Understanding Data Center Energy Intensity

A Dell Technical White Paper

John Pflueger

Dell │ Office of the CTO

August 2010

Contents

Tables

Figures

Executive Summary

The energy consumption of servers and data centers has become an important issue both for industry professionals responsible for data center and government entities responsible for energy-related issues. The industry has responded well. Data center professionals now have access to and have operationalized metrics that describe how well a data center's physical infrastructure delivers power from utilities to the data center's IT equipment. At the same time, the industry is also aware that it is in need of additional metrics that describe how efficiently IT converts energy into compute work.

Energy intensity metrics have been used in other fields to provide intuitive insights into energy consumption and some measure of work output. This type of metric has advantages over more traditional productivity metrics. Energy intensity can be easily broken down into smaller elements representing specific parts of the data center or specific types of equipment. This makes potential comparisons between different facilities and trends within an individual facility more meaningful and easier to analyze. This type of approach can work for data centers, however, calculating energy intensity still requires some means for estimating the amount of compute performed, or work output, by the data center.

The Green Grid has published a number of white papers on IT productivity. These define a number of possible approaches, or proxies, for estimating work output. Dell has leveraged both this work and its own internal research to produce a dataset that covers a production data center comprised of several thousand servers. This data, combined with regularly reported power data for the same data center, has enabled Dell to estimate the energy intensity for a production data center for every hour over a three-week period. The resulting dataset suggests strongly that this approach shows significant promise with respect to estimating data center work output.

This type of analysis creates an important feedback mechanism that can identify whether or not specific IT-related initiatives are affecting energy consumption, both in an absolute and relative sense. It also provides guidance as to how server and storage architecture choices affect data center power consumption. This includes policy choices such as the percent of data center work output that is performed within virtualization environments, hardware refresh rates, and the implementation of tiered storage.

Dell concedes that the approach and dataset are just a beginning. Collection and analysis of the initial dataset was more manual than automatic. Current data center performance tools, however, should easily be able to incorporate this approach into future management applications. While this is just the beginning for energy intensity, this is a beginning that shows promise. With future work, the industry will be able to further refine and develop these types of metrics, as well as develop new operational models for managing data centers with respect to both work output and resource consumption.

Introduction

Data Center Energy Consumption

As information technologies have become a larger and larger part of our economy, server and data center energy consumption has become responsible for a measurable fraction of all purchased electricity in the United States. By one estimate, servers and data centers accounted for 1.5% of U.S. electricity consumption in 2006, with this potentially growing to 2% to 2.5% of U.S. electricity consumption in 2011 (U.S. EPA, 2007).

In addition, energy availability and cost concerns, as well as infrastructure and building construction costs, are forcing organizations to take a close look at the energy profiles of their equipment and data centers (Judge, Pouchet, & Dixit, 2008). In fact, for some configurations, the cost of the energy consumed by a server may exceed the original purchase price of the equipment (Belady, 2007).

The policymakers' responses to these issues have been to put renewed focus and incentives around improving energy efficiency. Their responses have led to programs such as the EPA's ENERGY STAR® for Data Centers initiative (U.S. EPA) and the European Union's Code of Conduct for Data Centers (EC Joint Research Centre - Institute for Energy - Renewable Energies Unit, 2009).

The industry's response has been both additional focus within organizations on energy efficiency, and also development and support of industry organizations aimed at understanding and addressing these issues such as The Green Grid (http://www.thegreengrid.org) and the Climate Savers Computing Initiative (http://www.climatesaverscomputing.org/).

Regardless of whether the audience is regulatory or commercial, the industry requires strong metrics to understand the data center and provide guidance as to how best to attack the issues. The industry has had great initial success with Power Usage Effectiveness (PUE); however, PUE only provides insight and guidance into the data center's physical infrastructure—power distribution and cooling architectures (The Green Grid, 2007). PUE does not aim to and is not meant to provide guidance on IT equipment. It is the servers, storage and networking equipment, however, that performs the work of the data center. We must understand these if we hope to understand and improve data center energy efficiency.

Energy Intensity and the Data Center

To better understand data center energy consumption, data center professionals can pull a page from the Econometrics handbook. Energy Intensity is a metric that is used to relate resource consumption, typically Joules of energy, to a unit of work product, typically dollars of economic output. As a concept, Energy Intensity has been used to compare energy use across different industries, as well as understand how energy consumption changes as a nation develops its industries and economy.

To give one example, U.S. Energy Intensity in 2008 was estimated at 8.52 thousand BTUs per dollar of GDP—in 2000 *chained dollars*—with this figure declining from 12.28 thousand BTUs per dollar of GDP in 1988 (U.S. Energy Information Administration). In other words, U.S. Energy Intensity, or the energy required to produce value in the economy, decreased by 30% over this twenty-year period.

Why Consider Energy Intensity?

The main factors driving our interest in data center energy intensity are:

- Energy intensity provides a more intuitive feel for how the data center consumes resources.
- The math behind energy intensity provides an easier mechanism for understanding how different areas within the data center contribute to the energy intensity of the whole.
- Energy intensity enables meaningful comparisons to be made between some separate facilities.

Relationship to Energy Consumption

Energy Intensity metrics are the inverse (not the opposite) of *Energy Efficiency* metrics. To illustrate, we can look at fuel efficiency metrics.

In the U.S., we use *miles per gallon* (mpg) as our primary metric for fuel efficiency. In this metric, the result, *miles travelled* is the numerator and the measure of fuel used, gallons, is in the denominator. In Europe, however, the predominant metric is *liters per 100km* (L/100km). In this case, fuel consumption is the numerator and the result is in the denominator.

The European approach enables a more intuitive understanding of the relationship between improvements to the metric and fuel savings. Table 1 provides a comparison of how these metrics work. To save one gallon for every hundred miles driven, an automobile has to reduce its energy intensity by 2.4 liters per hundred kilometers. The required change in mpg to save one gallon for every one hundred miles driven is harder to calculate. Starting at 15 mpg, a car must improve to 17.6 mpg; starting at 25 mpg, a car must improve to 33.3 mpg. In the latter case, the relationship between the key metrics and fuel savings is not as clear.

Table 1. Comparison of Fuel Consumption Metrics

To save one gallon per 100 miles driven:

mpg (old)	mpg (new)	mpg change	L/100km (old)	L/100km (new)	L/100km change
15.0	17.6	2.6	15.7	13.3	-2.4
20.0	25.0	5.0	11.8	9.4	-2.4
25.0	33.3	8.3	9.4	7.1	-2.4
30.0	42.9	12.9	7.8	5.5	-2.4
35.0	53.8	18.8	6.7	4.4	-2.4
40.0	66.7	26.7	5.9	3.5	-2.4

Energy Intensity metrics provide more intuitive insight into improvements in resource consumption—a percent decrease in energy intensity directly corresponds to a similar percent decrease in resource consumption.

Understanding the Parts in Relation to the Whole

An energy intensity approach makes it relatively easy to understand how a subsystem or subprocess contributes to the system as a whole. When several subprocesses are required to produce a result, overall energy intensity is the sum of the energy intensities of the individual elements. When a given piece of work or product can take one of several different paths, overall energy intensity is a weighted sum of the energy intensities of the separate paths.

When the relationship between individual elements and the overall system has been identified, energy intensity can be calculated by focusing on smaller and smaller, more focused elements. An example of how this works for data centers is described as follows.

Making Meaningful Comparisons

In addition to enabling the calculation of energy intensity, the ability to decompose a large system or process into individual elements makes comparisons across different systems or situations easier.

Managers and regulators have been interested in metrics that enable the comparison of different data centers; however, the varied nature of these facilities frequently makes comparison difficult, if not impossible. For instance, differences in purpose, location, and age mean that comparing two facilities can be like comparing apples to oranges.

Easy decomposition of the higher-level system into smaller subsystems addresses some of these issues. For example, while there still may be important differences, it is far easier to compare two data center power distribution architectures than it is to compare two data centers.

Energy Intensity With Respect to Other Data Center Metrics

Of existing data center metrics, the one that is closest to Energy Intensity is Data Center Energy Productivity (DCeP). DCeP is an IT productivity metric first formulated, described, and proposed by The Green Grid. The relationship between energy intensity and DCeP is similar to the relationship between liters per 100 km and miles per gallon: Energy Intensity is simply the inverse of the DCeP metric. In should be noted that, in both cases—Energy Intensity and DCeP—the calculation of the metric requires an approach for measuring or estimating the work output of the data center. The Green Grid has published a number of documents on this problem (The Green Grid, 2009), and any approach to measuring or estimating data center work output will work equally well regardless of whether the high-level data center metric is Energy Intensity or DCeP.

An Energy Intensity Model for the Data Center

Direct application of energy intensity to the data center requires having a model for relating the different elements in the data center to each other and a meaningful denominator, or measure of work output. While for comparisons between industries, dollars of output may be useful; however, a data center is better measured by its computation work output. With this approach, an Energy Metric for data centers takes the form of Energy Consumption over Work Output.

In the sections that follow, we'll first look at how the Energy Intensity model enables the data center to be broken down into subsystems—first, between power, cooling and IT—then, within different classes of IT equipment.

Power, Cooling and IT

In an energy intensity calculation, the total system energy intensity is equal to the sum of the energy intensities of the required elements. At the highest level of the data center, this means that data center energy intensity is equal to the sum of the energy intensity of the power distribution architecture, the energy intensity of the cooling architecture, and the energy intensity of the IT equipment, as is depicted in Figure 1. Those industry professionals familiar with PUE will recognize the similarity between the basic formula for PUE and data center energy intensity.

Power distribution, data center cooling and IT are frequently addressed by different organizations within one company. The decomposition allows these groups to work within their disciplines to calculate their own energy intensities, and then roll that data up into an energy intensity for the facility as a whole.

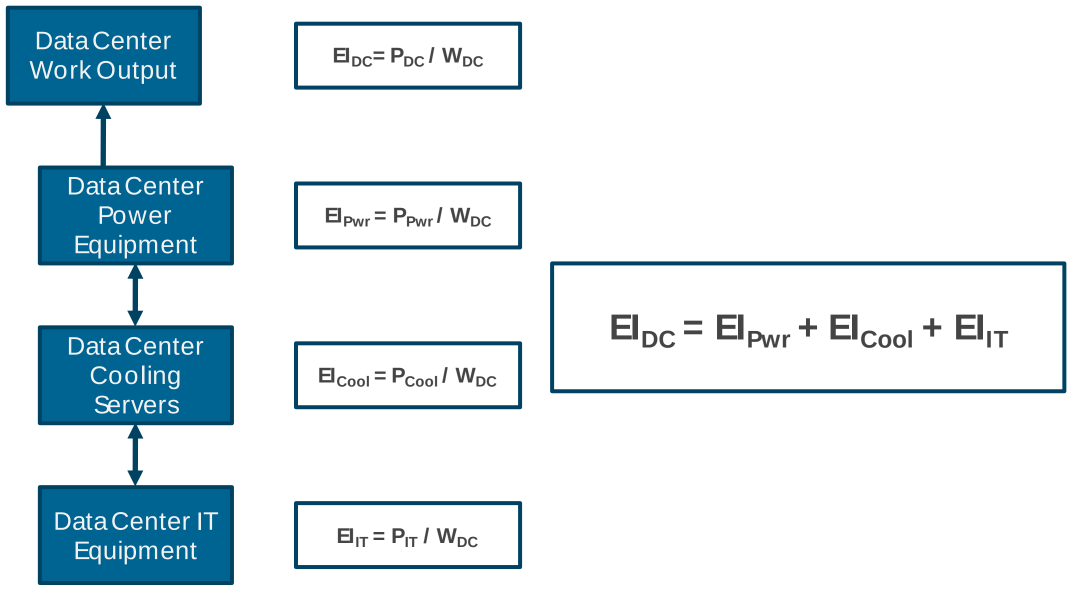

Figure 1. Basic Data Center Energy Intensity Decomposition

In addition to the basic analysis, energy intensity can be broken down further within each of these disciplines. One approach is to analyze the distinction between fixed losses (losses that are present regardless of workload) and variable losses (losses that are proportional to data center work output or server load). This type of analysis might work very well in determining seasonal variations in energy intensity due to outside environmental conditions. A second approach is to break down the analysis by subclass of equipment. This approach might be useful, for example, in understanding the losses inherent in different levels of a power distribution architecture.

Decomposing IT into Servers, Storage and Networking

Much in the same way, we can use this approach to understand the Energy Intensity of servers, storage and networking equipment. Figure 2 shows the breakdown of IT energy intensity into these separate areas. In this case, however, networking and storage energy intensity have some additional components. These components are domain-specific energy intensity metrics and additional factors relating each domain's work to the data center work as a whole.

Creating domain-specific metrics for storage and networking enables energy use to be measured in units appropriate for each domain[1]. The additional factors convert metrics of interest within these domains (the amount of transported data in the case of networking and the amount of stored data in the case of storage) into the overall metric's work denominator.

[1] Experts in both the networking and storage domains will consider this view to be somewhat simplistic. The example shown is meant to be one possible approach. Future approaches may include metrics such as IOPs or latency.

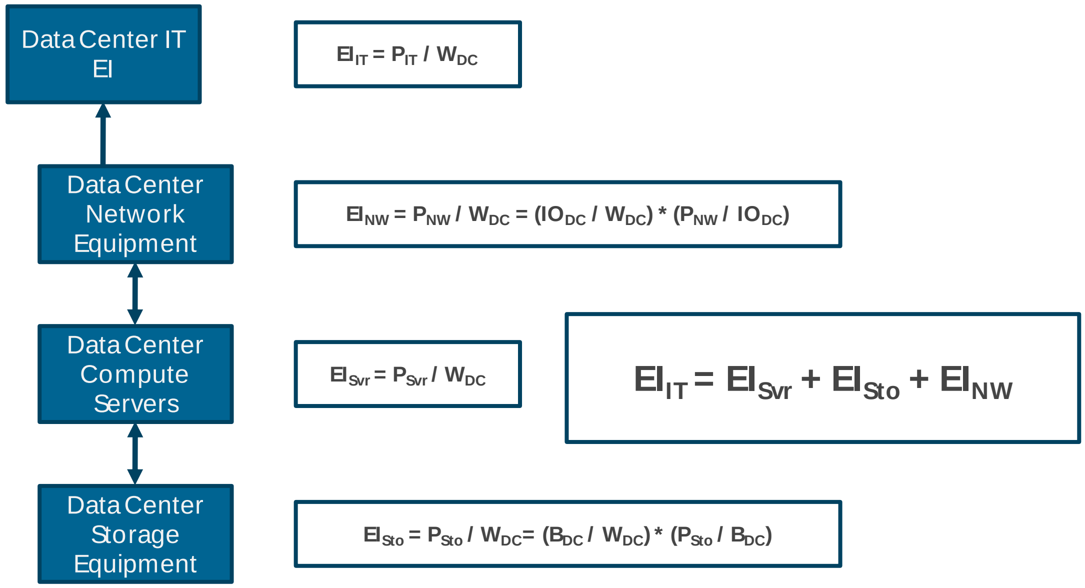

Figure 2. An Example of IT Energy Intensity Decomposition

Drivers of IT Component Energy Intensity

While there are a number of ways to perform a deeper analysis of data center power consumption, one approach is to consider it in the context of the data center’s load, its equipment, and its operations. These are directly reflected in a small set of key questions:

- How much work am I doing?
- What is the base efficiency of the equipment I’ve deployed?
- How well am I using that equipment?

A summary of how these questions relate to other domain-specific metrics is presented in Table 2, Key IT Performance Metrics, which specifies key metrics for each of the three key questions listed and for each of the three main classes of IT equipment: Compute Servers, Storage Equipment, and Network Equipment.

In this approach, energy intensity, or the power required to perform a set amount of work, is equal to 1/(Utilization * Efficiency). The total power consumed by that equipment is equal to the energy intensity times the workload. The general approach is the same whether we are looking at compute servers, storage or networking.

Table 2. Key IT Performance Metrics

	Compute Servers	Storage Equipment	Networking Equipment
How much work am I doing?	Compute Workload	Bytes Stored	Bytes Transported
What is the base efficiency of the equipment I've deployed?	Server Performance[2]/Watt	Storage Capacity/ Watt	Available Bandwidth/Watt
How well am I using that equipment?	Aggregate Server Utilization[3]	Storage Utilization	Network Utilization

Calculating Data Center Energy Intensity

The first application of data center energy intensity is to calculate the metric and determine what type of guidance it provides and how this data can be used to guide data center level decisions.

Estimating Data Center Work Output

While the preceding model is useful in describing the data center in terms of its constituent parts and providing a framework for calculating energy intensity in general, it is impossible to calculate a specific number until we have a means for estimating the *work output* of a data center.

The data center industry has been looking at this topic for some time now. The Green Grid published its first high-level paper on this subject in 2008 (The Green Grid, 2008). It followed this work in early 2009 by developing and publishing a number of different approaches to estimating the work performed by servers and the data center through proxies (The Green Grid, 2009) and in late 2009 by publishing and analyzing collected feedback on the proxies (The Green Grid, 2009). Each described proxy represents one approach to estimating work output by gathering key operational information and processing it in the context of other static information about the server and data center.

Dell currently favors a proxy that combines CPU utilization, an indicator of how hard a system is working, with system benchmark results, an indicator of the maximum possible performance of a system. While we are certain that this proxy will not be the ultimate approach, our initial experiences with this approach for estimating server and data center work output have been very positive. Figure 3 shows a work output chart for a data center over a period of three weeks generated by calculating this metric for several thousand servers and summarizing the results.

With data centers rapidly gaining the ability to report overall power consumption almost continually, and with tools in place to estimate data center work output, the data center manager can now calculate data center energy intensity. The next step is to understand the *why*. What types of guidance can this metric provide; how should it be used?

---

[2] Alternatively, this metric is Server Compute Capacity per Watt.
[3] Aggregate server utilization is the total compute workload divided by the sum of all available compute capacity. This metric is related to, but in general not equal to, average server utilization.

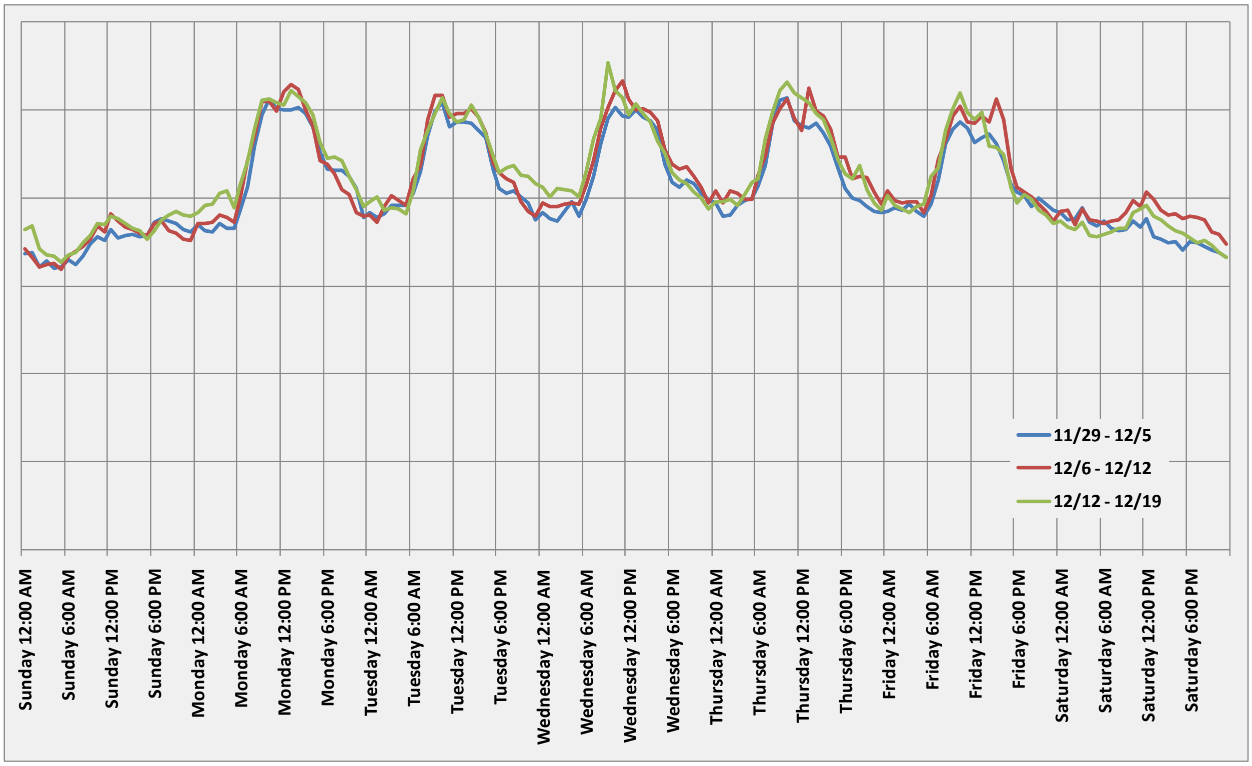

Figure 3. Work Output for a Production Data Center Over a Three-week Period[4,5]

Sample Data Center Energy Results

Figure 4 shows the calculated energy intensity for the production data center from Figure 3. In this chart, however, the facility power consumption and IT power consumption (measured from the PDU) are divided by the estimated work output of the facility.

A number of features from the graph are interesting:

- There are clear 24-hour cycles and 7-day cycles to the chart, and both IT Energy Intensity and Physical Infrastructure Energy Intensity experience the same general peak-to-trough behavior, where *peak* is worst-case energy intensity and *trough* is best-case energy intensity.
- Best energy intensity during the workday is around 75% of the peak weekday energy intensity, with weekend energy intensity, on the average, 10% worse than peak weekday energy intensity.
- Between November 30 and December 3, physical infrastructure energy intensity was close to half its average value during the balance of the time of the dataset.

The 24-hour and 7-day cycles evident on the chart in Figure 4 mimic the same cycles seen in the work output chart in Figure 3. This occurs because power consumption does not vary much with server work

[4] Data was collected over a three-week period from November 29, 2009 to December 19, 2009 and includes approximately 3900 servers.
[5] The function used to estimate a work output number from an individual server multiplies CPU utilization by either an actual or estimated benchmark result. In this case, therefore, the y-axis represents units roughly equivalent to SPECintRates2000 results.

output. This is indicative of IT systems with relatively large idle power components and physical infrastructure systems with corresponding high fixed losses (i.e., physical infrastructure losses do not scale well with IT workload). Both the IT equipment issues and physical infrastructure issues point toward future opportunities for improving energy intensity.

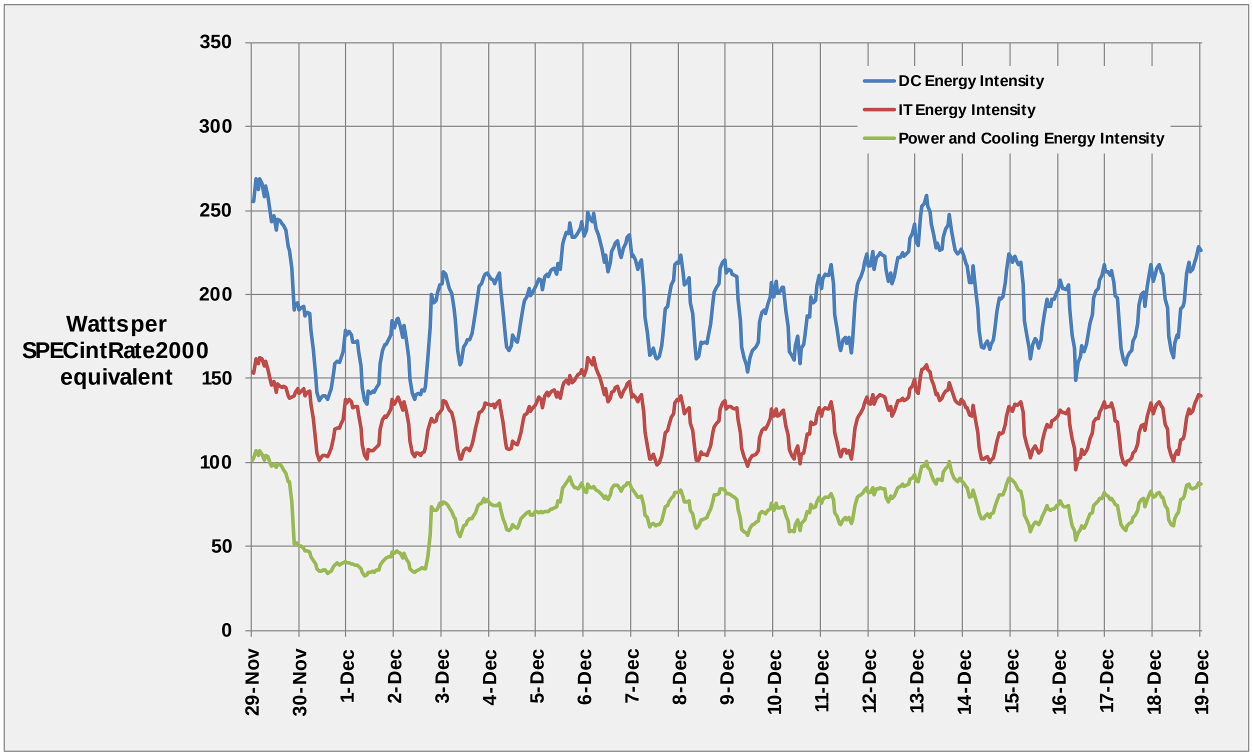

Figure 4. Data Center and Component Energy Intensity Over a Three-Week Period

The relationship between worst-case energy intensity (peak) and best-case energy intensity (trough) helps quantify the maximum potential savings available through operational management of workloads. An optimum strategy that maintains a set server utilization rate regardless of individual server workloads (by aggregating work and shutting unused servers off) would likely have, as best case, an IT energy intensity no lower than the lowest energy intensity achieved during the workday. In this case, that means a maximum possible improvement of ~25% during weekday evenings and ~35% during weekends.

The change in power and cooling energy intensity between November 30 and December 3 was due to a change in the data center. During this period of time, Dell was bringing air from outside of the facility into the data center. The data from this chart suggests that, in this case, a free-air cooling strategy reduced power and cooling energy intensity by close to half, with a corresponding reduction in overall data center energy intensity of between 15% and 20% (representing an absolute reduction in energy intensity of ~30 Watts per SPECintRates2000 equivalent work units). It should also be noted that, as would be expected, the temporary change in cooling architecture did not affect IT energy intensity. This demonstrates the value of an overall metric that is the sum of several smaller, somewhat unrelated, metrics.

Using Energy Intensity Metrics

The availability of energy intensity metrics for the data center provides new guidance on a number of key issues and decisions.

Evaluating Data Center Energy Programs

The primary use of these metrics should be as a tool for evaluating the usefulness of initiatives or programs aimed at reducing resource consumption. PUE and DCiE have already shown that when the proper metrics are available, data center managers will be successful in using them to improve operations. As mentioned before, however, the physical infrastructure metrics are not able to provide guidance pertaining to IT equipment.

The usefulness of this approach originates from its ability to normalize data collected from the data center at different levels of equipment deployment. The data center is not a static facility. Within each facility, new applications and servers are constantly deployed while older ones are decommissioned. As the data center is filled-out, its power consumption may naturally increase due to the deployment of new systems performing additional work. Any analysis of power data must be able to determine whether or not data center performance has actually improved or declined. By considering the amount of work being performed by the facility and including that in metrics calculations, data collected at different times in the data center's life can be compared.

Because the metric encompasses elements for all of the equipment in the data center—whether physical infrastructure or IT equipment—this approach can compare the effects of initiatives as different as installation of air-side economization and workload virtualization. This gives the industry a tool to understand not only how these sorts of initiatives relate to each other in terms of magnitude, but also how they may either help or interfere with each other.

Improving Server Energy Intensity

In addition to providing general guidance on the overall data center, this approach to estimating the work output of production servers also highlights key areas of focus for improving server productivity. As highlighted in Table 2, the energy intensity metric for compute servers relies on three components: the amount of work being performed, compute server utilization, and the performance per Watt of the server population.

Managing Data Center Workload

Managing the first component, the amount or work being performed in the data center, is difficult. Typically, the data center's workload is considered to be a *given*—the work that's done is the work that's required for the business to operate. In the short term, little of the work that is done within the data center is optional.

There are, however, other factors, decisions or policies pertaining to data center work that affect data center energy intensity. The management tools we select and our implementation of these tools can both affect resource consumption. These tools represent overhead; they are important for the continued operation of the data center, but are typically not the final work product of our most business-critical applications. Indeed, our choice of applications and how we architect them affects energy intensity as well. In addition, with this type of metric, we now have a tool that can help us understand the potential benefit of scheduling work, where possible, outside of normal business hours.

Another look at the data in Figure 3 and Figure 4 helps to highlight the potential benefits. The data center is at its lowest energy intensity and at its maximum productivity during the peak part of the business day. At the same time, the data center is also being asked to perform the most work. In the middle of the night, the data center is asked to do the least amount of work, and it also has its highest energy intensity. A policy that time-shifts appropriate workloads from peak hours to off-hours, simultaneously moving that work to servers that are lightly loaded during the off-hours, could help flatten the energy intensity curve and reduce overall resource consumption.

Improving Server Utilization

The single initiative an IT organization can undertake with the greatest effect on energy intensity is the consolidation of an existing server population using virtualization platforms. While the benefits from clearing out legacy servers are fairly easy to calculate, however, virtualization farms typically include significant quantities of SAN. The additional storage equipment consumes some of the resources that were freed through the decommissioning of the older servers. A rigorous analysis of the benefit of virtualization on resource consumption has to include both an analysis of the energy intensity of the compute server population and also the changes to the energy intensity due to the increased requirements for storage.

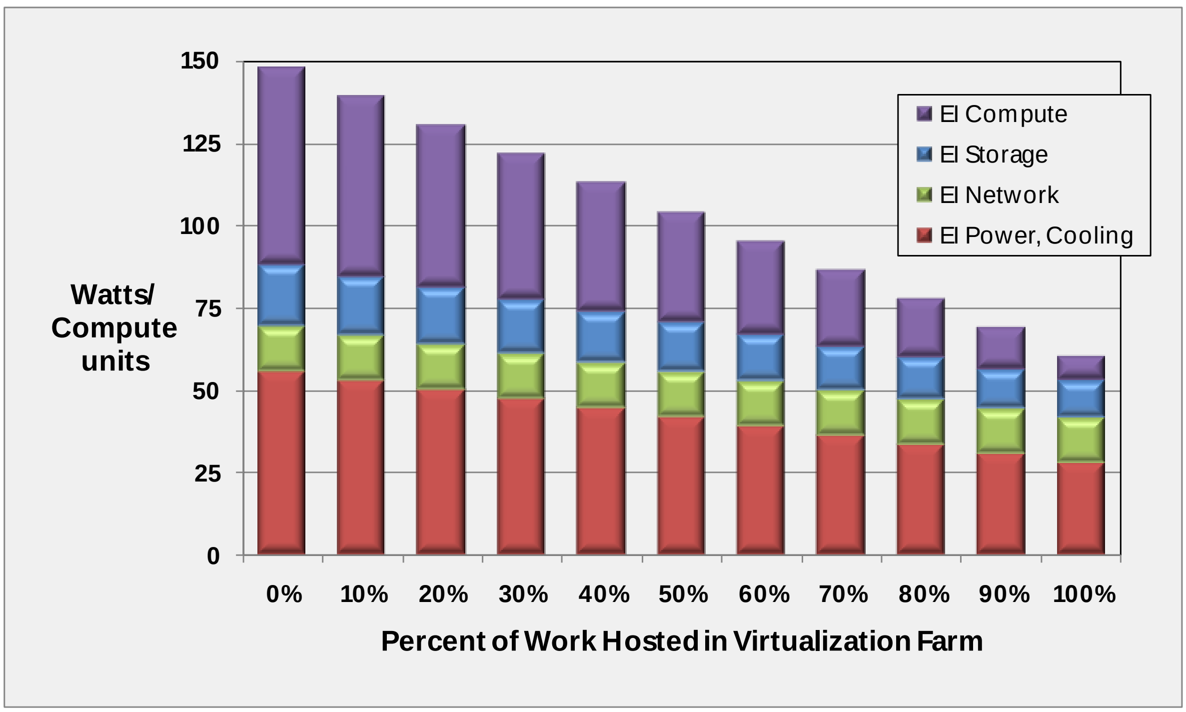

Figure 5. Virtualization Implementation and Data Center Energy Intensity

Figure 5 shows an example of the relationship between the scope of an organization's virtualization implementation and the resulting effect on data center energy intensity. In this example, the data center can reduce its energy intensity by close to half by virtualizing 80% of the work in the data center.

With respect to energy intensity, the main benefit comes from improved server utilization. For some workloads, however, virtualization is not an option. In these cases, the use of performance analysis tools and other means to architect solutions with significant improvements to utilization can result in similar benefits.

Improving Server Performance Per Watt

With every new generation of servers, average server performance and average server performance per Watt improve. For smaller, non-scalable applications or for non-virtualized environments, these benefits are difficult to capture[6]. For those servers, however, that are either used as hosts for virtualization platforms or are supporting large, scalable applications, the implementation environment allows those servers to leverage these natural performance improvements.

Figure 6 shows the result of some internal modeling work at Dell using Dell's Data Center Performance Estimation Tool. This chart provides the model's estimates for how the energy intensity of a data center's server population will change over time for three different policies relating to the rate at which server hardware is refreshed[7].

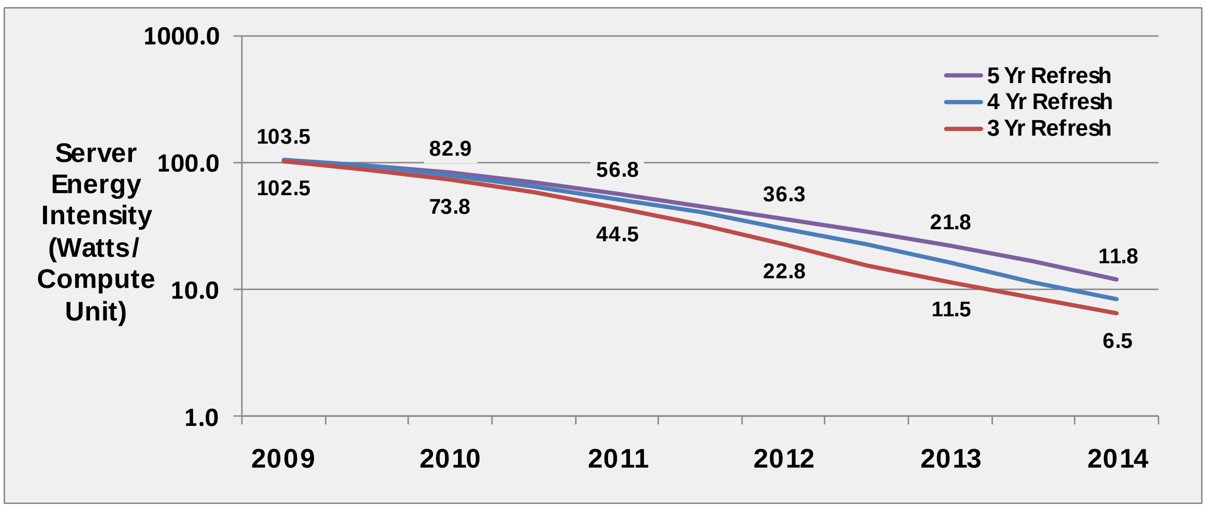

Figure 6. Server Energy Intensity as a Function of Hardware Refresh Rate

There are two key items to note from this chart. First, the model assumes that the data center specified in this case has implemented a program to virtualize a significant percentage of the data center's workload. This includes both new work required of the data center and old work previously supported on those servers that are being decommissioned as a result of the scenario's hardware refresh policy. The presence of a program almost guarantees that the data center will both better manage server utilization and be able to leverage generational improvements in server performance per Watt. Note that, for each of the three hardware refresh policies modeled in the chart, server energy intensity goes down dramatically over time.

---

[6] For this paper, scalable applications are those applications that are deployed on a large number of servers and for which application demand and compute supply can be measured and managed in such a way as to enable active control of server utilization.

[7] Server energy intensity is equal to the power consumed by the compute server population divided by the estimated work output of the same population. To calculate overall data center energy intensity, one would incorporate additional energy intensity numbers pertaining to the storage and networking equipment populations and data center power and cooling architectures.

In addition, the potential energy intensity benefits from implementing a more aggressive hardware refresh strategy are significant. The model estimates that reducing refresh time from five years to three years eventually reduces server energy intensity by more than 40%.

Improving Storage Energy Intensity

Traditionally, much of the attention paid to IT equipment, with respect to energy efficiency, has been focused on compute servers. While strategies for improving energy consumption in servers have been evolving for some time, storage equipment and storage management have only recently begun to address power-related issues.

Table 3. Storage Technologies and Energy Intensity

Component of Energy Intensity	Key Technology/Policy
Bytes Stored	Data Deduplication
Storage Capacity/Watt	Energy Efficient Storage Equipment Hardware Refresh Solid State Devices Tiered Storage
Storage Utilization	Thin Provisioning Tiered Storage

That being said, storage is catching up quickly. A number of technologies have been introduced that present significant opportunities for improving storage energy consumption. Technologies such as data deduplication, thin provisioning, and intelligent data management are rapidly becoming mainstream. In addition, there is an increasing interest and emphasis on understanding the percentage of data center power being consumed by storage systems.

In a manner similar to that for compute servers, the energy intensity of storage equipment is based on three elements: the total amount of data being stored, the baseline power-performance of a storage equipment population, and the utilization of that same equipment. Table 3 describes the relationship between these key technologies and the components that comprise storage equipment energy intensity.

Data Deduplication

Within the world of compute servers, it is relatively difficult to affect the amount of work performed by the server population. Most of the data center's work is at the request of the users. Within the world of storage, however, the emergence of data deduplication as a viable technology provides a means for reducing the total amount of data stored in the data center.

The power advantage comes from a reduction in the number of drives that are active, although this does come at a cost as the deduplication and data access processes require extra compute overhead to store and retrieve the compressed data. Data deduplication is also most appropriate for inactive, infrequently accessed information—it is not implemented across the entire data center.

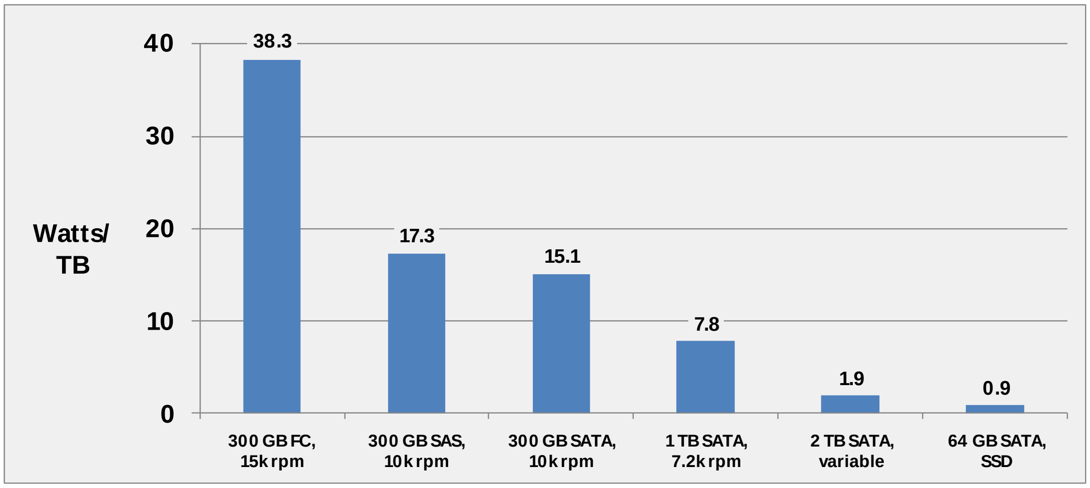

Figure 7. Minimum Storage Energy Intensity for Selected Enterprise Hard Drives[8]

Energy-Efficient Storage Equipment and Tiered Storage

Storage equipment provides a huge range of options when it comes to balancing power consumption with capacity and performance. The trade-off usually arises between power consumption and capacity/performance. For example, Figure 7 shows the power consumption of different drives per Terabyte of storage. In general, solid state drives are the most energy efficient, followed by those hard drives with the largest available capacity per spindle. Smaller drives are less power efficient; however, smaller drives usually provide faster access to data.

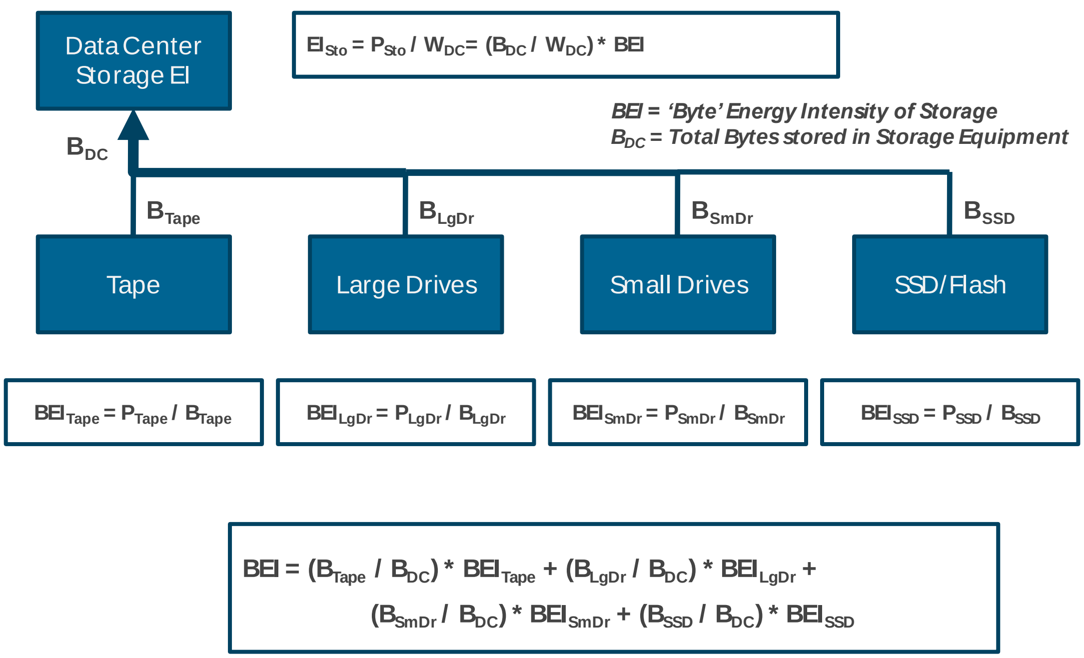

Figure 8. An Example of Tiered Storage Energy Intensity Model

[8] Actual storage energy intensity is determined by minimum storage intensity and the utilization of the storage equipment.

Tiered storage, however, provides a mechanism to leverage the strengths of each of these types of drives in improving both costs and resource consumption. The Storage Networking Industry Association's Data Management Forum (SNIA-DMF) defines tiered storage as the implementation of a hierarchical set of storage systems based on service requirements such as performance, security, retention, and cost (among others)(SNIA - Data Management Forum, 2006).

Figure 8 shows a model for calculating the energy intensity of a tiered storage architecture within a data center. Models such as these provide a means to understand either how different options for tiered storage affect energy intensity or how the energy intensity of a particular tiered storage system will change over time, as data is created and stored—and storage capacity is added.

An example of the use of this type of model to understand the potential benefits of tiered storage, with respect to power consumption and energy intensity, is provided in Figure 9. In this example, an initial implementation of 300GB Fibre Channel drives is augmented with a second tier of 300GB SATA drives as the data center's storage needs increase. Subsequent steps includes introduction of a third tier of 1TB SATA drives and, finally, replacement of the original 1TB 7200rpm drives with 2TB variable speed drives.

If only power consumption is considered, the magnitude of the improvement is hidden. While absolute power consumption goes up slightly as the second and third tiers are added, the energy intensity goes down due to the increased work that is being asked of the storage system (i.e., amount of data that is being stored). By the time that the 1TB drives are replaced with 2TB drives, the energy intensity of the storage system is less than half its original value—a significant improvement.

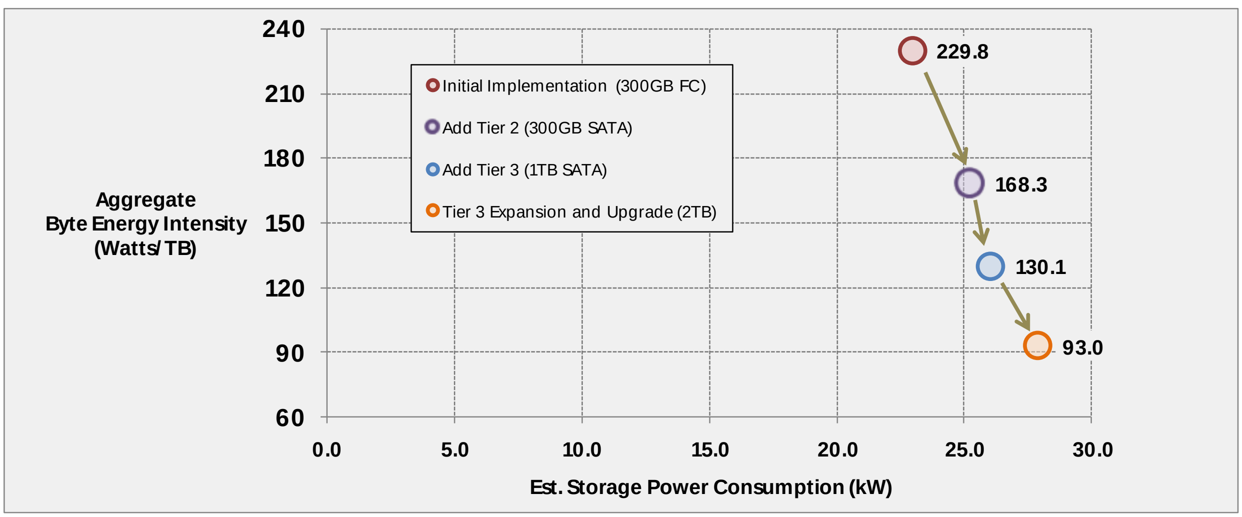

Figure 9. Changes in Energy Intensity and Power Consumption Through Tiered Storage

Thin Provisioning

Whereas data deduplication reduces the amount of stored data and tiered storage helps improve baseline power consumption per TB, thin provisioning directly affects the utilization of the data center's storage equipment. Thin provisioning is a forward planning tool that logically allocates all of the storage that an application may require upfront, while limiting the actual physical storage that is allocated until such time as it is needed (Dell, 2009). Figure 10 shows an example of the relationship between storage utilization and energy intensity. The effect of thin provisioning can be as dramatic with respect to power consumption as it is with total cost of ownership (TCO).

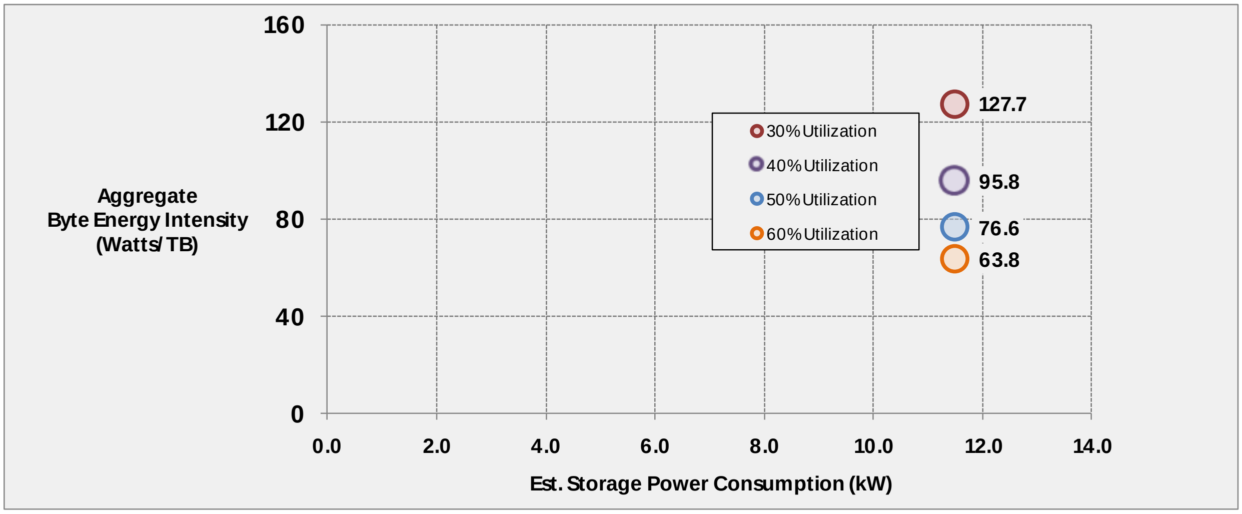

Figure 10. Effect of Utilization on Storage Energy Intensity

Conclusions

Data Collection and Metric Calculation

Certainly, the primary key to this type of analysis is the ability and opportunity to collect a meaningful and analyzable dataset. Unfortunately, this approach to estimating data center performance is new and not incorporated into today's tools. Fortunately, Dell's internal toolset of managing its data centers includes all of the different elements required to calculate an estimate of data center work output. Dell collects utilization information on thousands of servers on a regular basis.

Figure 11 shows the main classes of data that are required to run the estimate. In addition to utilization data, the IT productivity proxy requires configuration information on the deployed servers and static information estimating the maximum expected performance for each server[9]. While this dataset required some manual work to integrate the disparate data sources, the work involved is not technically difficult. If this approach provides value to the industry, it should be relatively easy for tool vendors to provide future support for these types of metrics.

In addition to the work estimate, however, calculation of energy intensity also requires collecting data center power consumption information at the same frequency with which server utilization data is collected. This may prove to be a more difficult task than estimating work output. Many larger data centers, however, have implemented power monitoring and reporting tools. In addition, their smaller counterparts are likely to have to integrate these types of tools in the future, if for no other reason than to address potential utility or regulatory reporting requirements.

---

[9] This dataset used publically available results using the SPECint2000rates and SPECint2006rates benchmarks.

The primary challenge for data centers wishing to pursue this type of analysis will be integration of data from server and data center IT management tools with data center power consumption tools.

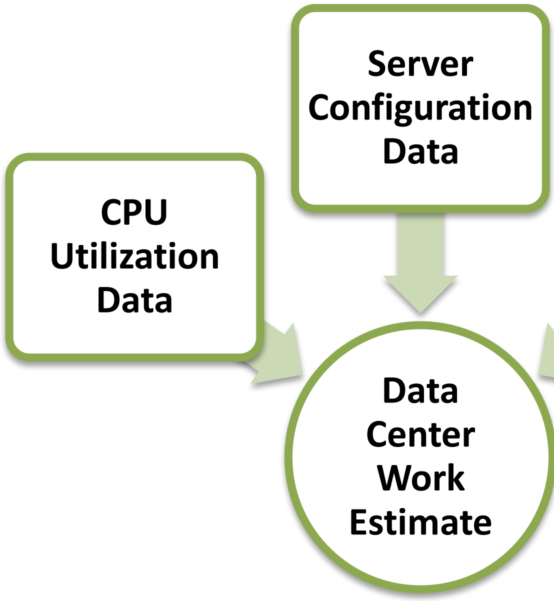

Figure 11. Data Sources for Server and Data Center Work Estimates

Future Metrics Refinements

This data represents the first known public estimate of an enterprise data center's aggregate work output. While the chart in Figure 3 looks qualitatively correct, it would be brash to assume that this is the final word on the topic.

One clue for further refinement can be found within the proxy itself. Although this is positioned as an estimate of work output, it is more accurately characterized as a measure of compute cycles. A simple CPU utilization report makes no distinction between compute cycles consumed by an application performing useful work, compute cycles consumed as a result of system or data center management requests, or compute cycles consumed on demand of the server's operating system. Tools that tie compute cycles to individual processes on the server and understand the context within which those processes operate (application, installed agent, O/S, etc.) can provide a better estimate of work output.

Additionally, compute cycles for one process or application on one machine may have greater or lesser value than compute cycles for a different process or application on a different machine. This is more within the scope of other IT productivity work being performed within The Green Grid than within the scope of this work.

The industry will also need better metrics around aggregate IT utilization. While most IT professionals are familiar with the concept of utilization as it applies to an individual piece of IT equipment, they have not had the tools to look at aggregate utilization of the entire data center. At best, overall data center utilization is calculated by collecting individual unit utilization and averaging the results. This does not take into account, however, that less powerful servers are easier to keep loaded than more powerful more recently deployed equipment.

Fortunately, approaches that estimate the work output of the data center can assist here directly. At the highest level, aggregate data center IT utilization can be thought of as the estimated work output of the data center divided by the total compute capacity of the data center[10].

$$Data\ Center\ IT\ Utilization = \frac{Estimated\ Work\ Output}{Total\ Compute\ Capacity}$$

Last, an energy intensity view of the data center, while extremely useful, provides a picture only of the amount of data processed or compute capacity provided by the data center. This metric does not take into account the inherent value of stored data.

Stored data in the data center can be viewed as a capital asset. The energy intensity metric discussed here does not help understand the value of the stored data in the data center, nor the efficiency with which the data center preserves the value of those assets. For that, the industry will need to develop additional metrics specifically focused on stored data and the energy consumed by storage equipment.

Additional Applications

One of the more intriguing possibilities for this type of analysis is the potential to compare different data centers within an organization. Today, these comparisons are difficult. A number of factors, including location, type of facility, and age of facility can affect data center performance. Without the proper approach, it may be difficult to understand why one data center outperforms another.

Because of the manner in which energy intensity can be decomposed into smaller components, it is possible to analyze and compare similar aspects. This could enable organizations to identify, for example, how location and climate affects differences in cooling architecture performance. It could also highlight how differences in server population or storage equipment populations affect IT energy intensity across different data centers.

The comparisons are important because they help to highlight what does and doesn't work for the data center. When an organization finds differences in performance across facilities (or across parts of facilities), the differences help provide focus on key decisions—does the organization invest to improve a non-performing facility, or does the organization look to leverage an efficient facility and decommission less-efficient ones?

A second interesting area of study is a look into how compute requirements are increasing over time. Data on the number of servers sold into the market is available. Data on the amount of storage sold into the market is available. What is harder to get is information on usage. How much actual compute power is being used today? How is the demand for compute growing?

As described in the paper, calculating energy intensity requires estimating data center work output. Tracking work output provides important data to the organization as to how rapidly its compute needs are increasing and how efficiently the organization is converting compute capabilities into value. This

---

[10] That is, the maximum possible performance of the data center if each piece of equipment were operating at maximum capacity.

will help IT organizations forecast and manage future IT capacity needs. It also provides additional data that can be used to refine the business case for potential service offerings.

Future Directions for Energy Intensity Analysis and Reporting

As mentioned before, this is the first public presentation and discussion of data center energy intensity. While Dell is excited about this approach, it is also open to refinements in approach and data collection. Dell expects that much of the work in this area will be performed in industry forums, such as The Green Grid.

With that in mind, there are a number of areas where future work would be valuable. First, the key piece of this work is the estimation of data center work output. Without the ability to collect meaningful data, the approach is nothing more than an academic exercise. As there are few tools and approaches for this today, validation of the data that can be collected will be difficult. The best test of the approach, however, is whether or not it is successful in providing accurate guidance. Accordingly, one key need for future study is analysis of before-and-after data where energy intensity has been measured alongside a data center initiative aimed at improving energy efficiency of some aspect of the data center.

Another important future direction for this type of analysis is improved data collecting and analysis. The dataset presented in this paper took some time to process and analyze. This delay limits the number of applications to which this type of analysis can be applied. Virtualization enables workloads to be moved across physical machines. Identification and execution of policies that dynamically move workloads in order to optimize resource consumption require real-time data.

One of the primary industry concerns with the public attention on server and data center power consumption has been a fear that regulators and policymakers might, in their need to address energy concerns, put a damper on the benefits that IT has been able to provide to the greater economy. Energy intensity provides a means for addressing issues that PUE cannot attack and for which PUE was not designed—the measurement and management of IT productivity. The approach presented here is a beginning. It offers a means to satisfy both policymakers and market participants, while ensuring that energy consumption analysis is tied back to work output.

Works Cited

Belady, C. (2007, February 1st). *In the data center, power and cooling costs more than the IT equipment it supports*. Retrieved March 4th, 2010, from Electronics-Cooling: http://www.electronics-cooling.com/2007/02/in-the-data-center-power-and-cooling-costs-more-than-the-it-equipment-it-supports/

Dell. (2010). *Intelligent Data Management*. Retrieved May 3, 2010, from Dell.com: http://www.dell.com/content/topics/topic.aspx/global/products/landing/en/idm?c=us&l=en&s=gen

Dell. (2009). *Thin Provisioning - Advanced Storage Allocation for Growing Applications*. Retrieved May 5, 2010, from Dell EqualLogic: http://www.equallogic.com/resourcecenter/assetview.aspx?id=4873

EC Joint Research Centre - Institute for Energy - Renewable Energies Unit. (2009, Oct. 30). *Code of Conduct on Data Centres Energy Efficiency Version 1.0*. Retrieved March 4, 2010, from European Commission Joint Research Centre - Institue for Energy: http://re.jrc.ec.europa.eu/energyefficiency/pdf/CoC%20data%20centres%20nov2008/CoC%20DC%20v%201.0%20FINAL.pdf

Judge, J., Pouchet, J., & Dixit, S. (2008, November). Reducing Data Center Energy Consumption. *ASHRAE Journal* , pp. 14-26.

Metcalf, G. E. (2008). An Empirical Analysis of Energy Intensity and Its Determinants at the State Level. *The Energy Journal , 29* (3), 1-26.

SNIA - Data Management Forum. (2006, January). *ILM and Tiered Storage*. Retrieved May 5, 2010, from SNIA: http://www.snia.org/forums/dmf/knowledge/DMF-ILM_and_Tiered_Storage_20060221.pdf

The Green Grid. (2008, April 29). *A Framework for Data Center Energy Productivity*. Retrieved April 30, 2009, from The Green Grid: http://www.thegreengrid.org/Global/Content/white-papers/Framework-for-Data-Center-Energy-Productivity

The Green Grid. (2009, October 7th). *Productivity Proxy Proposals Feedback - Interim Results*. Retrieved March 22nd, 2010, from The Green Grid: http://www.thegreengrid.org/en/Global/Content/white-papers/Productivity%20Proxy%20Proposals%20Feedback%20Interim%20Results

The Green Grid. (2009, January 14). *Proxy Proposals for Measuring Data Center Productivity*. Retrieved April 29, 2009, from The Green Grid: http://thegreengrid.org/Global/Content/white-papers/Proxy-Proposals-for-Measuring-Data-Center-Efficiency

The Green Grid. (n.d.). *The Green Grid*. Retrieved March 4, 2010, from The Green Grid: http://www.thegreengrid.org/

The Green Grid. (2007). *The Green Grid Data Center Power Efficiency Metrics: PUE and DCiE*. Retrieved October 28, 2008, from The Green Grid: http://www.thegreengrid.org/gg_content/TGG_Data_Center_Power_Efficiency_Metrics_PUE_and_DCiE.pdf

U.S. Department of Energy. (n.d.). *Energy Intensity Indicators: Efficiency vs. Intensity*. Retrieved March 4th, 2010, from U.S. Department of Energy - Energy Efficiency and Renewable Energy: http://www1.eere.energy.gov/ba/pba/intensityindicators/efficiency_intensity.html

U.S. Energy Information Administration. (n.d.). *Table 1.5 Energy Consumption, Expenditures, and Emissions Indicators, 1949-2008*. Retrieved March 4, 2010, from U.S. Energy Information Administration: http://www.eia.doe.gov/emeu/aer/txt/ptb0105.html

U.S. EPA. (n.d.). *Enterprise Server and Data Center Energy Initiatives - ENERGY STAR*. Retrieved March 4th, 2010, from ENERGY STAR: http://www.energystar.gov/index.cfm?c=prod_development.server_efficiency

U.S. EPA. (2007, August 2). *Report to Congress on Server and Data Center Energy Efficiency - Public Law 109-431*. Retrieved October 31, 2008, from ENERGY STAR: http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf