The State of Energy and Performance Benchmarking for Enterprise Servers

Andrew Fanara
United States Environmental Protection Agency
ENERGY STAR® Product Specifications Development Team
fanara.andrew@epa.gov

Evan Haines
ICF International
Associate – Energy, Climate, and Transportation Practice
ehaines@icfi.com

Arthur Howard
ICF International
Associate – Energy, Climate, and Transportation Practice
ahoward@icfi.com

The original publication is available at www.springerlink.com.

Abstract. To address the server industry's marketing focus on performance, benchmarking organizations have played a pivotal role in developing techniques to determine the maximum achievable performance level of a system. Generally missing has been an assessment of energy use to achieve that performance. The connection between performance and energy consumption is becoming necessary information for designers and operators as they grapple with power constraints in the data center. While industry and policy makers continue to strategize about a universal metric to holistically measure IT equipment efficiency, existing server benchmarks for various workloads could provide an interim proxy to assess the relative energy efficiency of general servers. This paper discusses ideal characteristics a future energy-performance benchmark might contain, suggests ways in which current benchmarks might be adapted to provide a transitional step to this end, and notes the need for multiple workloads to provide a holistic proxy for a universal metric.

1 Introduction

All day, every day, servers process and deliver increasing quantities of video, voice, and data through a vast global network to several billion devices, where that data is consumed and often stored for posterity. In this context, if computing is the heartbeat of a global network, servers are the muscle. It can be argued that the quality of life for the billions of people who rely upon ubiquitous computing would suffer without access to continually evolving computing technology. A variety of industries have invested tremendous resources to enhance the reach, richness, and speed of digital information, but the rapid growth of energy consumption by these enhanced services warrants increased scrutiny. As broad segments of the world economy increase their focus on energy efficiency, this scrutiny will help to ensure that continued increases in computing performance can be achieved without a run away increase in energy consumption.

The current market shows a discernable trend towards the improvement of operational productivity of computing systems, and data center operators around the world are taking an increased interest in energy performance when procuring IT equipment. While this has not yet become a universal management imperative, there is little doubt that organizations that embrace an energy efficiency strategy will minimize future risks to their business with the most sustainable data center operations. Building on the substantial progress made in this industry to date, additional tools are needed to uniformly assess and improve the efficiencies of IT equipment. One such tool would be a universal metric for server efficiency which is applicable to a majority of the server market. Such a generalized metric would provide end users with a window into the energy performance of systems under consideration and provide the data center industry with a stepping stone toward the smarter procurement of efficient servers.

1.1 Energy Constraints in the Current Data Center

The energy efficiency of information technology (IT) equipment and data center facilities has dramatically increased in importance over the past decade in response to the rapid growth in the number and size of data centers and the power and cooling constraints of the associated infrastructure. Consider the following:

- *Rising Data Center Costs*. McKinsey Consulting estimates that the cost of running data centers is increasing by as much as 20 percent a year, while overall IT spending is increasing by only 6 percent.[1]
- *Power Grid Capacity*. In a report to Congress, the EPA estimated that ten new power plants would be required to meet the additional energy demand from data centers by 2011.[2] Evidence of this trend is already mounting; a utility provider in Virginia estimates that by 2012, 10 percent of all the energy it supplies to northern Virginia will be consumed by data centers.[3]
- *Load and Demand*. EPA estimates 6X growth in server capacity and 69X growth in storage capacity in this decade.[4]

Available energy at the server-, rack-, row-, or building-level is often a bottleneck that hinders an organization's ability to meet the computing capacity demands of an increasingly digital economy. Ample supply of electricity is an important prerequisite for selecting the location of a new data center facility. Existing facilities can be haunted by the risk of grid congestion and peak power concerns. Moreover, if variable real time electricity pricing becomes commonplace data center operational expenses could rise well above current levels, especially during peak periods. While server compute performance may continue to be defined using contemporary rating systems, a clear metric for the work performed per unit of energy consumption has yet to be universally established. The development and adoption of such standard metrics would greatly improve the ability of data center operators to increase efficiency by maximizing the work completed by servers for a given energy consumption. Furthermore, greater access to detailed power information would facilitate better capacity planning for increased efficiencies in data centers. Breaking down the barriers obscuring this information is essential in order to provide clear indications of the energy-performance balance rather than the *perceptions* often reported on the market today.

1.2 Benchmarks, Metrics and Reducing the Total Cost of Ownership

The computing industry has long used software benchmarks as a basis for comparing the performance of competing server products. Such software benchmarks are developed to measure the output of servers as they perform standardized, representative workloads. The results of these benchmark tests allow products to be directly compared in a way not easily achieved in an actual operating environment. Software benchmarks output a metric indicating the server's ability to complete the workload's tasks, typically represented by the system's speed (e.g. operations per second). The resulting data provides the industry with a meaningful tool to compare competing systems or quantify improvements on a single system.

The rising cost of energy and corresponding increases in energy consumption together drive the need for server benchmarks with a broad focus on both speed-oriented performance and associated energy consumption. Existing benchmark methodologies vary in their ability to meet this need. Maximized computational performance will remain an important goal for server development, but a benchmark that solely focuses on compute performance does not easily fit into the total cost of ownership (TCO) calculation. In this case, analysis of both performance and energy require additional end user research or testing.

As an alternative to strict performance-based metrics, a second benchmark approach present on the market compares computational performance to a measure of TCO only including hardware and maintenance costs. This approach, which provides insight into how a server meets the day to day operational needs of the data center, makes it possible to compare cost-effective performance of

[1] Forrest 2008.
[2] US EPA 2007.
[3] Garber 2009.
[4] US EPA 2007.

various products. Still, this risks the under-representation of the broader operational cost of running a server; energy remains a missing component, and a significant omission: the 2007 *EPA Report to Congress on Server and Data Center Energy Efficiency* noted at that time that server energy costs would exceed the hardware purchase cost of a server by 2008.[5]

The addition of standardized energy measurement during benchmark testing expands the scope of a benchmark to include a more holistic view of the server in operation. A number of benchmarking organizations have undertaken efforts to include energy measurement methodologies within their processes; a few examples will be discussed in **Section 4**. The existence of these efforts points not only to the market's desire for this information, but also to the intrinsic strength of benchmarking organizations as trusted information resources. As performance benchmarks have evolved over time to serve a competitive and diverse market seeking standardized test methodologies, the development processes surrounding them have incorporated characteristics that support expansion into meaningful energy comparison:

- *Consortium-based development processes* provide input into workload development by a range of industry stakeholders with knowledge of available technologies, industry trends and developments in the market.
- *Pre-determined and transparent testing methodologies* ensure comparable results using agreed upon procedures for standardized energy measurements.
- *Structured versioning and revision schedules* allow for periodic updates to ensure continued applicability of energy metrics as technologies mature and change.
- *Established presence in the market* with well-understood workloads that provide context to associated energy measurements.

With these building blocks in place, there is clear context to provide the needed tools to address server efficiency and to contribute to the reduction of energy consumption, thereby reducing the long term TCO.

2 Current State of Server Performance Metrics and Benchmarks

2.1 Traditional Benchmarking: Determining Maximum Capability

Server benchmarks set a proxy by which computing performance can be consistently measured, quantified and understood. Benchmarks also facilitate ranking systems based on stable underlying testing conditions and settings. These two roles are closely tied. Though a particular workload may either be synthesized to exercise hardware under artificial conditions (synthetic workload) or designed to run a series of processes based on an end-use application mix (application-based workload), the repeatability and standardization of the process allows for direct comparisons of relative performance.

The proxy and ranking functions have traditionally been associated with maximized performance conditions. Vendors have responded by developing and aggressively marketing servers which can attain the highest benchmark scores. This focus leads to an emphasis on the highest *achievable* result instead of the *actual* performance that may be observed in a real end-use application in the data center. Though these benchmarks effectively illustrate maximum performance potential, they underemphasize the performance (and efficiency) of products as they would actually be used in the market. The maximum case does little for an end user seeking information on expected performance of the system once installed at their facility.

2.2 The Future Role of Benchmarks: Incorporating both Efficiency and Performance

Integration of efficiency measurements into performance benchmark results can effectively extend the applicability of existing benchmarks to more realistic end-use scenarios. In the hypothetical example presented in **Table 1**, three systems have completed a benchmark where data is presented in terms of performance (completed operations), efficiency (operations per watt consumed), and average idle

[5] US EPA 2007.

power measurements. Server 1 is the clear winner in terms of pure computational performance. However, a closer look at the data presented in this manner shows that Server 2 produced the more efficient completion of the workload per watt of power consumed. Server 3 was inferior to 1 and 2 in both completed operations and operations per watt but had a significantly lower idle power measurement.

Table 1. Example of holistic benchmark results.

	Completed Operations @ 100% Utilization	Completed Operations per Watt	Average Server Idle Power	Best-Suited Use
Server 1	400,000	1000	165	Maximum Performance
Server 2	250,000	1200	110	Efficient Operation
Server 3	200,000	950	70	Underutilized Applications

While hypothetical, this example illustrates how unique selection criteria by different audiences may yield diverse interpretations of the same set of data. An end user with business needs driven solely by computing performance might select Server 1, though they will have been made aware of the energy consumption penalties associated with this choice. A second user with similar computational needs but a tight power or density budget might choose Server 2, since it provides the best balance of energy use to workload performance. Finally, a third user with light application loads who expects long periods of idle time might find that Server 3 provides acceptable performance while also minimizing power consumption in the most common mode for expected applications. All three of these audiences are able to act on the cost-performance analysis most appropriate to their specific business needs.

With the growing emphasis on both energy and performance in the data center, the measurement of energy for existing benchmarks will be necessary to meet end-user expectations. Rather than highlighting only the fastest systems, there will also be demand to identify the most efficient systems, including models or configurations previously overlooked in benchmark results or by industry marketers. From a benchmark development organization's perspective, a greater demand for benchmarking data may result from a new audience looking for efficiency data rather than just maximum performance data. It is illustrative to consider a future scenario in which such benchmark development might result in a universal or generalized, metric for server efficiency. In the next section, ideal characteristics of such a unified approach are considered.

3 Development of a Generalized Energy and Performance Benchmark

In this section we consider a few important considerations for the development of a generalized metric for server energy efficiency and computing performance. These considerations include discussions of power versus energy measurement, synthetic versus application-based workloads, and other factors.

3.1 Power versus Energy Measurements

It is important to contextualize the differences between instantaneous power measurements, time-scaled energy measurements, and averaged values of each measurement as important elements of a power and performance benchmark. Marketing claims regarding energy efficiency for IT equipment are more prevalent in recent years; and while such efforts may meet the information needs of end users, marketing materials often use the terms energy and power interchangeably. It is important that the implications of each term are understood as they apply to benchmarks and metrics.

It is potentially less complicated to use instantaneous power measurements when performing a benchmark test, yet care must be taken to properly frame the periodic nature of a typical computing

workload. Averaged power reporting over time can be effective as a proxy for the expected power consumption of a workload exhibiting stabilized or cyclic behavior. Selecting an appropriate sampling rate for power measurements is critical to the quality of the measurement; if readings are not taken frequently enough, one risks overlooking important system events that have a significant effect on average power consumption.

In contrast, measuring energy accumulated over time requires either (1) that instantaneous readings be abstracted to apply to an expected usage case, or (2) that the selected workload is truly representative of actual server operation. One risk with the accumulated energy approach is that end users may make incorrect assumptions about the relationship between watt-hour output and utility pricing. To mitigate this risk, data on the time taken to complete the workload and the instantaneous power consumption during the test should be provided along with the accumulated energy data to ensure that the test results are taken in proper context.

In general, it is critical to the success of the benchmark metric that workload weightings and measurement inputs are made available to the end user. Transparency preserves the context of the data and enables end users to assess the relevance of the results to their specific application environment. As an example, the Version 5.0 ENERGY STAR Computer Specification includes an efficiency metric based on kWh ratings.[6] In addition to publishing the calculated kWh "score," the ENERGY STAR program makes transparent the equation used to calculate the score and requires vendors to report the measured power inputs entered into this calculation. While the standard efficiency equation is weighted based on statistically relevant data, this transparent reporting structure provides a means for end users to estimate their own energy costs based on the specifics of their application.

3.2 Synthetic versus Application-Based Workloads

Two common workload structures for benchmarks are *synthetic workloads* that drive the server to complete as many artificially-derived tasks as possible in a given amount of time, and *application-based workloads* that measure the server's ability to complete predetermined operations based on real applications. A generalized server efficiency benchmark could make use of either type of workload, but any results would have to be carefully annotated to ensure that they are interpreted properly by end users. The impact of each structure on the marketing of power and energy results is considered in **Table 2**.

Table 2. Comparison of different benchmark characteristics.

	Synthetic Workload	Application-Based Workload
Power Measurement	• Opportunity to meet the steady-state condition necessary to support averaged power measurement. • The number of operations most likely varies between tests. It is reasonable to report averaged power, but also to frame the power levels with information on utilization during the test.	• May not meet the steady state condition since transitions between applications or the realistic variations in power necessary to complete tasks will vary from test to test. • The number of operations may vary similarly to the situation for a synthetic workload. Averaged power is again useful to report. Associating the average power measurements to the applications in the workload can provide insight into architecture's ability to handle elements within the workload.
Energy Measurement	• A set time period can provide structure to energy measurement, but results are best weighted with the number of operations completed during the time period. • Since operations vary from test to test, this workload structure is not easily positioned to report a generalized "expected energy consumption."	• As systems improve in performance, a task may initiate and conclude too rapidly to derive a meaningful energy measurement. • Since the server is completing the same set of tasks and may vary in utilization during the workload, energy data provides more of the expected variety important for development of a generalized energy consumption model.

[6] ENERGY STAR 2009.

3.3 The Use of a Generalized Server Efficiency Benchmark

Widely used server performance benchmarks typically mimic or replicate intended workloads in the data center. An effective *generalized* benchmark – one applicable for a wide variety of system applications – should give end users an indication of how a particular server ranks compared to others in general operation through an assessment relevant to different workload types. Because workloads within data centers vary widely, there will never be a perfect correlation between the work performed in a benchmark workload and that which is performed in an end-use application. There is no true substitution for testing a server with an actual application workload, but for buyers without the resources to conduct such in-depth testing, an effective generalized benchmark should provide insight into server performance under a variety of operating conditions.

Examples of typical workloads run by servers and represented by available benchmarks are *high performance computing (HPC)*, *web services or other accessed services*, *email services*, *database management*, and *shared file services*. These five categories represent a broad cross-section of server uses and illustrate the types of workloads that could be assessed by a generalized benchmark. Examples of benchmarks used to approximate these workloads are available in **Section 4**. Although these workloads are expected to cover the majority of the server market, other common and niche application workloads may exist.

An available benchmark that produces general indication of broad-based server efficiency and performance is currently missing from the market. Such a benchmark – capable of representing more than one workload type – might be thought of as a *first-order* approximation of the energy efficiency of a server; benchmarks based on one of the five referenced workload types might be thought of as a *second-order* approximation, providing greater accuracy for a specific workload type. A *third-order* approximation of energy efficiency could be achieved by testing a server in its intended application, affording more precision at the cost of additional testing resources. Server purchasers might rely on a mix of first, second, and third-order approximations depending on available resources.

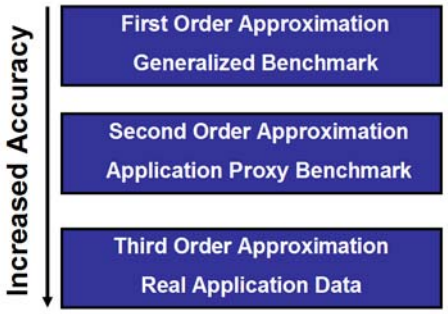

Figure 1. Hierarchy of Different Benchmarking Approaches.

For example, large organizations might use a first-order approximation to narrow down a list of hardware platforms for more detailed benchmarking or application testing, while a smaller buyer looking for a general workhorse server to run a number of different applications might use a first-order approximation as their sole purchasing criteria.

3.4 Technical Characteristics of a Generalized Benchmark

Most servers can be thought of as consisting of a few key components and capabilities that will affect the performance and energy consumption of that server, which have been summarized in **Table 3**.

Table 3. Capability factors in server performance benchmarks.

Capability	Component(s)	Description
Compute	Processors and system memory	Performing operations, i.e. switching 1s and 0s

Storage	Hard drives, solid state drives, etc.	Long term storage of data, i.e. keeping 1s and 0s
Input and Output (I/O)	Network cards, RAID/SAS controllers, etc.	Transferring data in and out of devices, i.e. moving 1s and 0s

Different workloads require a different mix of these basic capabilities. For example, an HPC application will be almost all compute, while file services, in contrast, will be very storage and I/O intensive. A conceptual illustration of this concept is included in **Figure 2**.

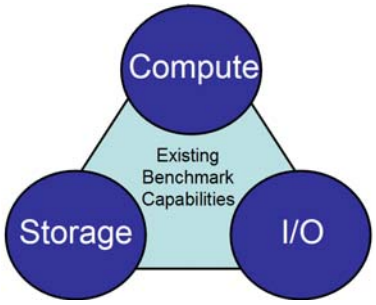

Figure 2. Capability factors in server performance benchmarks.

A truly generalized benchmark would test relative energy and performance efficiency for each of the three factors, using a combination of the relative efficiencies of each capability to arrive at a generalized system efficiency. A server with high compute efficiency (e.g., with high efficiency processors and/or memory) and a low efficiency I/O device would receive a moderate efficiency rating on the generalized scale, while a server with high efficiency in all three factors would rate much higher. If the specific efficiencies of each capability could be separately assessed, this benchmark could also be used to identify servers ideal for more specific workload scenarios. A generalized benchmark capable of evaluating a server in this way could be developed with either a synthetic benchmark designed to stress each factor in turn, or with a carefully-selected set of application code designed to concurrently assess the performance of each factor.

3.5 Other Important Elements of a Generalized Power and Performance Benchmark

A benchmark is only useful if there is a low barrier to entry for its use and it is adopted by a large segment of the industry it is intended to serve – there must be a critical mass of test results available to allow purchasers to make meaningful comparisons to support their purchasing decisions. To lower this barrier to entry, there are many other criteria a successful benchmark must meet to maximize its effectiveness in the market:

- Able to operate on a wide variety of system architectures and operating systems.
- Low cost to run and report data in a standard way.
- Scalable with system size.
- Easily configured for consistent, repeatable results.
- Consistent with current standards for operation of equipment in data centers.
- Able to assess the relative efficiency of multi-node and blade systems.

4 Using Existing Benchmarks to Assess Generalized Server Efficiency

Many benchmarks exist in the current market to measure the performance of systems under various workloads. This section will focus on benchmarks intended for general servers and how such benchmarks might be combined to create a generalized metric for server efficiency.

4.1 Selection of Current Benchmarking Organizations

Transaction Processing Council (TPC)

TPC is a non-profit corporation and industry consortium which focuses on benchmarks for data base systems and transaction processing. Transactions measured and tested by TPC involve common business processes. A typical transaction, as defined by the TPC, would include the updating of information in a database system for purposes such as inventory control, airline reservations, or banking transfers. Systems relevant to TPC benchmarks are often large database systems composed of many subcomponents (e.g., servers, external storage, and networking) which create the larger systems. Certain TPC benchmarks already include metrics for $/operation, and the organization is currently engaged in ongoing efforts to include energy measurements for the benchmarks, so that their metrics include a true measure of TCO (including energy costs) for all benchmarks. Draft energy measurements are expected in 2009.[7] Further information on TPC and their benchmarks can be found at www.TPC.org.

Standard Performance Evaluation Corporation (SPEC)

SPEC is a non-profit corporation and industry consortium which focuses on the creation of server benchmarks for a variety of standard data center applications. The SPEC benchmarks are typically aimed at individual server systems and specific subsystems. A SPEC subcommittee has recently developed a standard protocol for measuring and reporting power consumption as part of the measurement and reporting process for its benchmarks. SPEC released the first such benchmark (SPECpower_ssj2008) in 2008 and the second (SPECweb_2009) in 2009, and will continue to revise its other benchmarks to include power consumption measurements.[8] Further information on SPEC and their benchmarks can be found at www.SPEC.org.

Green 500

The Green 500 is a ranking of the most energy efficient super computers in the world. The Green 500 uses the LINPACK benchmark along with associated power measurement techniques to measure floating point operations per watt.[9] Further information on Green 500 and their benchmarks can be found at http://www.green500.org/

[7] Transaction Processing Performance Council.
[8] Standard Performance Evaluation Corporation.
[9] The Green500.

4.2 Available Benchmarks by Data Center Workload category

Table 4. Typical data center workloads and available benchmarks. Additional details of the available benchmarks are included in **Appendix A** of this report.

Data Center Workload Category	Available Benchmarks
High performance computing (HPC)	LINPACK, Green 500*, SPEC_CPU2006
Web services or other accessed services	SPECpower_ssj2008*, SPECweb2009*, TPC-App
Email services	SPECmail2009
Database management	NNA Server Power Efficiency*, NNA Server Transaction Throughput Benchmark, TPC-C, TPC-E, TPC-H
Shared file services	SPECsfs2008

(*) denotes benchmarks that currently integrate power measurement into results/procedures.

4.3 Measuring Power using Existing Benchmarks

If existing benchmarks are to be used as a proxy to measure the energy efficiency of servers, it will be necessary to develop standardized procedures for adding power and/or energy measurements to some existing benchmarks. The EPA set the stage for this work in the 2006 release of an initial *Server Energy Measurement Protocol*[10] and in the 2009 release of the *ENERGY STAR Test Procedure for Determining the Power Use of Computer Servers at Idle and Full Load*, as Appendix A to the ENERGY STAR specification for Computer Servers[11]. As described in the SPEC procedures and *Server Energy Measurement Protocol,* benchmark tests should, where possible, be performed at a number of different load points, including at a minimum full load (100%) and idle (0%), in order to allow for the development of a power and performance load curve. An example load curve from a SPECpower_ssj2008 result has been included in **Figure 3** to illustrate this approach. In order to use existing benchmarks to assess generalized server efficiency, more investigation may be necessary to ensure that existing practices can be applied to some current benchmarks which do not yet include energy or power measurements.

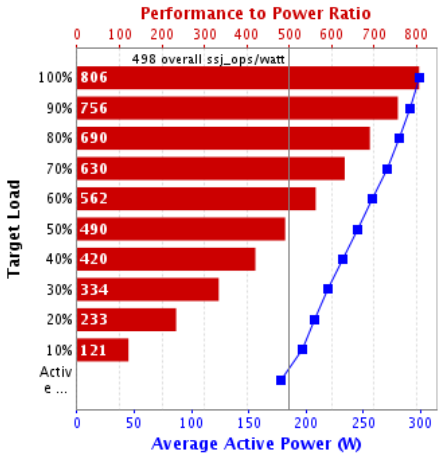

Figure 3. Example SPECpower_ssj2008 result showing a measured load curve.[12]

[10] Koomey, *et al* 2006.
[11] ENERGY STAR 2009.
[12] Standard Performance Evaluation Corporation. *SPEC and the benchmark name SPECpower_ssj2008 are registered trademarks of the Standard Performance Evaluation Corporation. For the latest SPECpower_ssj2008 benchmark results, visit http://www.spec.org/power_ssj2008/results/.*

4.4 Creating a Generalized Server Efficiency Metric from Existing Benchmarks

The development of an ideal generalized efficiency benchmark for servers as described in **Section 3** could be a lengthy and challenging process. However, the recent emphasis on efficiency and energy management in the data center illustrates that there is momentum in both the manufacturer and end-user communities to support such an effort.

In the short term, this suggests an opportunity to bring together the efficiency metrics referenced above to develop a hybrid metric to assess server energy efficiency. Since servers can be expected to operate under a variety of applications and workloads, this hybrid metric would integrate elements from a variety of workloads. These workloads, as well as appropriate benchmarks which act as proxies to server performance, could be chosen from **Table 4** in **Section 4.2**. A possible scenario would be to select a single, representative benchmark from each category for inclusion in the hybrid metric; this scenario would minimize the testing burden on manufacturers and ensure uniformity in results between systems. Once a list of appropriate workloads and benchmarks was selected, data could be collected to assess different options for a generalized efficiency metric. The following approaches could be considered:

- Measure the relative efficiency of each benchmark separately, to allow end users to determine which metric is most suited to their particular application;
- Weight each benchmark to calculate a single hybrid efficiency metric based on the combined test results; or
- Identify a preferred benchmark that served as the best proxy for all additional benchmarks (i.e. select the single benchmark that best preserves the relative ranking of server efficiency for all benchmarks).

Data gathered during the development and implementation of a metric based on existing benchmarks could then form the basis for development of a more advanced generalized efficiency metric that meets the intent and ideal requirements identified in **Section 3**.

Regardless of the approach used to leverage existing benchmarks, a new emphasis should be placed on testing a wider variety of servers, as configured for shipment to the end customer, with their associated benchmark scores disclosed. Greater disclosure of consistent, accurate performance data – including energy consumption – across a diverse set of server product lines will enable smarter procurement practices and stimulate competition while continually propelling market transformation.

5 Conclusion

Performance benchmarks have traditionally focused on measuring maximum computing performance without regard to energy efficiency. However, the importance of environmental issues related to computing is prominently discussed in the business community today. While the use of TCO as a purchasing tool has increased, more transparency is needed to identify operating costs that are specifically attributable to energy consumption, and to highlight the role of inefficient computing practices in exacerbating these costs.

The community responsible for server performance benchmarks is well-positioned to contribute to the development of new metrics which include energy efficiency in addition to computing performance. The consortium-based development structures and open process for sharing performance data that are the hallmarks of performance metrics would also serve the development of energy efficiency metrics. Numerous benchmark organizations have already recognized this opportunity by developing independent methods to collect energy or power information as a standard practice.

This paper reviewed the current state of server energy and performance benchmarking, highlighting important issues for consideration in further benchmark development. The server industry as a whole, however, continues to focus primarily on setting new benchmark records for maximized workloads. By incorporating energy measurement into benchmark results, the industry can help mainstream product

configurations become more competitive in the marketplace based on optimized operational and efficiency performance.

While it can be argued that data derived from a discrete set of workloads is not representative of actual server performance in all cases, the very nature of benchmarks as a standardized evaluative set of methodologies does provide a means for end users to make meaningful comparisons of different server products. It will be important for benchmark development organizations to continue efforts to standardize energy measurement methodologies in a manner that is consistent with how products are actually operated in the field, so that benchmark results are repeatable and relevant to real world conditions.

A generalized benchmark, applicable for a wide variety of data center applications, will remain a valuable objective for the server industry. Current and forthcoming efforts to enhance existing performance benchmarks will provide the foundation on which to build a generalized assessment tool, and will provide an ongoing catalyst for continued energy efficiency improvements in servers. The benchmark community should continue to seek out opportunities to integrate energy measurement into standard benchmark procedures, and should standardize the collection of power and/or energy data in benchmarking procedures. By making energy measurements a common and accepted part of performance measurement, the benchmarking community will be able to reach a wider audience, broaden the scope of systems that can be measured with existing benchmarks, and serve their customers needs for insight into expected energy performance.

References

1. ENERGY STAR Program. ENERGY STAR Program Requirements for Computer Servers (2009)

2. Forrest, W., Kaplan, J., and Kindler, N.: “Data Centers: How to cut carbon emissions and costs.” McKinsey on Business Technology 14, 6 (2008)

3. Garber, K. The Internet’s Hidden Energy Hogs: Data Servers. US News and World Report (2009)

4. The Green500 List, http://www.green500.org

5. Hass, J., Monroe, M., Pflueger, J., Pouchet, J., Snelling, P., Rawson, A., Rawson, F.: Proxy Proposals for Measuring Data Center Productivity. The Green Grid: Beaverton, OR (2009)

6. Koomey, J., Belady, C., Wong, C., Snevely, R., Nordman, B., Hunter, E., Lange, K., Tipley, R., Darnell, G., Accapadi, M., Rumsey, P., Kelley, B., Tschudi, W., Moss, D., Greco, R., Brill, K.: Server Energy Measurement Protocol. Analytics Press: Oakland, CA: (2006)

7. Koomey, J., Brill, K., Turner, W.P., Stanley, J.R.: A Simple Model for Determining True Total Cost of Ownership for Data Centers. The Uptime Institute: Santa Fe, NM (2007)

8. Neal Nelson & Associates, http://www.nna.com

9. Standard Performance Evaluation Corporation, http://www.spec.org

10. Stanley, J.R., Brill, K., Koomey, J: Four Metrics Define Data Center “Greenness” Enabling Users to Quantify “Energy Efficiency for Profit” Initiatives. The Uptime Institute: Santa Fe, NM (2007)

11. Transaction Processing Performance Council, http://www.tpc.org/information/about/abouttpc.asp

12. United States Environmental Protection Agency: Report to Congress on Server and Data Center Energy Efficiency, Public Law 109-431 (2007)

Appendix A – Available performance benchmarks

Benchmark Name (*Organization*)	Intended Workload	Workload Category	Performance Metric	Power/ Energy Meas.?*
LINPACK (*N/A – Public*)	Floating point operations	High Performance Computing (HPC)	MFLOPs	No
LINPACK (*Green 500)*	Floating point operations per Watt	HPC	MFLOPs / Watt (peak performance divided by average power)	Yes
NNA Power-Efficiency Benchmark (*Neal Nelson and Associates*)	WWW transaction requests	Database Management	Watts for a given transaction rate	Yes
NNA Server Transaction Throughput Benchmark (*Neal Nelson and Associates*)	WWW transaction requests	Database Management	Transactions / minute	No
SPEC_CPU2006 (*SPEC*)	Integer speed (*SPECint2006*), integer rate (*SPECint_rate2006*) and floating point speed (*SPECfp2006*), floating point throughput (*SPECfp_rate2006*)	HPC	N/A – unitless mix of various performance measurements from multiple workloads	No
SPECmail2009 (*SPEC*)	Corporate mail server workloads based on number of users	Email Services	Sessions / hour	No
SPECsfs2008 (*SPEC*)	File server throughput and response time	Shared File Services	Throughput (ops/sec), response time (msec)	No
SPECpower_ssj2008 (*SPEC*)	Java based applications	Web/Accessed Services	Operations / watt (ssj_ops/watt)	Yes
SPECweb2009 (*SPEC*)	Http transactions including: Banking, ecommerce and support	Web/Accessed Services	Simultaneous user sessions (SUS) / watt	Yes
TPC-App (*TPC*)	Application server and web services	Web/Accessed Services	Web Service Interactions / second (SIPS), price / interaction ($/SIPS)	Pending
TPC-C (*TPC*)	New-order transactions	Database Management	Transactions / minute (tpmC), price / transaction ($/tpmC)	Pending
TPC-E (*TPC*)	On-Line Transaction Processing (OLTP): workload of a brokerage firm	Database Management	Transactions / second (tpsE), price / transaction ($/tpsE)	Pending
TPC-H (*TPC*)	Decision support benchmark of business oriented queries	Database Management	Query-per-Hour (QphH@Size), price / query ($/QphH@Size)	Pending

***** Denotes status of power/energy measurement as an integral methodology within the benchmark