SERVER ENERGY MEASUREMENT PROTOCOL

Contact: Jonathan Koomey, jgkoomey@stanford.edu

Authors: Jonathan Koomey (Lawrence Berkeley National Laboratory and Stanford University), Christian Belady (Hewlett-Packard), Henry Wong (Intel), Rob Snevely (Sun Microsystems), Bruce Nordman (LBNL), Ed Hunter (Sun Microsystems), Klaus-Dieter Lange (Hewlett-Packard), Roger Tipley (Hewlett-Packard), Greg Darnell (Dell), Matthew Accapadi (IBM), Peter Rumsey (Rumsey Engineers), Brent Kelley (AMD), Bill Tschudi (LBNL), David Moss (Dell), Richard Greco (California Data Center Design Group), and Kenneth Brill (Uptime Institute)

November 3, 2006

Version 1.0

According to a Ziff-Davis survey of almost 1200 key players in the data center industry,[1] more than two-thirds (71%) of the data center operators/owners identify information technology (IT) power, cooling, or both as the primary issue facing their data center, but 62% of the respondents reported that neither IT power consumption nor cooling concerns had affected server purchases in the preceding twelve months. One important reason for this lack of response to a broadly perceived problem is that no standard method exists for manufacturers to measure and report server efficiency.

Each manufacturer has its own way of measuring energy use associated with different performance levels (or doesn't measure it at all), so that server buyers are unable to easily compare computing and energy performance. This document should help resolve that problem until a formal industry standards process produces more permanent and detailed guidelines.[2]

This draft protocol encapsulates recent thinking of a group of researchers and technologists related to measurement of the energy efficiency of servers. This effort grew out of two meetings held on January 31st, 2006 and March 27th, 2006 with the support of the U.S. EPA <http://www.energystar.gov/datacenters>. At the second of these two meetings, a group of about fifty representatives of major server and data center companies wrestled with how to structure measurements of energy use of servers in conjunction with performance measurements.

[1] <http://kz.sun.com/teleweb/x64-factor/Ziff_Power_and_Cooling_IT_survey.pdf>.

[2] For example, see <http://www.spec.org/specpower/pressrelease.html>.

OVERALL APPROACH AND PHILOSOPHY

Server performance is a complex issue, and arguments over performance benchmarks are as old as the computer age itself. This document presents a method for attaching energy metrics to existing performance benchmarks. As a generic method, it avoids the difficult issue of which benchmarks better represent real-world applications or conditions. Server purchasers already use performance benchmarks to make purchasing decisions—this method merely adds additional information on energy consumption to those existing benchmarks so buyers can make informed purchasing decisions based on comparable energy data developed using a consistent measurement framework.

Performance-based benchmarks are generally used to determine maximum performance of a given compute configuration applied to a specific set of operations. The many possible uses of these general-purpose machines make it difficult to choose a single metric. Performance based benchmarks are also focused on particular portions of the server's configuration and may tax servers differently depending on the particular characteristics of the benchmark and the server system.

Because many servers operate at a fraction of their maximum processing capacity (Nordman 2005), and because many servers can scale down power use when work loads are below their maximum capacity, it is important to be able to measure energy use for servers at different levels of work load. Efficiency for a server (as opposed to a dedicated client) is achieved in the modulation of the hardware's power consumption over a variety of processing loads well below the peak operation. Highly efficient designs tend to configure their hardware, firmware, and middleware to optimize power utilization across various work loads in the system. This initial effort focuses on transaction-based benchmarks (e.g. web pages served per second) because such benchmarks readily allow estimation of power used by servers at different work load levels. Future efforts will be focused on extending such measurement protocols to cover a broader range of performance metrics.

The focus of this protocol is on measurements of energy use that reflect performance of servers in real data centers. Most servers in business applications run at very low processing loads (typically 10-20% of maximum computing output when in non-virtualized environments), while in scientific computing applications most servers run closer to full capacity. Energy measurements should allow users to determine the loading levels appropriate to their applications so they can compare the energy use of servers accurately.

BOUNDARIES

Because of the short-term focus of this effort (i.e. to establish a credible measurement protocol in advance of more detailed efforts from established benchmarking institutions), the measurement protocol focuses on the direct power use of the simplest small (1U and 2U) servers, which are generally AC powered. While it may be possible to apply the method to medium servers, large servers, and blade servers, the current focus of this protocol is only on this simplest class of servers.

Energy is but one issue facing the buyers of server hardware. This protocol does not explicitly address differences in availability, system management, reliability, response times, or other non-energy related attributes of server hardware. Server purchasers using data based on this protocol will need to choose servers that are comparable in these other attributes, then use the energy data described here to rank order those servers on an energy and performance basis to help inform their purchase decisions.

MEASUREMENT METHODS, EQUIPMENT, TEST CONDITIONS, OUTPUTS

All measurement conditions and outputs must follow the guidelines below, and be reported clearly and accurately in the final measurement reports for every server tested. This protocol calls for energy use measured over some time period, with power levels calculated as the average power over the time period of the measurement.

Environmental conditions

Measurement conditions (e.g. temperature, relative humidity, dew point) should fall within the ranges specified by the ASHRAE Class I standards for data centers. Table 2.1 in ASHRAE (2004) gives a recommended dry bulb temperature range of 20-25 degrees C, a recommended relative humidity (non-condensing) of between 40% and 55%, a maximum dew point of 17 degrees C, a maximum rate of change of temperature of 5 degrees C per hour, and a maximum rate of change of relative humidity of 5% per hour. Measurements should be conducted near sea level, up to 300 meters.[3] Measurement conditions should generally fall in the middle of these ranges and all should be reported in the final measurement document for each server.

Other test conditions

Total Source Impedance: < 0.25 ohm

Total Harmonic Distortion of source voltage (loaded), based on IEC standards: < 5%

For products to be qualified in markets using 110-120V input:

- Input AC Voltage: 115 VAC RMS ± 5V RMS
- Input AC Frequency: 60 Hz ± 1%

For products to be qualified in markets using 100V input:

- Input AC Voltage: 100 VAC RMS ± 5V RMS
- Input AC Frequency: 50 Hz ± 1%

---

[3] The Class I standards apply up to 3050 meters but because the energy use of servers is affected by altitude and the density of air, it is important to do measurements at sea level so they are comparable.

For products to be qualified in markets using 208V / 230V input:

- Input AC Voltage: 230 VAC RMS ± 10V RMS
- Input AC Frequency: 50 or 60 Hz (depending on the market) ± 1%

Meter characteristics

Meters used for these measurements should have the following attributes[4]:

- Power resolution of 1 mW or better;
- An available current crest factor of 3 (or more) at the meter's rated range value;
- Minimum current range of 10mA (or less).
- Frequency response of at least 3 kHz;
- Calibrated within one year of the test date with a standard that is traceable to the National Institute of Standards and Technology (NIST).

Measurement instruments should be able to average power accurately over any user-selected time interval (this is usually done with an internal math calculation dividing accumulated energy by time within the meter, which is the most accurate approach). In this method, the measurement instrument should be capable of integrating energy over any user selected time interval with an energy resolution of less than or equal to 1 Wh and integrating time displayed with a resolution of 10 seconds or less.

Accuracy of meters

Measurements of power of 0.5 W or greater shall be made with an uncertainty of less than or equal to 2 % at the 95 % confidence level. Measurements of power of less than 0.5 W shall be made with an uncertainty of less than or equal to 0.01 W at the 95 % confidence level. The power measurement instrument shall have a resolution of:

- 0.01 W or better for power measurements of 10 W or less;
- 0.1 W or better for power measurements of greater than 10 W up to 100 W;
- 1 W or better for power measurements of greater than 100 W.

All power figures should be in watts and rounded to the second decimal place. For power levels greater than or equal to 10 W, three significant figures shall be reported.

[4] Desired characteristics of meters taken from IEC 62301 Ed 1.0: Measurement of Standby Power, as summarized in the EPA's draft test procedure for measuring Energy Star compliance of personal computers, draft 2, version 4.0, with some loosening of those requirements to better suit this protocol.

Test Configuration

The server under test should be tested as sold and shipped, with all attributes reported so that server purchasers can compare servers on a consistent basis. The boundaries of the system being tested are defined by the needs of the performance benchmark.

Measurements should be taken between the main power outlet and the server, assess total box power, and include all power connections simultaneously. For example, servers with redundant power supplies should have both cords plugged in and the power flow through each cord summed to get total system power. If any external equipment (e.g. disks) is part of the system being tested for purposes of the performance benchmark, power used by those devices should also be included and reported in the output materials.

The configuration and tuning parameters used to achieve maximum benchmark performance in the 100% work load case must be maintained through the process used to generate power levels at lower performance levels. If the system in its "as shipped" state automatically modifies system and configuration parameters at lower work loads, that is not a violation of the procedures of the protocol, but no manual changes by those individuals conducting the test are allowed in those tuning parameters at lower work loads.

The server under test should be connected to an Ethernet network switch capable of the server's maximum network speed. The network connection should be live during all tests. If there are multiple network ports, all should be connected during the test.

SUMMARIZING OUTPUTS FROM THE MEASUREMENTS

Detailed outputs

Frequency, voltage, power factor, and total harmonic distortion (THD) conditions of the input power should be reported. Environmental conditions defined above should also be reported.

For each data point measured, CPU utilization[5] should be reported (along with the key data of processing work load and the power at that work load).

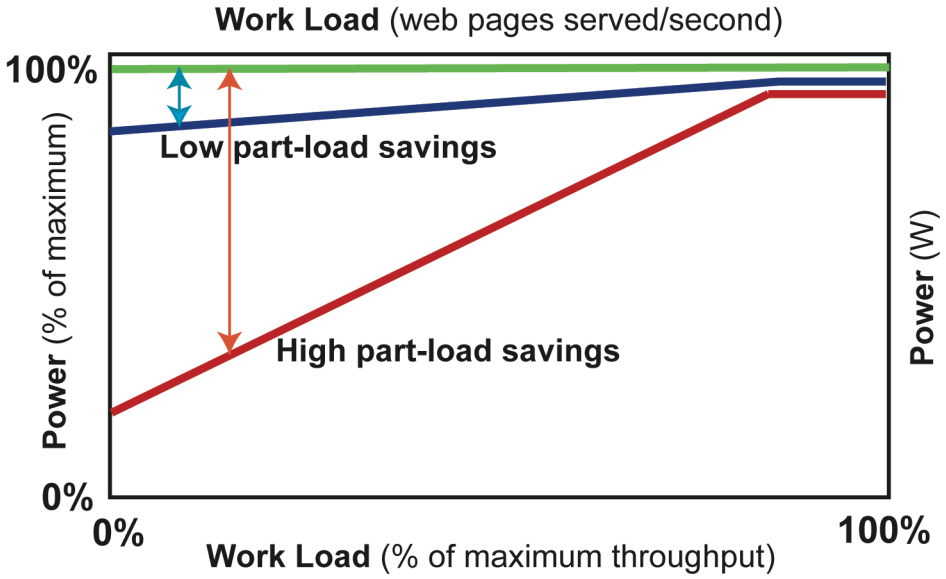
To the extent possible, measurements should include power use and efficiency of subcomponents of the server (e.g., power supply, disk drive, memory, processors, etc). At a minimum, the power supply model, maximum (rated) power, and efficiency curve should be specified and reported along with the power measurements described below. Power factor and THD for the server should also be reported.

[5] Note: the percentage of maximum work load is not generally the same as CPU utilization of the server.

Key summary output

The key output needed from the measurements should be summarized by plotting power consumption as an absolute amount[6] and as a percentage of maximum measured power use on the y-axis, and performance or processing load as an absolute amount and as a percentage of the maximum possible work load on the x-axis. Work load is measured in terms of the performance output of the benchmark, such as web pages served per second. The actual power use of the server as a function of performance levels should then be plotted on the graph. A server that powers down when less computation is required will be represented by a curve or a set of points that decline more steeply as work load decreases

Figure 1: Conceptual Diagram of Energy vs. Computation Metric



Source: Nordman 2005.[7]

In practice, a finite number of points for a range of work loads will be measured. Those points (i.e., the measured data) should be plotted on a graph like Figure 1, but the dots should not be connected because only the dots are real data.

[6] Not the rated power of the power supply but the actual power used by the server at full load under the performance benchmark being used.

[7] For more examples of graphs of this type, see Belady, Tipley, and Lange: <http://thegreengrid.org/pdf/Efficiency_slides_for_General_Distribution_Final.pdf>.

Manufacturers: Creating the power plot

Creating the power plot using a transactional benchmark should involve the following steps:

1) Run the benchmark at full work load and measure the energy use (E_{100} in Watt-hours) over the time needed to run the benchmark (T_{100} in hours), then divide E_{100} by T_{100} to calculate the average power (P_{100} in Watts) associated with the benchmark at full work load. Record the full work load L_{100} (e.g. web pages served over the course of the benchmark test) and P_{100}.

2) Run the benchmark at a different work load (e.g. L_{90} for 90% load) but over the same time period (T_{100}) as for the full load test. In practice, this means that the server will serve only 90% of the web pages served in the 100% case over the time period T_{100}. The pages served should be spread in equal intervals over the time T_{100}. Measure E_{90} and then calculate P_{90} as E_{90} divided by T_{100}. Repeat for 10% work load increments (e.g. 80%, 70%, 60%) all the way down to 10% load.

3) To measure idle power (i.e. power at 0% work load) use the same server connected to the network, but do not run the benchmark program that otherwise would be simulating web pages served and measure E_0 over the same time period as all the other benchmark tests are run (T_{100}). Calculate P0 in the same manner as before, dividing E_0 by T_{100}.[8]

   For purposes of testing and qualifying servers under this protocol, "idle" is the state in which the operating system and other software have completed loading, the machine is not asleep, and activity is limited to those basic applications that the system starts by default.

4) Divide work load at each loading level by L_{100} and divide P at each loading level by P_{100} to calculate the percentages needed to create the power plot. Be sure to put the absolute levels of L_{100} and P_{100} on the graph itself so that it contains complete information about the power use of the server.

Server purchasers: Using the power plot

Metrics involving simple averages across work load levels do not reflect the actual energy use of servers in real-world applications. Most non-virtualized servers in business applications operate only at 10-20% work load while servers in scientific computing applications operate at close to 100% work load. Servers utilizing virtualization in

[8] An alternative would be to run the benchmark program at a load level that was a few tenths of a percent of the maximum load and make the same measurements and calculations. As more experience is gained with measurements of server power, we expect the preferred method for measuring 0% load will become clearer.

business applications are likely to operate at work loads higher than 10-20%, but no generalizable quantitative data demonstrating this effect are now available

Using the power plot, server purchasers can specify the performance levels at which they expect their servers to operate, based on their facility's experience and then read off the power use (as a % of peak power use) at that server work load.

Average power use per server can then be calculated as

$$\text{Avg power use (W/server)} = \frac{\text{Avg output performance}}{\text{Server}} \cdot \frac{\text{Power (W) at avg load}}{\text{Avg output performance}}$$

Output performance (work load) should be measured in whatever units are most appropriate to the server application at the loading level at which the server is expected to operate (say web pages served per second).

Power (in W) can be read off the y-axis for a given work load level. Output performance per watt is calculated by dividing the absolute performance of the server at any point by the power level of the server at that work load level.

Maximum power use per server (which may be useful for sizing cooling and auxiliary equipment) would be calculated as

$$\text{Maximum power use (W/server)} = \frac{\text{Max. output performance}}{\text{Server}} \cdot \frac{\text{Maximum Power (W)}}{\text{Max. output performance}}$$

Average energy use over a year would be calculated as

$$\text{Energy use (kWh/year)}^{9} = \frac{\text{Avg power use (W)}}{\text{Server}} \cdot \frac{1\ \text{kW}}{1000\ \text{W}} \cdot \frac{8766\ \text{hours}}{\text{year}}$$

Those server purchasers who are unable to specify at what work load level their equipment will operate should assume (in the absence of data) a work load level of 15% and create the calculations outlined above for that work load level. If these users know they will run scientific applications most of the time, then they should use the power data for 100% work load to calculate power and energy use.

Once these power plots are developed by manufacturers as a matter of course, then server purchasers can use them to specify equipment. For example, a server purchaser could rank by energy use all servers that meet certain performance criteria at a given load level, and then consider more carefully the top 20% of servers by that metric. It will also allow

---

[9] Note that 8766 hours per year is the average including leap years (non leap years have 8760 hours, while leap years, which occur every fourth year, have 8784 hours).

a purchaser to ask more pointed questions about the energy use of the components of a given server than they are now able to do (such as "What would happen to the power plot and the capital cost of that server if we put in a more efficient power supply?").

ISSUES FOR FURTHER RESEARCH

The development of this initial measurement protocol highlights several potentially fruitful areas of future research.

First, there is the general issue of the applicability of any specific benchmark to particular applications. For example, High Performance Computing (HPC) workloads are measured using different benchmarks than web hosting applications. The extent to which power plots differ when different benchmarks are used to create them (which is in part a function of how well each benchmark accurately reflects real-world applications) will require further investigation. Consistent comparisons should eventually be created with real data in order to understand this issue and develop a strategy for addressing it (if it turns out to be important to the accuracy of the results).

The general method and measurement criteria presented in this protocol can be applied easily to non-transaction-based benchmarks that measure peak performance of servers (such as the SPEC CPU2006 benchmark suite). The result would not be a power plot, because these benchmarks do not lend themselves to part-load measurements, but instead would be an estimate of the measured (actual) maximum and idle/zero load power use of the server. Data of this type have been developed for a small number of servers for an ASHRAE report (2004) and could easily be expanded to become common practice. This would make energy consumption estimates available for a broader range of commonly used benchmarks, so effort should be focused on expanding the protocol to cover those benchmarks.

The rapidly growing market in blade servers, which in many cases compete with the 1U and 2U servers covered by this protocol, makes it imperative that the issues associated with measuring power used by blades be addressed soon.

More data are needed on typical workloads in different data center environments. The work load levels for business and high performance computing users cited in this document are anecdotal estimates from discussions with many industry experts, but more accurate data are needed to better characterize work loads for specific applications and industries.

There is a potential issue with improved response time at lower workload levels. Processor management strategies may vary regarding response times as the workload scales back. In order to ensure a consistent comparison, explicit specifications may be necessary to establish response time constraints at various load levels. The need for such constraints will depend on the actual variance of such response times. In the protocol as it now stands, response times should be reported for each measured data point, which will provide the information needed to analyze this issue and determine the course ahead.

The distinction between 0% work load and idle/standby conditions may need to be investigated further. The 0% load case could be implemented by sending a very small number of client requests to the server over the testing cycle (say 0.1% of the maximum work load). Alternatively, it could be measured by sending no client requests to the server over the testing period. If a server is able to power down after a certain period of inactivity, it might be able to achieve a lower 0% work load power level in the second case than it could in the first. At this point, very few servers include the ability to reduce power use in standby, but such a feature could become more important in the future, in which case an energy measurement protocol would need to more precisely specify how to conduct the measurements at 0% work load.

Finally, many current and upcoming technological developments may complicate the creation of future performance metrics, and while the measurement protocol is somewhat insulated from that problem (because it merely attaches energy measurements to whatever performance metric exists) some of these developments may have implications for how the metric should be structured in the future. For example, advanced power management technologies, virtualization, multiple cores on a chip, inactive and non-high performance system states (among other issues) may all affect characteristics of servers in a way that has important implications for energy use. These developments will need to be followed closely as measurement protocols evolve in coming years.

SUMMARY

The graph of server power vs. performance/load is the key output from the metrics process, which allows the user to calculate average power use at a given work load level (as well as peak power use). The default server work load for typical business uses should be about 15%, while for scientific applications it should be close to 100%. The widespread use of the power plots described above should encourage server buyers to demand and manufacturers to supply servers that deliver equivalent performance at lower energy use than would otherwise be the case.

REFERENCES

ASHRAE. 2004. *Thermal Guidelines for Data Processing Environments*. Atlanta, GA: American Society of Heating, Refrigeration, and Air-Conditioning Engineers. <http://www.ashrae.org>

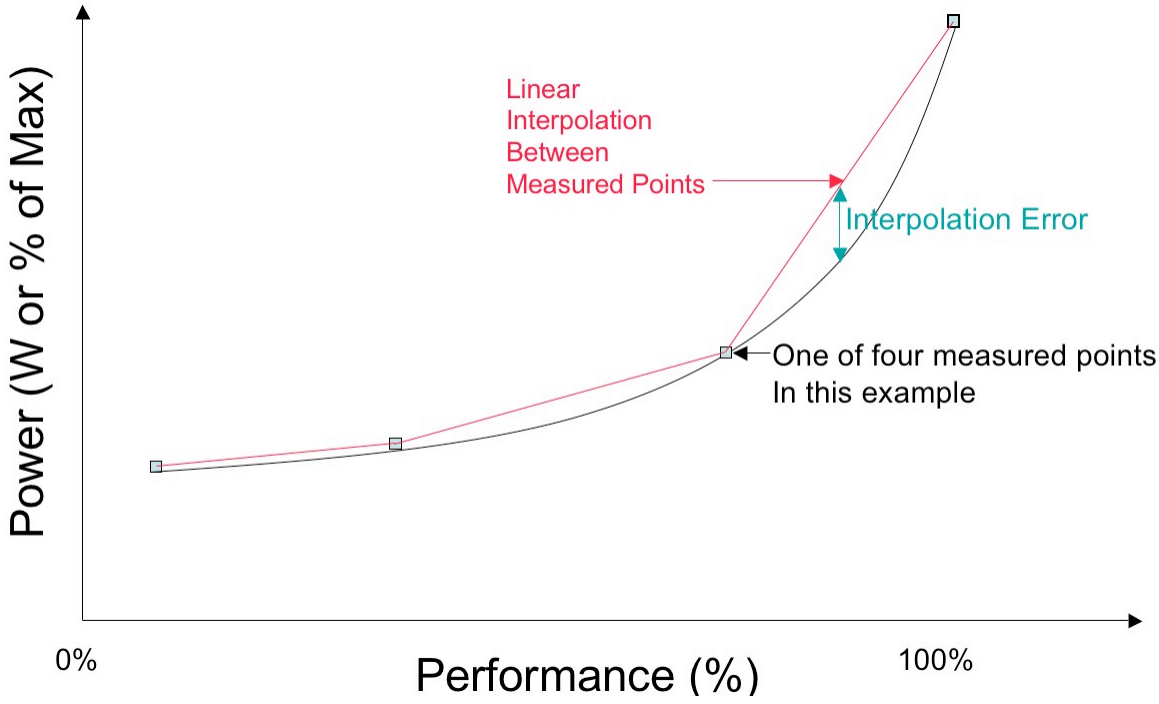
Nordman, Bruce. 2005. *Metrics of IT Equipment — Computing and Energy Performance*. Berkeley, CA: Lawrence Berkeley National Laboratory. Draft LBNL-60330. July 26. <http://hightech.LBL.gov/datacenters.html>

APPENDIX A: HOW MANY MEASUREMENTS ARE NEEDED?

How many data points are needed to accurately specify the power plot? We would like to collect just enough data to achieve some pre-specified accuracy level. This choice will involve balancing the time and money needed to conduct the measurements with the accuracy required.

Figure A-1 shows one example of how such an error limit might be specified. In this example, four points are measured, and are connected with straight lines. The curved line represents the actual power plot, and the blue double arrow labeled interpolation error is the error level we want to measure (the error can be expressed as a percentage of the absolute power level on the actual power plot). In the beginning phases of measuring power use, it will be important to measure a large number of points (say at 5% or 10% intervals, starting at 0% and going to 100%). With such data, one can easily calculate errors introduced by reducing the number of points to some smaller level. Once enough data are collected, one can then make more accurate determinations of just how many points need to be measured.

Figure A-1: Assessing measurement errors in the power plot data



Source: Nordman (2005).