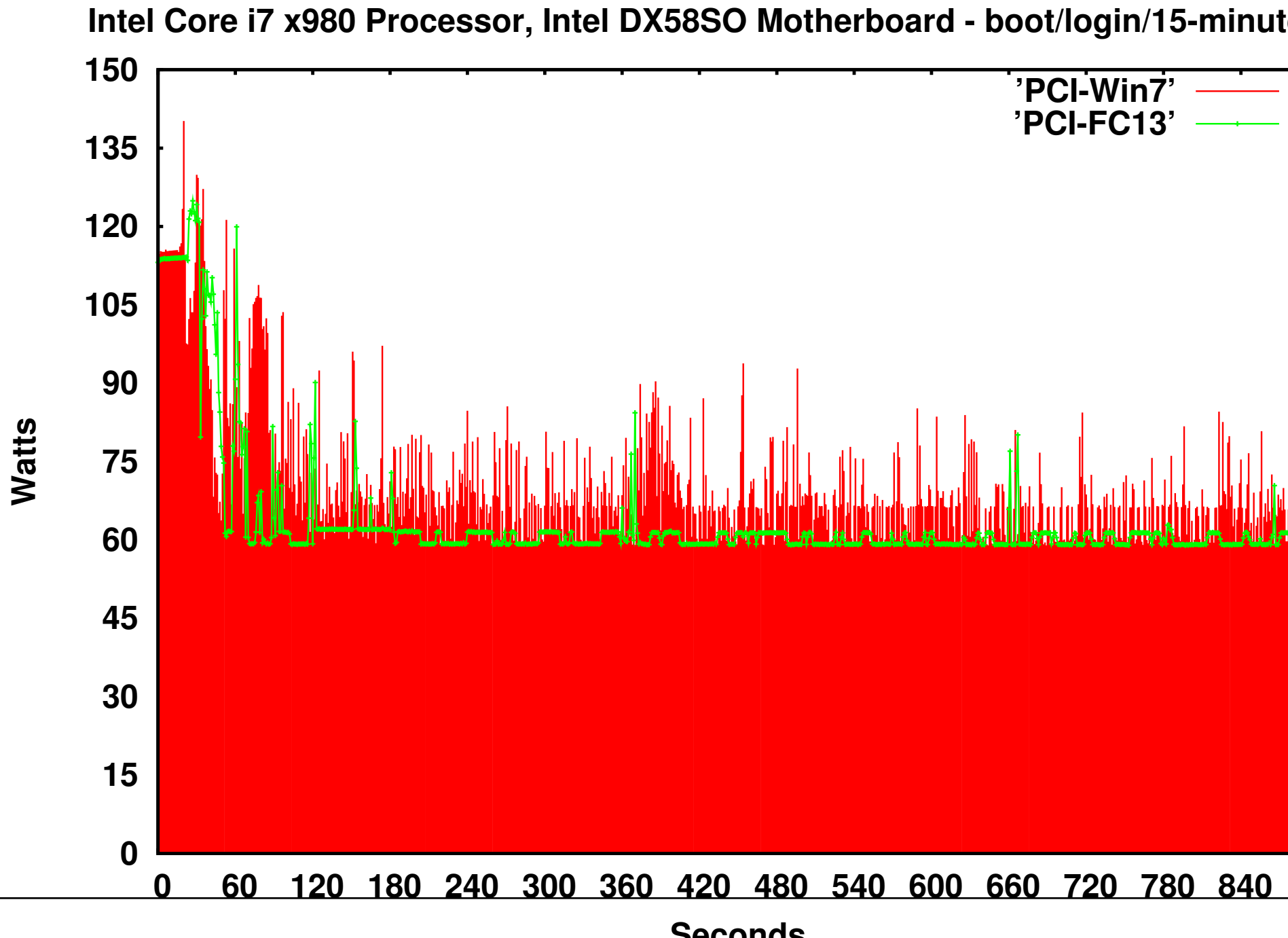
Idle Measurements

- Methodology: "as shipped"
- Configuration:
  - Wired-Ethernet live
  - WiFi, BT disabled
  - DT w/ USB KBD/mouse, monitor not included
  - Notebook display bright
- Cold Power-on, log-in, walk away
- AC Power meter: Yokogawa WT210, 1-sec integral

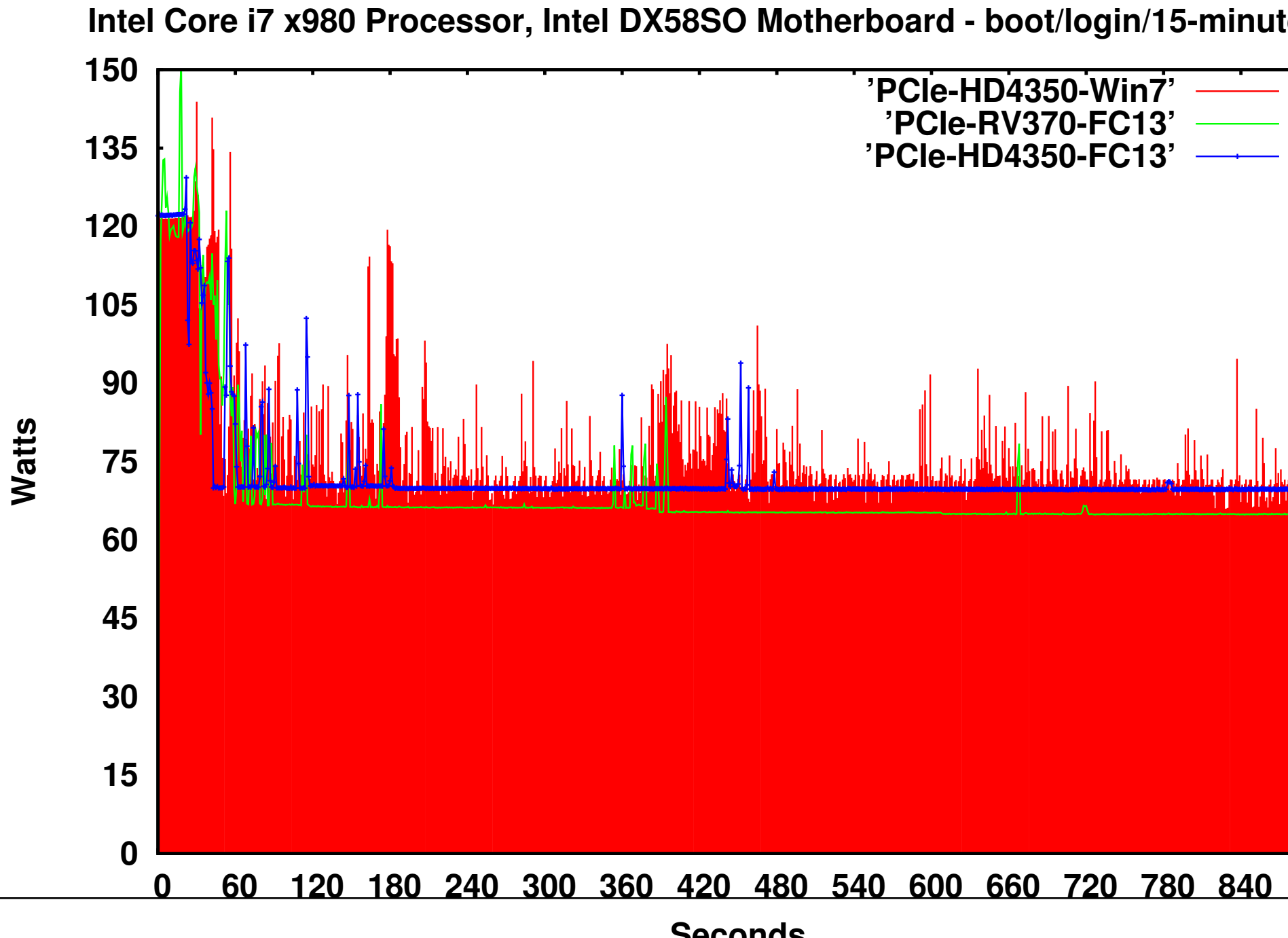
Desktop - FC13 v. Win7



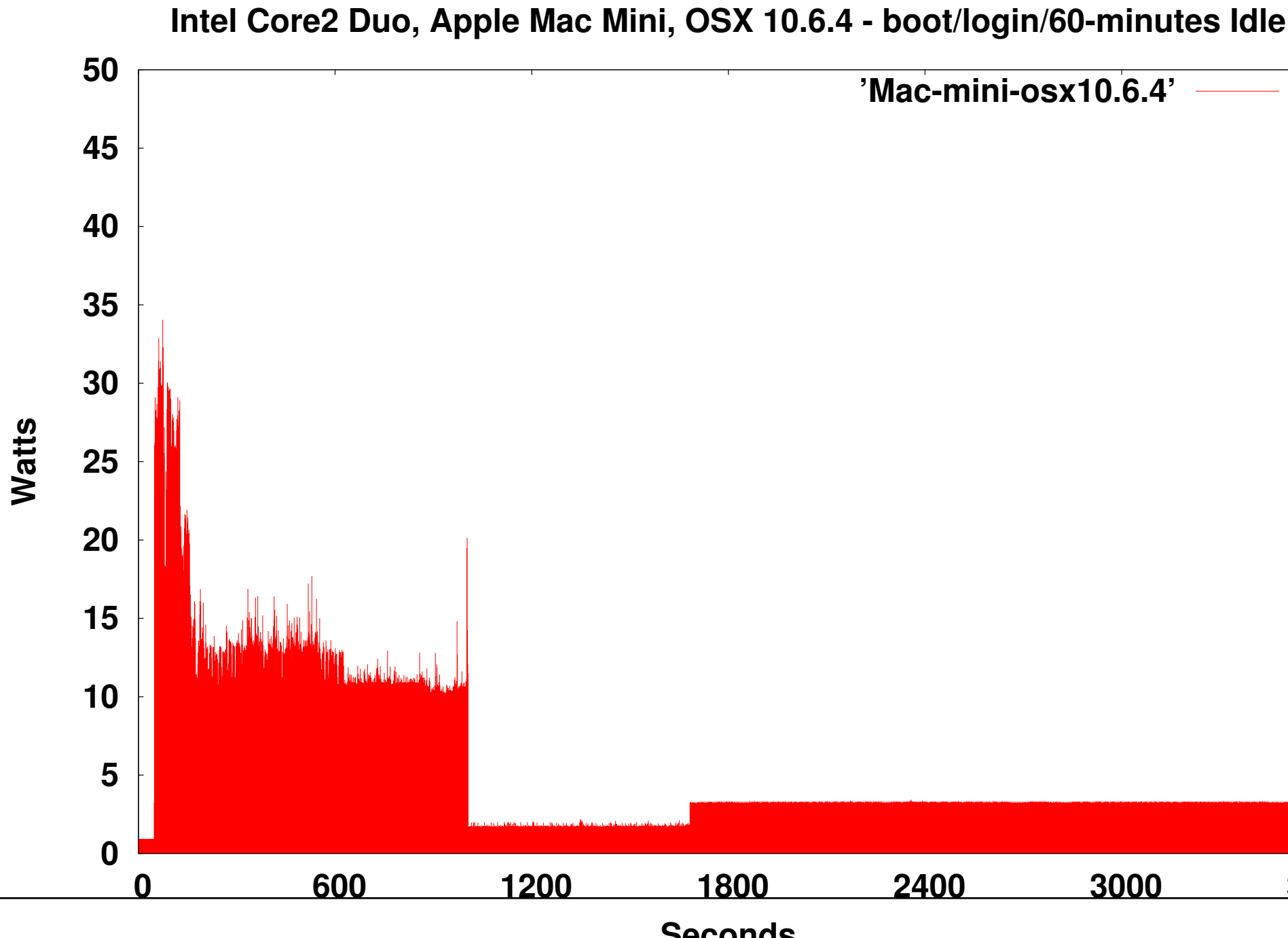
Desktop w/ PCI graphics - FC13 v. Win7



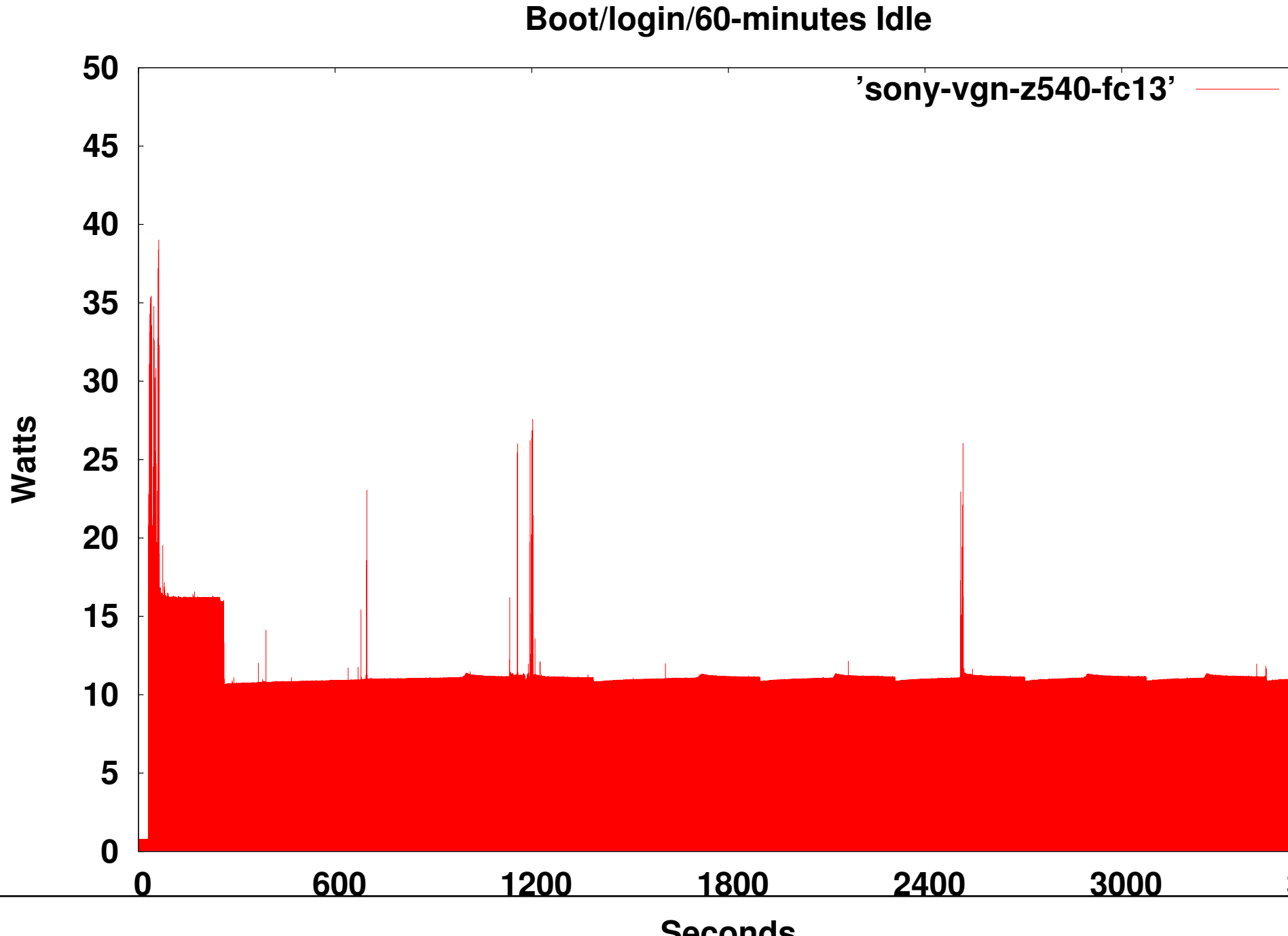
Desktop w/ PCIe graphics - FC13 v. Win7



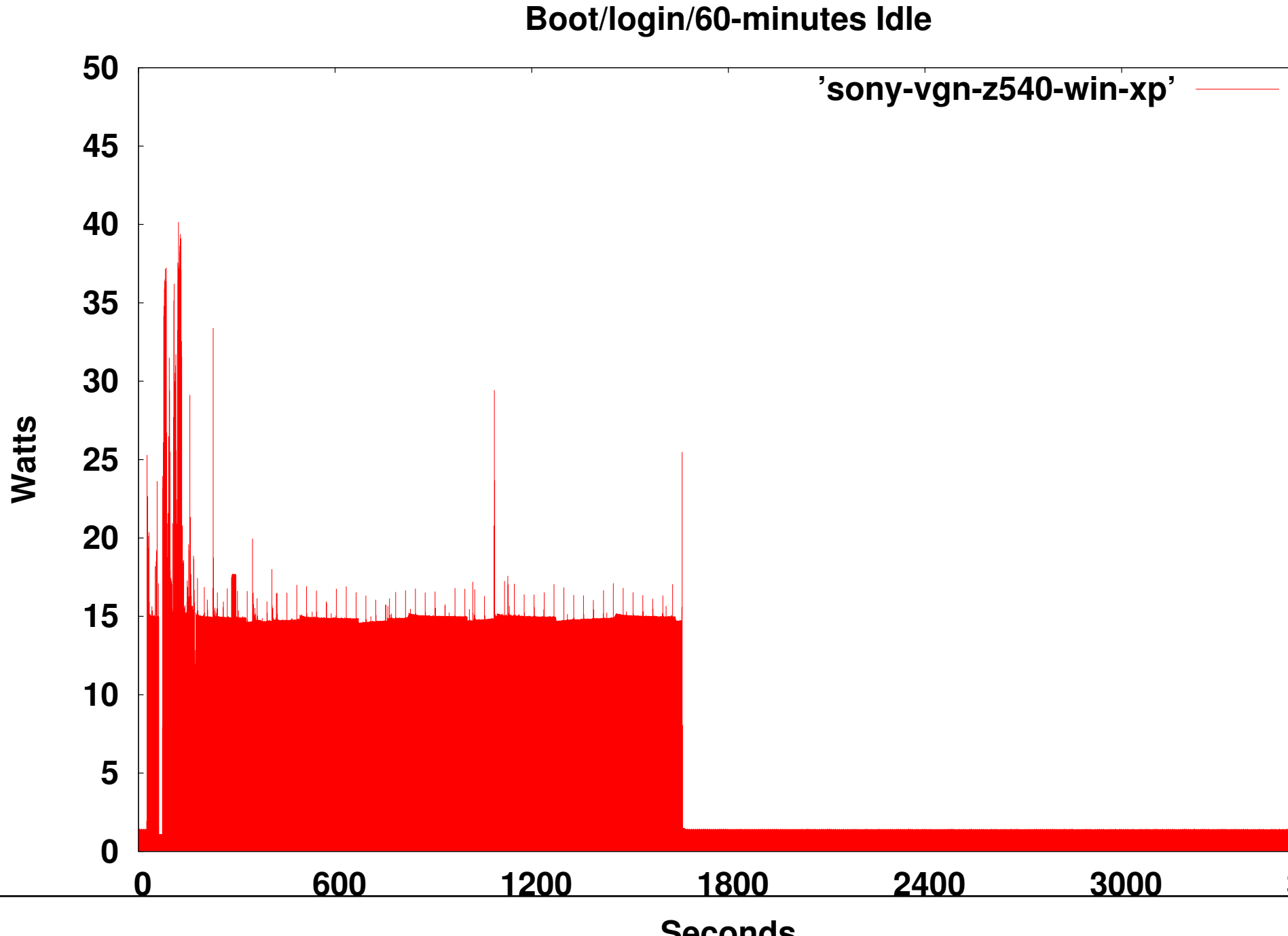
Apple Mac mini 3,1 - Snow Leopard



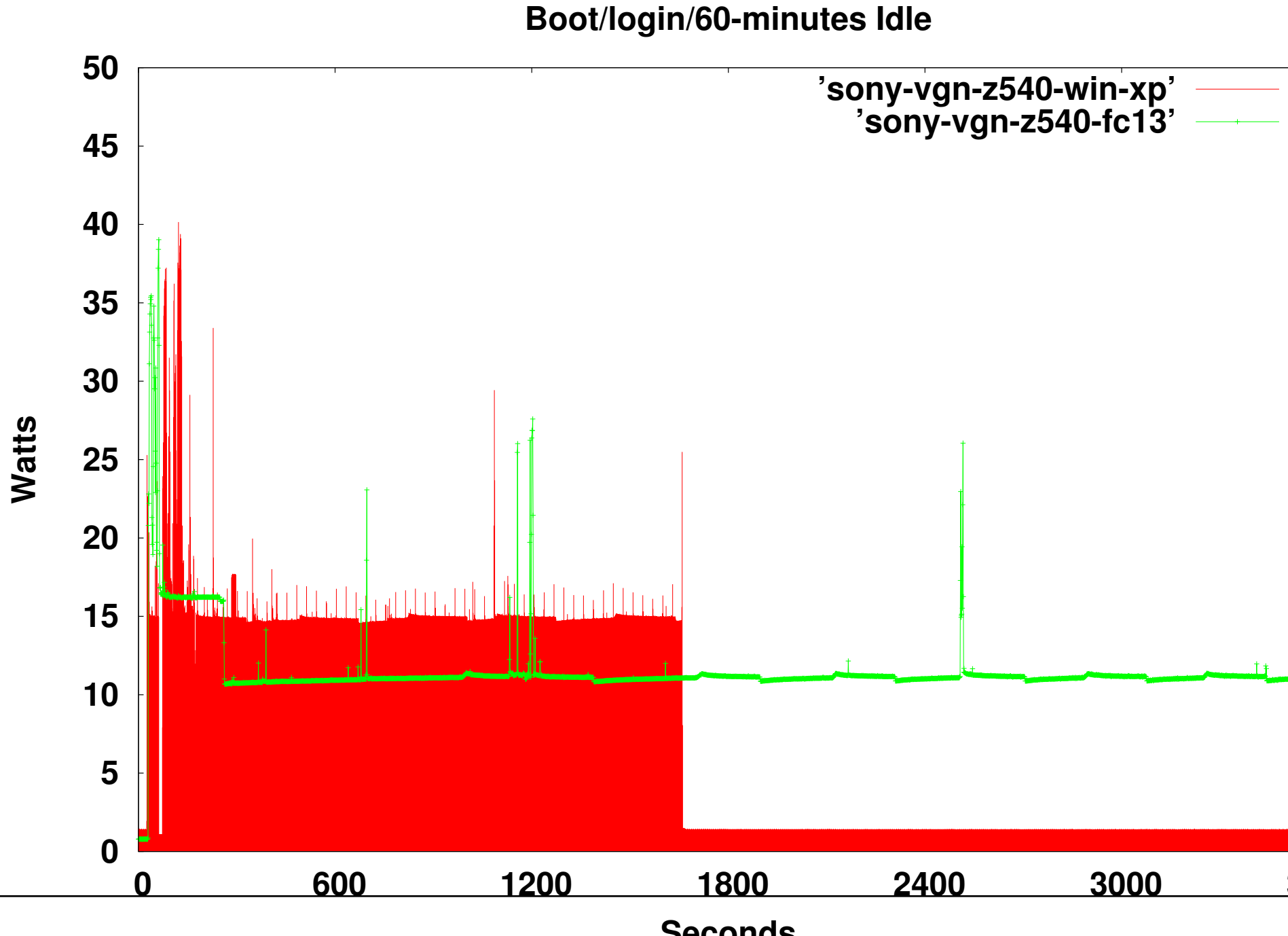
Sony Vaio Notebook - Fedora 13



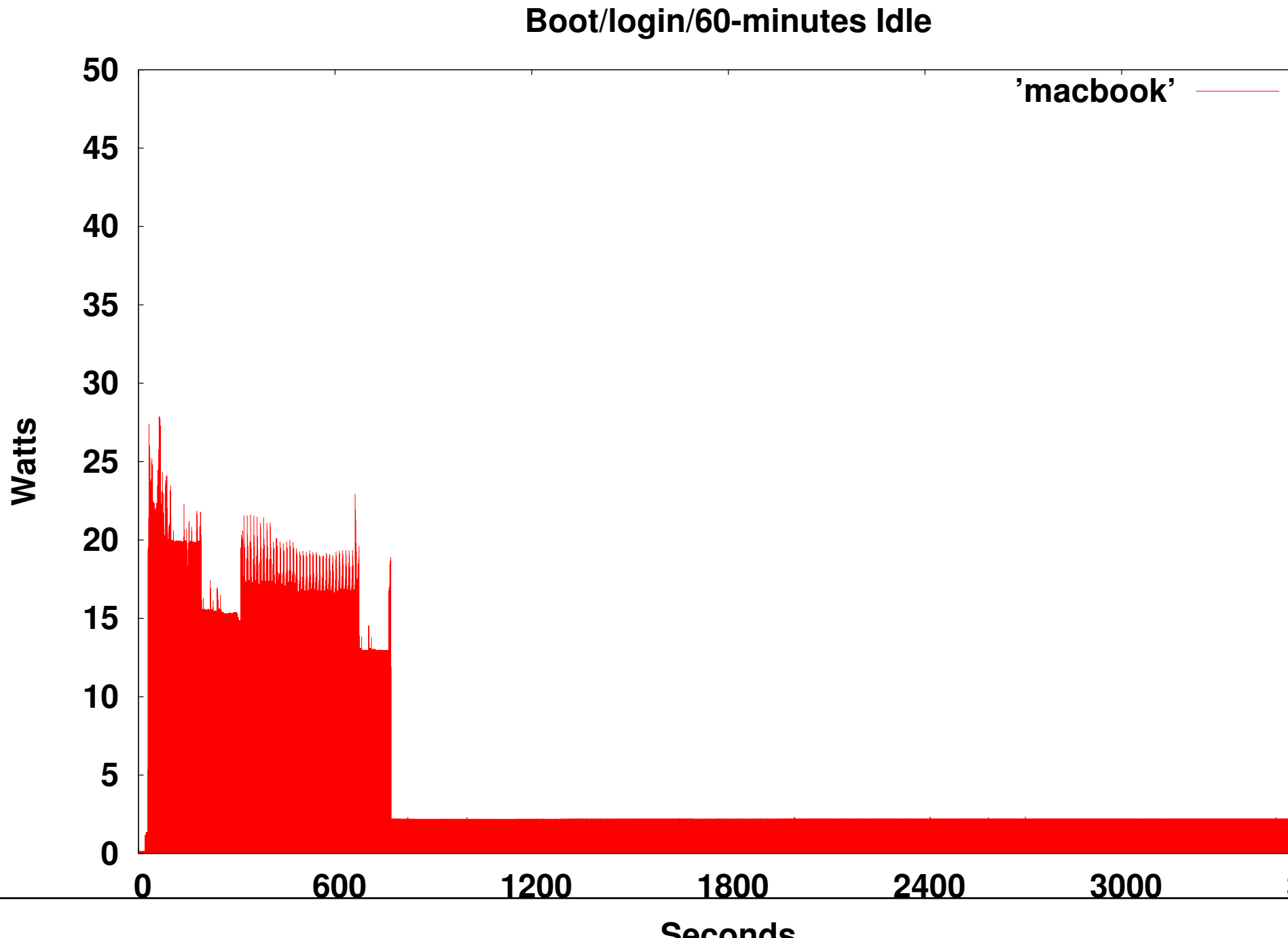
Sony Vaio Notebook - Windows XP



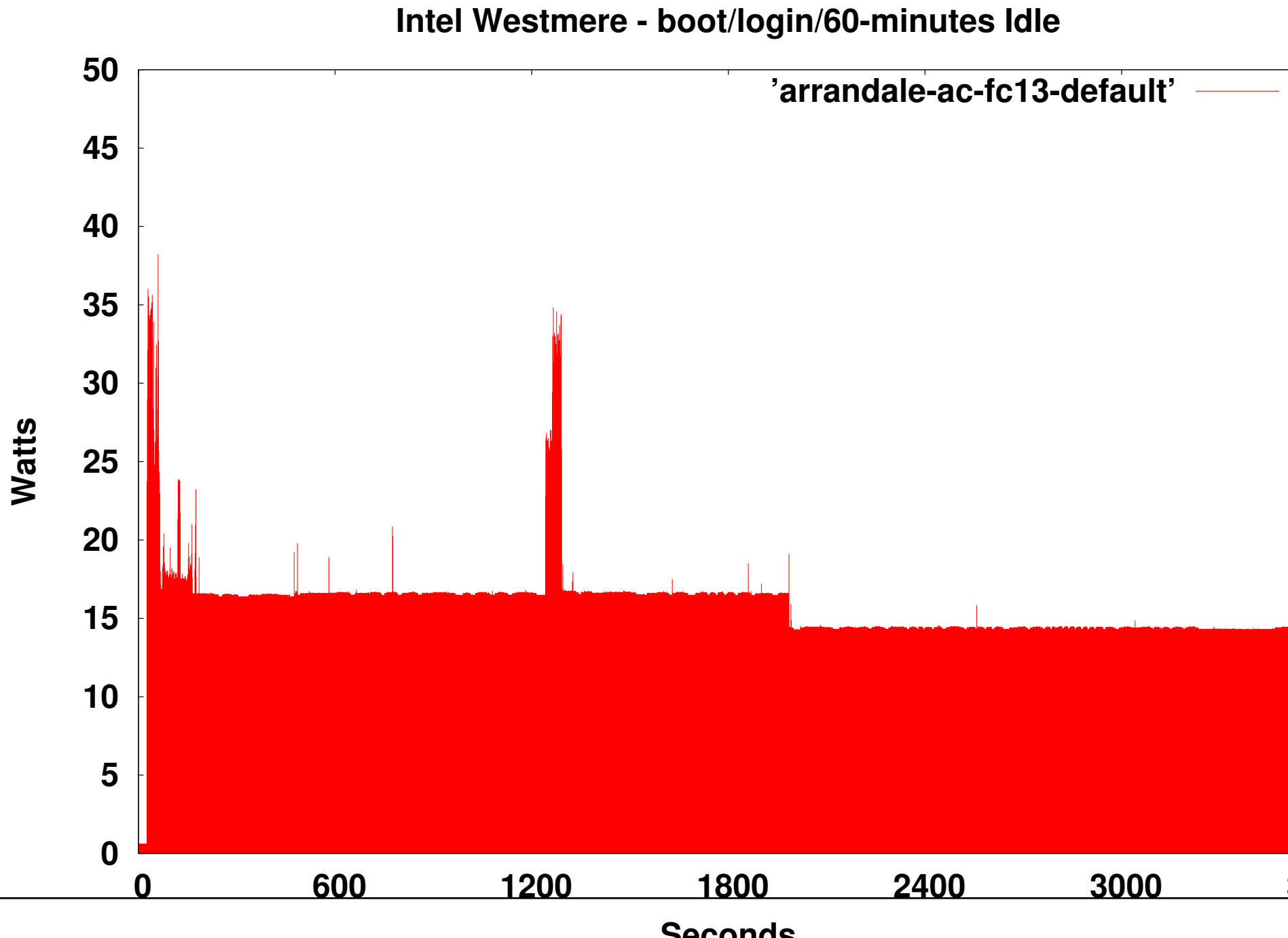
Sony Vaio Notebook - WinXP v. FC13



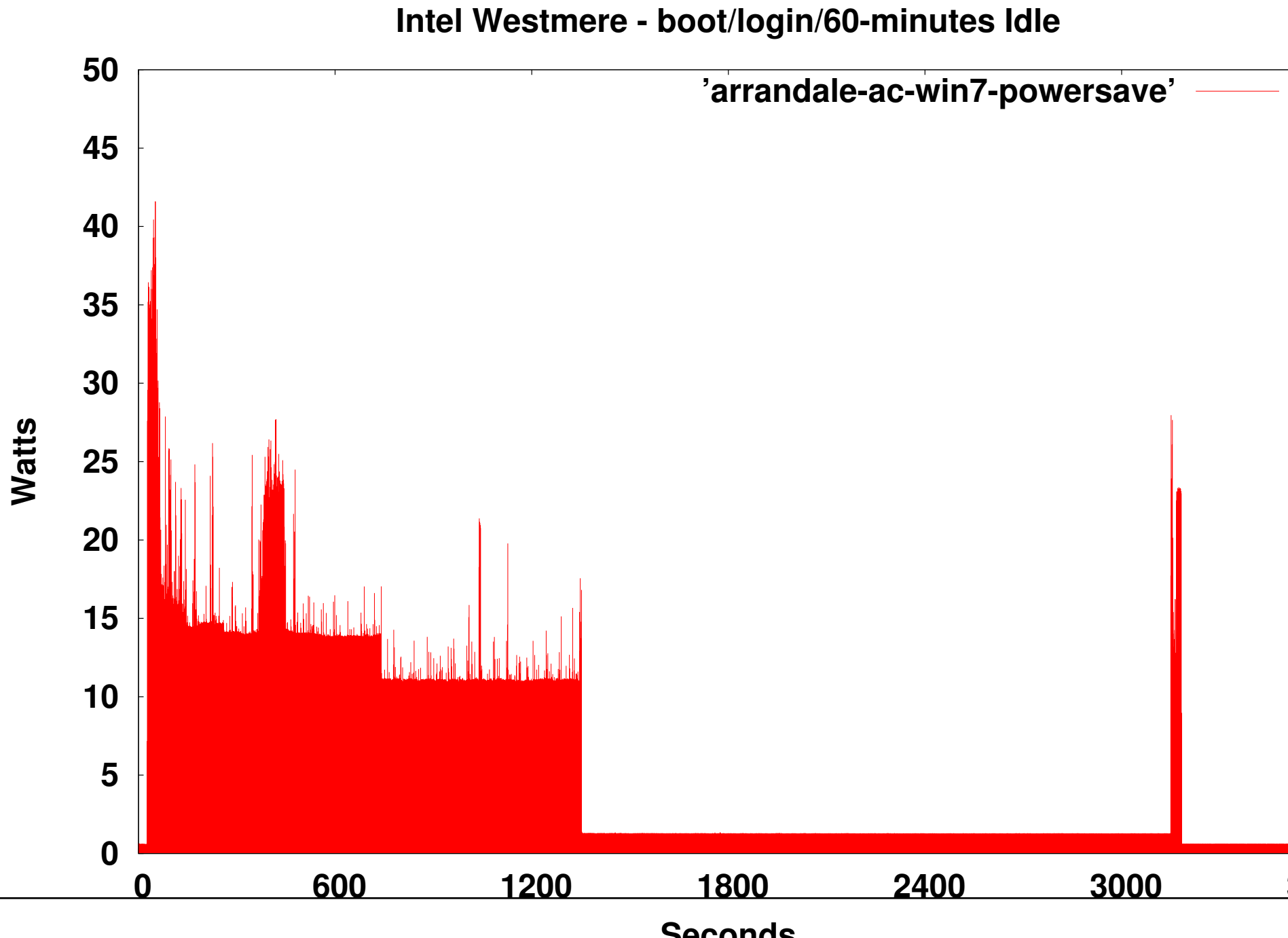
Apple Macbook - Leopard (circa 2006)



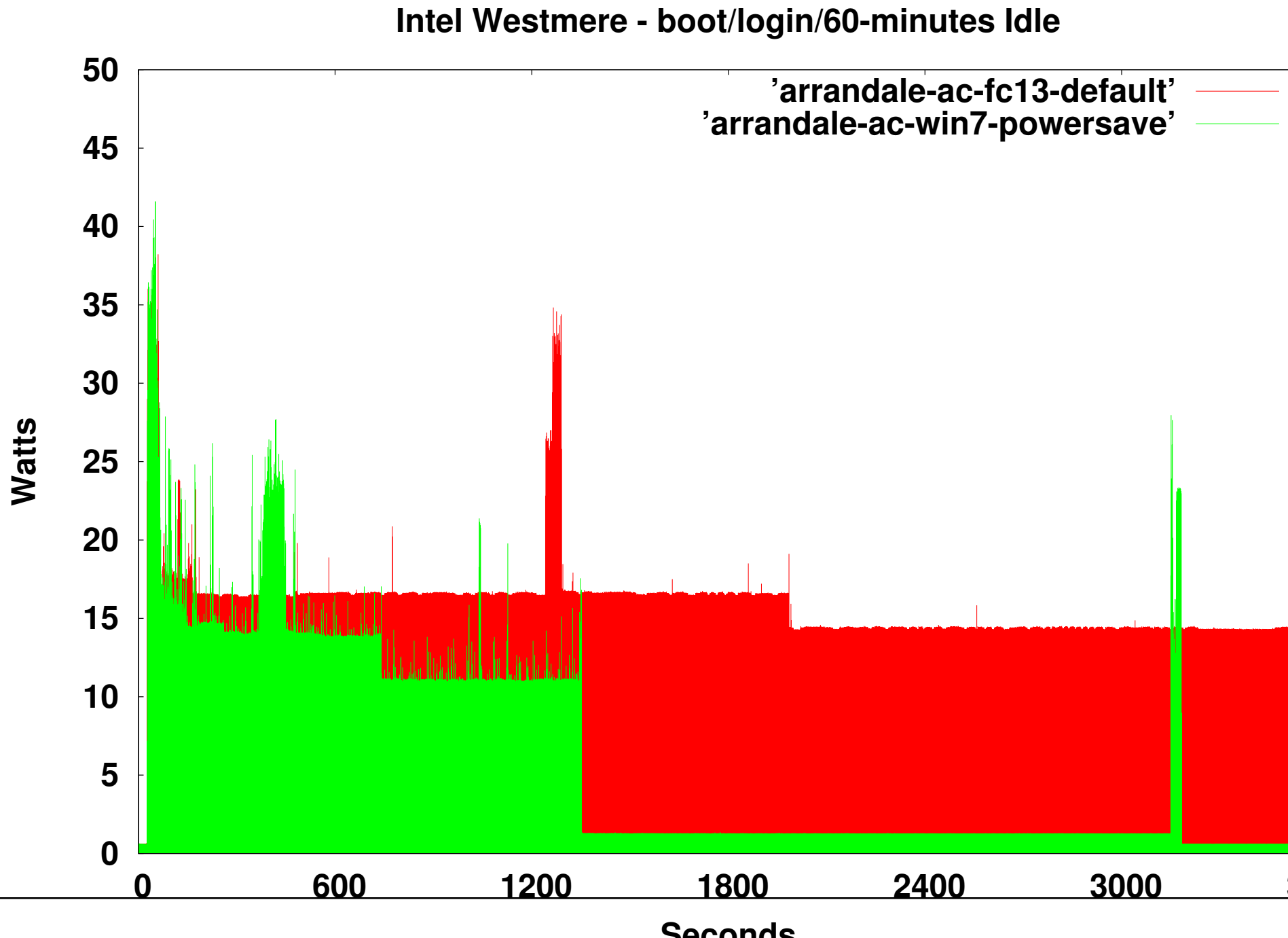
Notebook - Fedora 13



Notebook - Windows 7



Notebook - FC13 v. Win7



Summary

- Aggressive suspend is a game-changer
  - Apple benefits from system integration
- Linux is competitive on Desktop Idle Power
  - Unless competition suspends...
- Linux trails on notebook Idle Power
  - Both Core2 and Core i7 generations

References

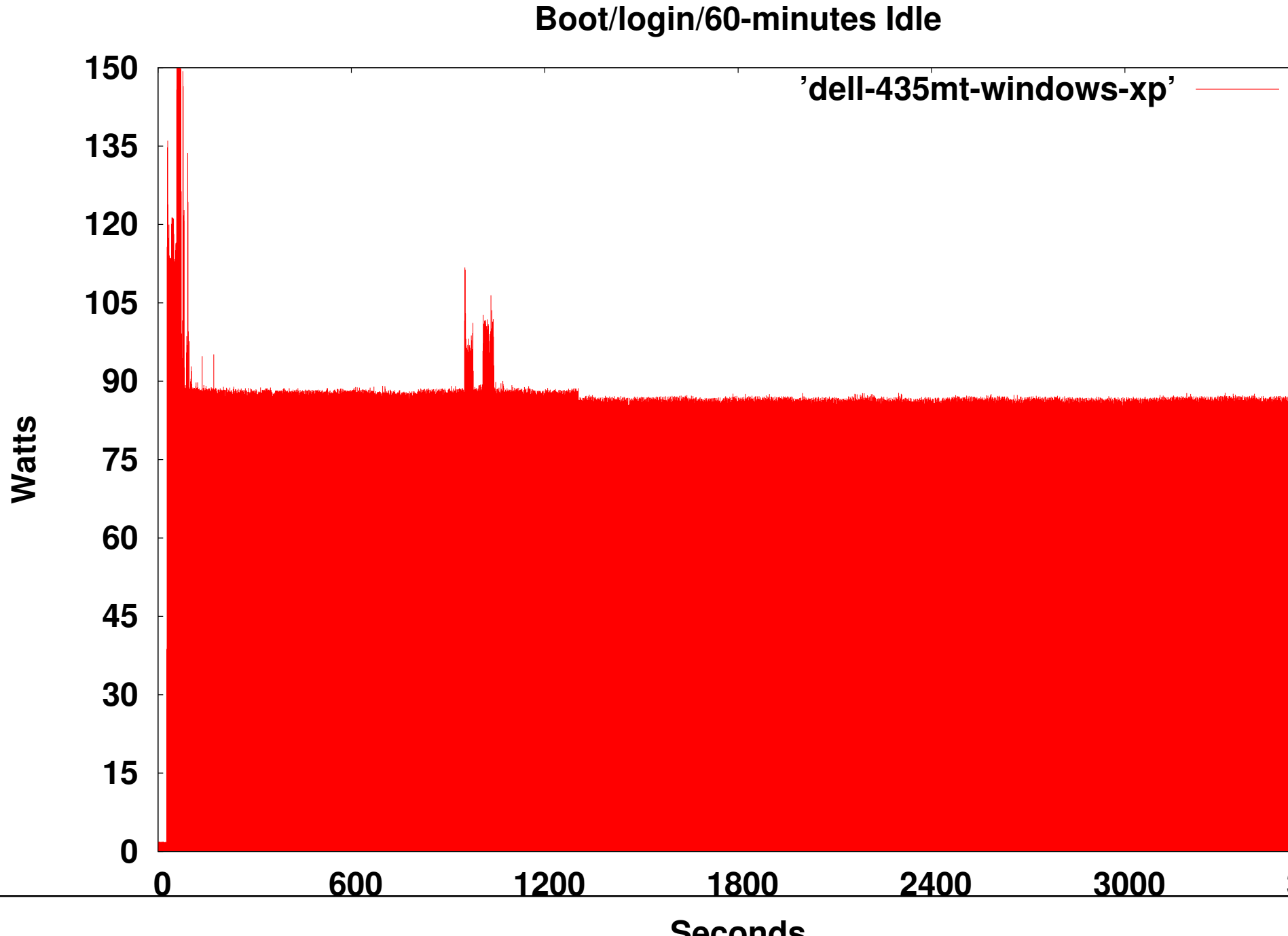
cpuidle:

http://lwn.net/Articles/384146 (April 26, 2010)
http://www.kernel.org/doc/ols/2007/ols2007v2-pages-119-126.pdf

turbostat:

http://www.kernel.org/pub/linux/kernel/people/lenb/acpi/utils/pmtools

Dell 435MT Desktop - Windows XP



Dell 435MT Desktop - Fedora 13