the green grid™
get connected to efficient IT

FIVE WAYS TO REDUCE DATA CENTER SERVER POWER CONSUMPTION

EDITOR:

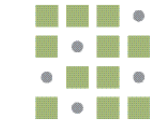
MARK BLACKBURN, 1E

 Rev 2008-0

ABSTRACT

A significant reduction in energy usage can be made by moving away from a mindset that all servers have to be powered on at all times regardless of use, to one where a service is always available regardless of server state. With use of the commonly available tools, server energy consumption can be reduced without impacting ongoing operations, capital budgets or system reliability.

TABLE OF CONTENTS



INTRODUCTION

This document only addresses changes that can be made at the server1 level. Other white papers from The Green Grid will address power, cooling, airflow, consolidation, virtualization and a host of other mechanisms to increase efficiency elsewhere in the data center. Reducing energy use at the point of consumption (the server) provides benefits at all other levels by reducing load on power and cooling facilities which in turn reduces their own energy use.

The bulk of installed servers in data centers today consists of x86 commodity servers. These servers consume much2 of the power allocated to IT server equipment. Therefore, the x86 servers present the largest opportunity for saving power in the data center. A significant reduction in energy usage can be realized if data center professionals move away from a mindset that all servers need to be powered on at all times.

The conventional wisdom is that servers must be kept running 24x7x52 because restarting them poses a potential downtime risk. However, research data suggests that this perception is false. Mean Time Between Failure (MTBF) statistics for components are now measured in hundreds of thousands to millions of hours.

In a series of 3 laboratory tests over a 5 month period, a total of 123 servers were restarted several times daily by disconnecting and reconnecting the power utilizing an automated power strip outlet. Out of 18,826 restarts, not a single component failure occurred.

By utilizing scripting and systems management tools such as Wake-on-LAN capabilities, most organizations can implement key energy saving processes, without impacting ongoing operations, capital budgets or system reliability.

Listed below are five key recommendations that will allow data center professionals to reduce their overall data center energy consumption by making changes at the server level.

1. IDENTIFY THE CULPRITS

In order to understand the impact on energy consumption of implementing new practices, it is necessary to identify and document all running servers within the data center, determine their business purpose and measure their power consumption. Organizations do not currently measure power consumption on a per server basis. However, it is possible to generate estimates without too much difficulty.

The latest generation of servers feature built-in power monitoring via their out-of-band management capabilities. However, the vast majority of currently installed (older) servers do not have this ability. Therefore, this cannot be the only measurement method used.

It is possible to instrument the power delivery infrastructure (e.g. ‘smart’ power strips) which can monitor power usage for each server in real time and provide accurate power usage statistics. Be aware, however, that this will require investment in new hardware, will impact operations during installation, and will add overhead when implementing and monitoring the solution.

A low cost, low disruption method bases power usage calculations on a server’s CPU utilization. A study4 which included a comparison of different power consumption based on thousands of servers with differing workloads concluded that power consumption tracks very closely with CPU utilization. This single metric therefore can be used as a relatively accurate estimate of power consumption.

Internal disks spin and draw power all the time, the only additional power they draw when being accessed is

to move the read/write head. The dynamic differential between idle and fully utilized is only around 30% of the disks power draw, and as a fraction of overall system power that is negligible.

Memory is constantly being refreshed and drawing power regardless of it being read or written to - the change in memory power draw with use is also not significant when taken as a fraction of overall system power use.

Most I/O and memory use also comes with some CPU activity, since the CPU is used to manage and monitor the progress of the task, and as such disk and memory use correlates to CPU utilization.

the green grid™

The CPU varies dramatically in its power draw, since the architecture has been optimized to enable large parts of the silicon to shut down when in idle states - as such it is unique in being the only component of the system that has a marked effect on system level power draw based on its utilization.

Figure 1 illustrates a model where server power consumption scales linearly with CPU utilization

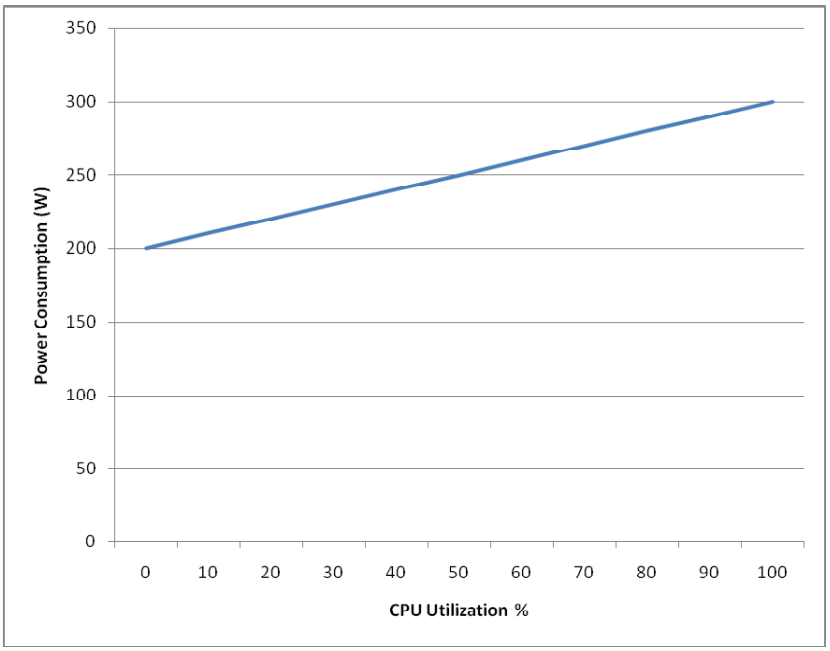


Figure 1: CPU Utilization to Power Consumption

Most servers are already collecting CPU utilization information via systems management software; however few organizations make use of this data other than for capacity planning. By taking average CPU utilization over a defined period of time, it is possible to calculate an estimate of the power consumed for that period.

Since our model scales linearly from idle to maximum utilization, once we know the power draw of a server at peak usage and at idle it becomes a simple arithmetic operation to estimate power usage at any utilization rate.

Until recently the only power figure published for servers was the rating of the power supply, which is typically much higher than the actual power consumed. However, an increasing number of server manufacturers are now publishing actual power utilization figures for current models at idle and at maximum CPU utilization. This is being driven by the adoption of the ASHRAE Thermal Guideline5 or similar manufacturers report which provides power ratings for minimum, typical, and full configuration. The ASHRAE Guideline was published by ASHRAE's TC 9.9 and was developed by the majority of server manufacturers and other stakeholder ASHRAE

members.

Most organizations standardize on server specifications. Therefore it is likely that a limited number of differing server models exist at any particular site. Therefore, measuring the power consumption of a single server of each type at full load and at idle will not be complicated or time consuming and will provide sufficient accuracy to make informed decisions.

Once these figures are available, an estimate of power consumption (P) at any specific CPU utilization (n%) can be calculated using the following formula:

Pn = (Pmax – Pidle) * n/100 + Pidle

Example:
If a server has a maximum power draw of 300 Watts (W) and an idle power draw of 200W, then at 5% utilization the power draw would approximate to:

Power Utilization at 5%
= (300 - 200) * 5/100 + 200
=100 * 0.05 + 200
=205W

If the server was running at that average utilization for a 24 hour period, then the energy usage would equate to the following:

205W * 24 = 4920 Watt hour (Wh) = 4.92 kilowatt hour (kWh)

Through empirical measurement of various servers using a power analyzer6 this approximation has proven to be accurate to within ±5% across all CPU utilization rates.

A baseline of current power usage throughout the data center can be created by adding up the power usage for all the servers in the data center. This data can then inform later decisions regarding which changes will have the most positive impact on overall server power usage.

2. ENABLE SERVER PROCESSOR POWER SAVING FEATURES

In recent years, x86 server processors have begun to incorporate the power saving architectures that have been common in both desktop and laptop computers. Enabling this feature can result in overall system power savings of up to 20%.

The power saving is achieved by reducing the frequency multiplier (Frequency identifier or FID) and the voltage (Voltage identifier or VID) of the CPU. Intel's version of this technology is known as either Enhanced Intel SpeedStep Technology (EIST) or Demand Based Switching (DBS) and AMD's version is marketed under the name Cool'n'Quiet or PowerNOW! The combination of a specific CPU frequency and voltage is known as a performance state (p-state). Moving forward, this paper will utilize the term 'p-state control' to refer to the capability to reduce frequency and voltage.

Altering the p-state can reduce a server's power consumption when at low utilization but can still provide the same peak level of performance when required. The switch between p-states is dynamically controlled by the operating system and occurs in micro-seconds, causing no perceptible performance degradation.

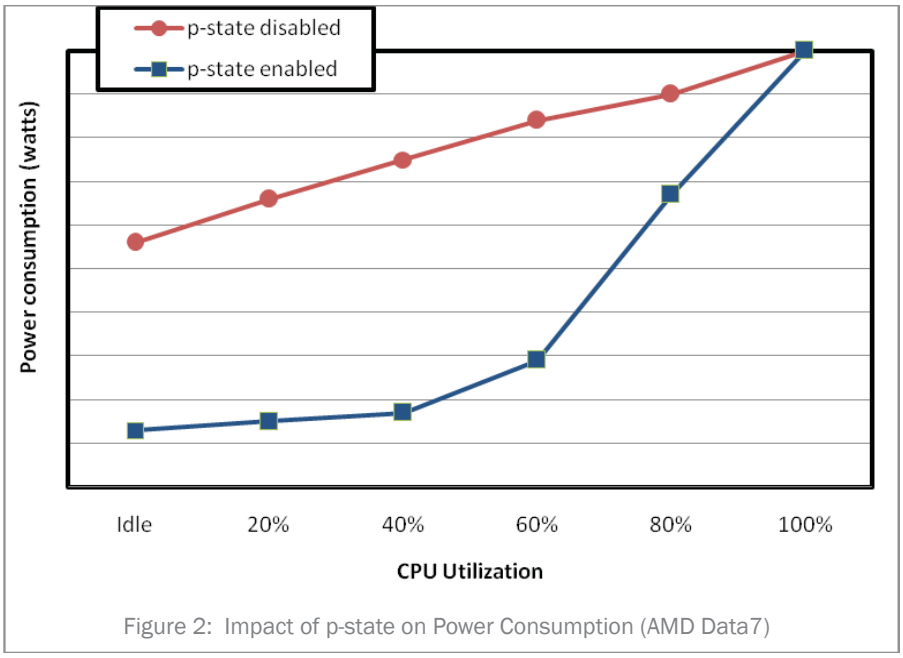


Figure 2: Impact of p-state on Power Consumption (AMD Data7)

Although a processor may be p-state capable, both the system Basic Input-Output System (BIOS) and the operating system must be capable of enabling the feature to make use of it. Check the BIOS of a representative sample of each server model to find out if the server supports the relevant version of the p-state technology.

Further information regarding the impact of p-state control on system power consumption can be found in documentation from both AMD8 and Intel9.

Instructions on how to implement p-states on the three main x86 commodity server operating systems can be found in Appendix A.

3. RIGHT-SIZE SERVER FARMS

In recent years, Web services have driven the growth of server farms in data centers. In many cases these server farms are vastly over-provisioned. The analysis of server farm usage patterns will reveal the potential for 'right-sizing'. Unneeded capacity can be turned off, but the server farm can still provide sufficient resiliency for agreed upon service levels.

Data center owners should perform analysis of server utilization data (CPU, disk and network) across all servers in a server farm. Average utilization across the farm is likely to follow a daily, weekly and/or monthly pattern.

If enough utilization data has been collected to demonstrate the trends over time, an informed decision can be made regarding how many actual servers are required to provide peak service levels plus resilience. It is likely that this number is lower than the actual number of servers in the farm, meaning that the surplus capacity can be powered down to conserve energy.

For example, if a server farm consisting of 10 servers has a maximum utilization (max of CPU, disk or network utilization per server) across the farm of 50% this is an aggregate utilization of 500%, which equates to 5 servers running at 100%. To provide sufficient headroom and still allow for resilience, the farm could easily run with 7 servers (peak utilization of 500/7 = 71%). Under this scenario, if one server failed sufficient

capacity would still exist (with 6 servers peak utilization would be at 83%) and 3 warm standby servers would still be available to rapidly recover the availability levels should one of the active servers fail. In this example, a power saving equivalent of up to 3 servers is possible for this farm.

It is possible to automate the restart of servers by using either built-in out-of-band power management capabilities or Wake-on-LAN tools. Out-of-band management capabilities can be controlled via vendor specific software or through standard SNMP methods.

4. POWER DOWN SERVERS WHEN NOT IN USE

Not all servers need to be operational 24x7x52. Individual servers may be powered down for certain periods of the day. For example, servers executing backup software are normally only required at night and branch-based servers are normally only used during the day.

Certain types of servers will regularly go unused for random, lengthy periods of time. These should be targeted for powering down. Typical examples are servers found in test and development environments. The test team will know when a test run has finished. These particular test machines should then be powered down until they are needed. In addition, development build systems should be powered down until a build run is required.

CPU utilization statistics will show that certain machines have a consistently low (typically an idle server will run at <1%) CPU utilization for large periods of time. Analysis of server utilization over time will normally reveal a pattern to when the servers are busy. These machines could be scheduled to power down for the periods of time that they are idle and then powered up in time to perform their useful work.

For example, a server executing backup software which is only busy from 10PM until 6AM could be scheduled to power itself down at 8AM every day (checking beforehand to ensure that backups had completed for the day) and then be powered up by operations management tools or a job scheduling system at 9PM ready to perform the next night's backups. If the server were required for a restore during the day, the operator could run a script that would power the machine back up, run the restore and then power the machine back down.

5. DECOMMISSION OLD SYSTEMS THAT PROVIDE NO USEFUL WORK

Anecdotal evidence suggests that a significant number of installed servers are not used at all by anyone. These are older servers that have fallen out of use but have not been decommissioned. No one has tracked whether anyone still uses them or not. These machines can be identified by analyzing their use (or lack thereof).

Servers of this type will usually have very low utilization rates all the time, with only the occasional spikes of utilization when standard housekeeping tasks run (backups, virus scans, etc.). The machines are however performing no useful purpose and are just consuming power and heating the data center for no good reason.

Once a machine has been identified as "unused" it is possible to confirm this status by analyzing network statistics. This exercise will ensure that all connections to the machine in question are from management systems and not from other business systems or from end users. If end users are indeed linked to the server in question, these end users should be contacted to determine how the server is providing useful work. It is highly likely that the connections are merely legacy in nature and can be terminated.

Once the server has been categorically confirmed as unused it can either be decommissioned, or turned off and put aside as stock ready for deployment should users develop a relevant requirement.

Keeping a legacy server around simply because it is available may be poor efficiency practice. New servers available today offer better performance with significantly reduced energy demands. If the decision is made to retire legacy servers, they should be processed for recycling and/or repurposing. Most server manufactures have global recycling programs available. In addition there are numerous third-party groups who embrace environmentally benign practices in the recycling of e-waste10.

SUMMARY

It is not necessary to invest in large scale hardware refresh programs or consolidation exercises to start making a positive impact on energy efficiency. Identifying energy wasters, enabling power saving features, right sizing, powering down underutilized servers and decommissioning legacy servers all represent a major entitlement for energy reductions. An immediate difference can be made by adjusting the way existing servers are running and moving away from traditional thinking. A positive impact can be made on the environment, on energy consumption, and ultimately on the bottom line with minimal effort.

APPENDIX A: HOW TO ENABLE P-STATE SUPPORT IN THE MOST PREVALENT SERVER OPERATING SYSTEMS

WINDOWS OPERATING SYSTEMS

Microsoft operating systems from Windows Server 2003 onwards automatically include Intel p-state support. Add-in drivers are available from server hardware vendors that enable support in Windows 2000.

AMD p-state support requires an add-in driver11 (available from AMD) for all Microsoft operating systems. Although the relevant drivers may be installed, the correct Windows power scheme must also be selected for p-state control to be put into effect.

the green grid™

The power scheme to enable this function in Windows 2000 Server and Windows Server 2003 (up to service pack 1) is called 'Minimal Power Management'. In Windows Server 2003 service pack 2, the scheme is called 'Server Balanced Power and Performance'.

When Windows Server 2008 is released it will automatically enable p-state control where available in hardware12.

OPENSOLARIS

The Sun operating systems also include p-state support. One such tool is called Project: Tesla: OpenSolaris Enhanced Power Management. Information on this tool can be found at the following URL: http://www.opensolaris.org/os/project/tesla/
This tool features AMD/Intel CPU frequency / voltage scaling support (PowerNOW!/Speedstep), CPU throttling and support for suspend to RAM on x86-based systems.

The Solaris operating system takes advantage of p-states by default (no user configuration required). The following command will list the available p-states for the CPU on a server: $ kstat -m cpu_info -s supported_ frequencies_Hz. If more than one supported frequency is listed for each of the CPU instances, this means that Solaris supports frequency scaling of those CPUs.

CPU power management can be enabled on certain Intel systems and CPU idle threshold can be set by adding the following two entries to the power.conf(4) file: cpupm enable and cpu-threshold 15s. After the desired idle threshold has been reached, verification that the CPU(s) are running at the lowest supported frequency can be validated via the following command: $ kstat -m cpu_info -s current_clock_Hz.

LINUX

The Linux kernel, starting from version 2.6.18, also enables power saving settings. The following commands enable access to the settings:

- /sys/devices/system/cpu/ controls the Multi-core related setting
- sched_mc_power_savings
- Setting – 0 (default, optimal performance; e.g. no power saving)
- Setting – 1 (power saving enabled)
- /sys/devices/system/cpu/ controls the multi-threading setting
- sched_smt_power_savings
- Setting – 0 (default, optimal performance; e.g. no power saving)
- Setting – 1 (power saving enabled)

For more information please access the following URL: http://oss.intel.com/pdf/mclinux.pdf

REFERENCES

A 'server' in the context of this document refers to an 'enterprise server' as outlined by the EPA EnergyStar specification currently under development http://www.energystar.gov/index.cfm?c=new_specs. enterprise_servers

68% according to the EPA "Report to Congress on Server and Data Center Energy Efficiency Public Law 109 431", Page 28 http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf

Pai, Vinay. "Yes, It's Still Safe to Power Off and Power On That Server," http://vinaypai.blogspot.com/2007/09/yes-its-still-safe-to-power-off-and.html

Fan, Weber, and Barrosso. "Power Provisioning for a Warehouse-sized Computer," Pages 3-4, June 2007 http://labs.google.com/papers/power_provisioning.pdf

ASHRAE Technical Committee 9.9, "Thermal Guidelines for Data Processing Environments 2004," http://resourcecenter.ashrae.org/store/ashrae/newstore.cgi?itemid=21074&view=item&categoryid=174&categoryparent=174&page=1&loginid=595434

Voltech PM100 Single Phase Power Analyzer calibrated to NIST standards.

Power and Cooling in the Data Center, page 5 http://enterprise.amd.com/Downloads/34146D_PC_WP.pdf

Power and Cooling in the Data Center, page 7 Figure 6 http://enterprise.amd.com/Downloads/34146A_PC_WP_en.pdf

Addressing Power and Thermal Challenges in the Datacenter, Page 5 http://www.intel.com/products/services/intelsolutionservices/success/techdocs/wp/thermal.pdf

http://www.dell.com/recycle http://

www.hp.com/hpinfo/globalcitizenship/environment/recycle/index.html

http://www.sun.com/aboutsun/ehs/weee.html

http://www.ibm.com/ibm/environment

http://www.amd.com/us-en/Processors/TechnicalResources/0,,30_182_871_9033,00.html

http://www.microsoft.com/whdc/system/pnppwr/powermgmt/ProcPowerMgmt.mspx

