Power Based Performance and Capacity Estimation Models for Enterprise Information Systems

Meikel Poess
Oracle Corporation
Meikel.Poess@oracle.com

Raghunath Nambiar
Cisco Systems, Inc.
rnambiar@cisco.com

Abstract

Historically, the performance and purchase price of enterprise information systems have been the key arguments in purchasing decisions. With rising energy costs and increasing power use due to the ever-growing demand for compute capacity (servers, storage, networks etc.), electricity bills have become a significant expense for todays data centers. In the very near future, energy efficiency is expected to be one of the key purchasing arguments. Having realized this trend, the Transaction Processing Performance Council has developed the TPC-Energy specification. It defines a standard methodology for measuring, reporting and fairly comparing power consumption of enterprise information systems. Wide industry adaption of TPC-Energy requires a large body of benchmark publications across multiple software and hardware platforms, which could take several years. Meanwhile, we believe that analytical power estimates based on nameplate power is a useful tool for estimating power consumption of TPC benchmark configurations as well as enterprise information systems. This paper presents enhancements to previously published energy estimation models based on the TPC-C and TPC-H benchmarks from the same authors and a new model, based on the TPC-E benchmark. The models can be applied to estimate power consumption of enterprise OLTP and Decision Support systems.

1 Introduction

In the last decades, performance and purchase price of hardware and software were the dominant concerns of data center managers. Even though the performance of hardware and software have improved substantially, and their price have dropped significantly over the years, the competitive business environment continued to demand more and more performance and compute capacity. Consequently, over the last few years, especially due to the increase in energy prices, the cost of owning and maintaining large data centers has become a serious concern for data center managers.

Hardware and software vendors have been investing in the development of energy efficient versions of their systems. Low power versions of system components have been developed or made power aware, such as processors that are equipped with demand-driven clock speed adjustments. Reducing power consumption is also at the top of the priority list of government agencies as they challenge data center managers and system developers to reduce power consumption globally. The U.S. Environmental Protection Agency (EPA) has been

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

Table 1: Adaption Rate of Energy Reporting in TPC Results.

Benchmark	Total number of results	Results since December 2009	Results with energy Information	Adaption Rate
TPC-C	750[1]	10	3	1.3 %
TPC-E	42	14	2	14 %
TPC-H	168	10	4	40 %

1. This includes benchmark results from revisions 1, 2, 3 and 5.

working with various organizations to identify ways in which energy efficiency can be measured, documented, and implemented, not only in data centers, but also in the equipment they house [3]. Furthermore, standard organizations have responded to the growing demand for energy benchmarks such as the Transaction Processing Performance Council (TPC), the Standard Performance Evaluation Corporation (SPEC) and the Storage Performance Council (SPC) [18]. All major computer and system vendors are members of these organizations. Each of these consortia addresses different, often unique aspects of computer system performance. While SPEC and SPCs benchmarks focus on subsets of large enterprise information systems, TPCs benchmarks address complex On Line Transaction Processing (OLTP) and Decision Support (DS) systems, often involving hundreds of processors and thousands of disk drives.

The TPC offers currently two benchmarks to measure OLTP systems, namely TPC-C [23] and TPC-E [9], and one to measure DS performance, TPC-H [18]. TPC benchmarks are widely accepted in the industry for disseminating objective and verifiable performance data using well designed, long lasting benchmarks. In the last 18 years over 750 TPC-C benchmark results were published across a wide range of hardware and software platforms, representing the evolution of transaction processing systems [15]. TPC-C results were published by over two dozen unique vendors and over a dozen database platforms. In its first three years 42 TPC-E were published. In the last eight years 168 TPC-H results were published. System vendors publishing benchmark results are referred to as benchmark sponsors.

In 2009 the TPCs step to add an optional power measurement methodology (TPC-Energy [5]) uniformly to all its benchmarks marked a significant milestone towards designing and deploying standards to measure energy efficiency in servers and storage sub-systems. So far three TPC-C, two TPC-E and four TPC-H results have been published with power consumption information (see Table 1). Depending on the benchmark specification the adaption rate of benchmark publications, which include energy measurements, varies between 1.3 % and 40 %. It is expected that the adaption rate will increase over the next years, as seen with previous benchmarks.

As of now, however, the majority of TPC-C, TPC-E and TPC-H benchmark results are still published without any energy consumption information. This is not surprising because it usually takes several years for system vendors to implement and tune their systems for new benchmarks. Wide industry adaption of TPC-Energy requires a large number of benchmark publications across multiple software and hardware platforms. Meanwhile, the authors believe that previously published analytical power estimation models for x86 based systems [14],[13], which are based on nameplate power consumption, are useful tools for TPC-C and TPC-H benchmark publications, whose full disclosure report (FDR) supplies all necessary information for these models.

Each server and storage sub-system is tagged with a nameplate power rating, which is typically estimated by its manufacturer simply by adding up the worst case power, drawn by all components in a fully configured system. The purpose of nameplate rating is to aid the buyer of a particular component in provisioning the power infrastructure for it. In general, the nameplate rating is a very conservative number that is guaranteed not to be reached. Because of safety and economic reasons components commonly use only a percentage of the power reported in their nameplate specification during maximum load. Estimates of this percentage vary, but 20 % to 30 % are not uncommon. Personal computers are even reported to use two to four times less power than specified in their nameplates. Hence, provisioning power based solely on the nameplate specification results in drastically over sizing power infrastructure and supporting systems [11].

In this paper we present updated versions of our analytical power estimation models for OLTP and DS sys-

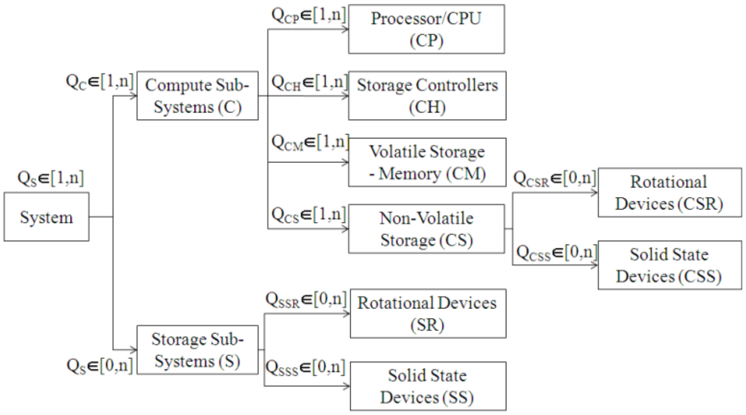

Figure 1: Hierarchy of System Components of the Analytical Power Estimation Model

tems, which are based on TPC-C and TPC-H benchmarks. Recent trends, such as the use of solid state memory technologies and in-memory database technologies due to the deployment of very large memory configurations require some modifications to our models, which were previously published in [14] and [13]. Additionally, we extend our suite of analytical power estimation models with one that is based on the TPC-E benchmark. We verify our models with measurements taken from fully scaled, optimized and published TPC-C, TPC-H and TPC-E configurations, including client systems, database server, and storage sub-system. Although estimated, the numbers obtained from our models are very close to those of the measured configurations. Hence, they can be applied to historical TPC data to perform power consumption trend analysis, they can help identifying the most power intensive system components, they can enhance dozens of performance and sizing tools that are based on TPC workloads and, to some extent, they can also be applied by data center designers to estimate power consumption of OLTP and DS systems that show system utilizations similar to those of TPC-C/E and TPC-H.

The remainder of this paper is structured as follows. Section 2 introduces a generalized power consumption model that can be applied to all current TPC benchmarks. This model generalizes and refines the previously published power estimation models [14] and [13]. Section 3 shows how this model can be applied to OLTP workloads and Section 4 shows how it can be applied to decision support workloads. Section 5 validates the power estimation models using TPC benchmark publications. The paper is summarized in Section 6.

2 Power Estimation Using Nameplate Information

Our analytical power consumption models, presented in [14] and [13], are based on the assumption that the peak power consumption of an entire system during steady state can be derived from the aggregate of the nameplate power consumptions of its individual components. Each model follows the same general approach: The nameplate power information of major system components, such as processor (CPU), volatile storage, internal non-volatile storage devices, i.e. rotational disks and solid state memory, and external storage sub-systems, i.e. enclosures with non-volatile storage devices, are aggregated discounting the nameplate overhead. Additional power of supporting components, such as motherboards and fans, is calculated with a combination of a fixed overhead and a percentage of the power consumption of the components they support. The models do not account for the power necessary for the air conditioning systems of data centers.

Figure 1 shows the hierarchy of system components that are used in our power estimation models. Each component in this hierarchy is abbreviated with up to three capital letters, as indicated in parenthesis. TPC systems may consist of two types of sub-systems, namely compute sub-systems (C) and storage sub-systems (S). In case of a clustered system and systems that have multi-tier architectures, there can be multiple compute sub-systems. We also refer to the compute sub-systems as servers. Each server consists of one or more compute units (CP), i.e. processors/CPUs, a number of storage controllers to connect to external storage enclosures (CH), some sort of volatile storage, usually memory DRAM DIMMs (CM), some non-volatile memory (CS), traditionally rotational devices (CSR), but recently also Solid State Devices (CSS) and supporting components, such as the main board and cooling fans. We also refer to the supporting devices of servers as chassis. The storage sub-system (S), which is used to store data persistently, consists of non-volatile storage devices, traditionally rotational devices (SR), but recently also Solid State Devices (SS). We also refer to the storage sub-systems as storage enclosures.

Each of these components may occur multiple times in a system and each occurrence may have different nameplate characteristics. Hence, we enumerate them with an index on each level. For instance, the second CPU in the first compute sub-system is labeled CP$_{1,2}$. The 5th rotational device in the second supporting component of the first storage sub-system is labeled SSR$_{1,2,5}$. We refer to the quantities and power consumptions of these components with Q and P respectively. E.g., the number of CPUs in the first compute sub-system is $Q(C, P)$ and the power consumption of the second CPU in the first compute sub-system is $P(CP_{1,2})$.

There are two key requirements that need to be met for our power models to correctly estimate power consumption. Firstly, only workloads that observe steady state can be used. The second requirement is system balance. Depending on the application and system, an optimal component ratio has to be maintained to keep all components (CPU, disks, controllers etc.) utilized during the measurement interval. If a system does not have the optimal ratio between these components, the power consumption model will not produce accurate estimates for the same reason that the system needs to be fully utilized. For each of the components listed in Figure 1 we determine its peak power consumption as follows.

2.1 Power Consumption of Compute Units

We obtain the peak power consumption of compute units, i.e. processors/CPUs, from their manufacturers specifications [10]. The power consumption is usually specified as Thermal Design Power (TDP). Table 2 shows the peak power consumption of selected CPUs. They range from 50W to 150W. The power consumption of all processors in compute sub-system i can be calculated with the aggregate of the nameplate power consumption of all individual CPUs:

$$P(CP_i) = \sum_{j=1}^{Q(CP_i)} Q(CP_{i,j})$$

2.2 Power Consumption of Storage Controllers

The power consumption of all storage controllers can be estimated using the maximum power consumption as defined in the Peripheral Component Interconnect (PCI) standard. It specifies that a PCI card draws at most 25 W of power. The power consumption of all storage controllers in compute sub-system i can then be calculated multiplying the number of storage controllers by 25:

$$P(CH_i) = Q(CH_i) * 25$$

2.3 Power Consumption of Volatile Storage

Similarly to the compute node nameplate power consumption, we obtain the peak power consumption of volatile storage, usually DRAM memory DIMMs, from the manufacturers website. 3 shows the power consumption for

CPU Description	TDP [W]	CPU Description	TDP [W]
Intel Pentium III Xeon - 900 MHz	50	AMD Opteron 2.2GHz Dual Core - 2.2 GHz	93
Intel E5420	50	AMD Opteron Dual Core 1 MB L2 - 2.4 GHz	95
Intel Pentium Xeon MP - 1.6 GHz	55	Intel Xeon X5650	95
Intel Xeon MP - 1.6 GHz	55	Intel Itanium2 - 1 GHz	100
Intel Xeon MP - 2.0 GHz	57	Opteron 6176	105
Intel E5506	60	Intel Itanium 2 Processor 6M - 1.5 GHz	107
Intel E5530	60	Intel DC Itanium2 Processor 9050 - 1.6 GHz	130
Intel Xeon MP - 2.8 GHz	72	Intel Dual-Core Itanium2 1.6Ghz	130
Intel Xeon MP - 2.7 GHz	80	Intel Itanium2 - 1.6 GHz	130
AMD Opteron - 2.2 GHz	85	Intel Xeon 7350 2.93GHz	130
AMD Opteron - 2.4 GHz	85	Intel Xeon X5680	130
Intel Xeon MP - 3.0 GHz	85	Intel Xeon X5670	130
AMD 8220SE 2.8 GHz	93	Intel Xeon X5560	130
AMD Opteron - 2.6 GHz	93	Intel X7560	130
AMD Opteron - 2.8 GHz	93	Intel Xeon 7140 3.4GHz	150

Table 2: Thermal Design Power (TDP) of x86 CPUs (Intel and AMD)

DDR	Size [GByte]	Density [GByte]	TDP [W]
2	4	1	9.3
2	4	2	5.4
2	8	2	5.5
3	8	2	5.6
3	16	1	8

Table 3: Power Consumption of Volatile Storage (Memory)

various types of DRAM used in TPC benchmarks. The power consumption of the entire volatile storage in compute sub-system i can be calculated as the aggregate power consumption of all individual memory DIMMs as follows:

$$P(CM_i) = \sum_{j=1}^{Q(CM_i)} Q(CM_{i,j})$$

2.4 Power Consumption of Non-Volatile Storage

We distinguish non-volatile storage between rotational storage, i.e. disk drives, and solid state storage, i.e. SSDs. Peak power consumption levels of disk drives vary widely with the disks form factor, size, and rotational speed. 5 summarizes the peak power consumption of disk drives used in TPC benchmark publications. The energy consumption of SSDs depends on the underlying technology. Most SSDs use NAND-based flash memory, a non-volatile chip that can be electrically erased and reprogrammed. There are two different types of SSDs: Single-level cell (SLC) and multi-level cell (MLC). One major difference between these technologies is that SLC hold one data bit while MLC hold two data bits. SLC provides higher write performance and reliability while MLC allow for higher storage density and lower cost. Most of todays SSDs use the same interface as traditional hard disk drives, namely SAS or SATA, so supported in traditional SAS/SATA storage arrays. Another type of SSDs that is becoming popular is PCI Express-based flash storage card (e.g. ioDrive from Fusion-io and WrapDrive from LSI. Todays PCI Express-based flash storage cards can hold between 160 GByte to 1.2 TByte. Their nameplate power consumption is 25 Watts, equal to the nameplate power consumption of the PCI slot. The nameplate power of SSDs used in TPC benchmarks are listed in Table 4. They are obtained from manufacturers web sites [20]. FF refers to the form factor of the drive. The power consumption of the entire non-volatile storage in compute sub-system i can be calculated as the aggregate power consumption of its rotational devices and solid state devices:

Type of Device	Size [GByte]	Power [W]
SATA 2.5 inch Form Factor Solid State Drive	60	2
SATA 2.5 inch Form Factor Solid State Drive	120	2
PCI Express-based flash storage cards	160-1200	25

Table 4: Power Consumption of Solid State Devices

FF=2.5		FF=2.5		FF=3.5		FF=3.5		FF=3.5	
RPM=10K		RPM=15K		RPM=7.2K		RPM=10K		RPM=15K	
GByte]	[W]	[GByte]	[W]	[GByte]	[W]	[GByte]	[W]	[GByte]	[W]
36	17	36	10.0	240	11.35	9	9.7	18	13.2
36	7.2	36	9.2	465	13	9	10.0	18	10.0
72	8.4	72	9.2	500	3.5	36	12.5	18.2	9.7
73	10.5	146	5.7			36	10.0	32	10.0
146	10.0					72	12.6	36	14.5
146	9.0					73	11.0	36	15
300	4.8					146	14.2	72	13.2
						146	11.4	73	16.2
						160	12.8	146	14.2
						250	11.35	146	19.0
						300	16.4	300	17.6
								300	14.4

Table 5: Power Consumption of Rotational Devices

$$P(CS_i) = \sum_{j=1}^{Q(CSR_i)} P(CSR_{i,j}) + \sum_{j=1}^{Q(CSS_i)} P(CSS_{i,j})$$

Similarly, the power consumption of the entire non-volatile storage in storage sub-system i can be calculated as the aggregate power consumption of all disk drive devices and solid state devices in all storage enclosures:

$$P(SS_i) = \sum_{j=1}^{Q(CSR_i)} P(CSR_{i,j}) + \sum_{j=1}^{Q(CSS_i)} P(CSS_{i,j})$$

2.5 Power Consumption of Compute Sub-System (servers)

In addition to compute units, storage controllers, volatile and non-volatile memory, we need to account for the supporting components of the compute sub-system to estimate their total power consumption. Supporting components are the main board, cooling fans, caches etc. They are also referred to as the Server Chassis. Recent studies [7] and [21] suggest that the power consumption of the server chassis can be expressed as a percentage (30 %) of the nameplate power consumption of its main components plus a fixed overhead (100W). Hence, we compute the power consumption of the entire compute sub-system i as follows:

$$P(C_i) = [P(CP_i) + P(CM_i) + P(CS_i)] * 1.3 + 100$$

2.6 Power Consumption of Storage Sub-Systems (Enclosure)

Similar to the power estimate of the compute sub-systems chassis case, we approximate the power consumption of the disks enclosures as a percentage of the nameplate power consumption of its main components. In our model we express it as 20 percent of the aggregate power consumption of all non-volatile storage devices. [7] and

[21] assume that the enclosures are fully populated. This is true for the majority of TPC results. Consequently the power consumption of an entire storage sub-system that contains only fully populated enclosures can be calculated as follows:

$$P(S_i) = [P(SS_i)] * 1.2$$

With the introduction of solid state drive technologies the above assumption is not true anymore and the power overhead of storage sub-systems might be under estimated using Formula 6. This is due to the increased performance of SSDs compared to rotational devices. Benchmark sponsors are able to substitute up to 12 rotational devices with one SSD using existing disk enclosures. Since the bandwidth of disk enclosures is limited by their controllers, only few of their slots are filled. Still the power infrastructure of these storage enclosures has been sized for many more devices. In order to accurately estimate the overhead of disk these enclosures, we need to assume that each enclosure is fully populated (Q$_{max}$) with rotational devices using the maximum allowable power for this storage device (P$_{max}$). The power consumption of storage enclosures that use some or only SSDs, can be estimated as:

$$P(S_i) = P(SS_i) + 0.2 * \sum_{j=1}^{Q_{max}(SS_i)} P_{max}(SS_i)$$

2.7 Power Consumption of the Entire System

Finally, the total power consumption of the entire system (P$_S$) can be estimated as the aggregated power consumption of the compute and storage sub-systems discounted by a factor of 0.2, which has also been validated in [14] and [13].

$$P = [\sum_{i=1}^{Q(C)} P(C_i) + \sum_{i=0}^{Q(S)} P(S_i)] * 0.2$$

3 Energy Consumption of On Line Processing Systems

Online Transaction Processing (OLTP) systems facilitate and manage transaction-oriented applications, typically for data entry and retrieval systems, used in industries such as banking, airlines, mail-order, supermarkets, and manufacturing. While some understand a transaction in the context of computer or database transactions, the TPC defines it as a business or commercial transactions. The TPC defines two benchmarks, TPC-C and TPC-E to measure the performance of large scale transactional systems.

3.1 The TPC-C Benchmark

TPC Benchmark C (TPC-C) [23] models an On Line Transaction Processing (OLTP) workload. In order to keep up to date with technology and system requirements in general, TPC-C has undergone three major revisions since its establishment in 1992. The first two revisions, published in the first 18 month were comparable due to their minimal effect on existing results. While revision 4 failed to get the necessary support revision 5 of TPC-C was approved in October 2000. This revision contained substantial changes with a large impact on existing results, which made it non-comparable to previous revisions under the rigid TPC rules. However, all revisions fulfill the requirements of the power consumption model. During its tenure each revision has been accepted in the industry as the most credible transaction processing benchmark with a large body of results across all major hardware and database platforms. Modeled after actual production applications and environments, it

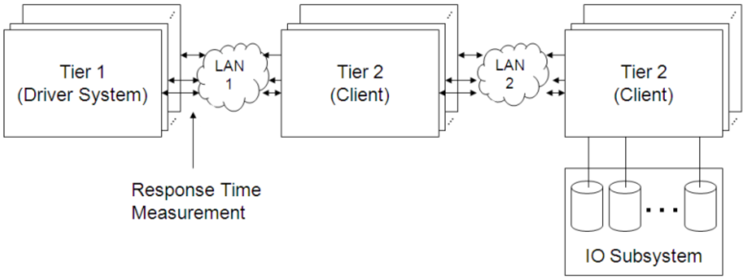

Figure 2: Typical TPC-C System Setup

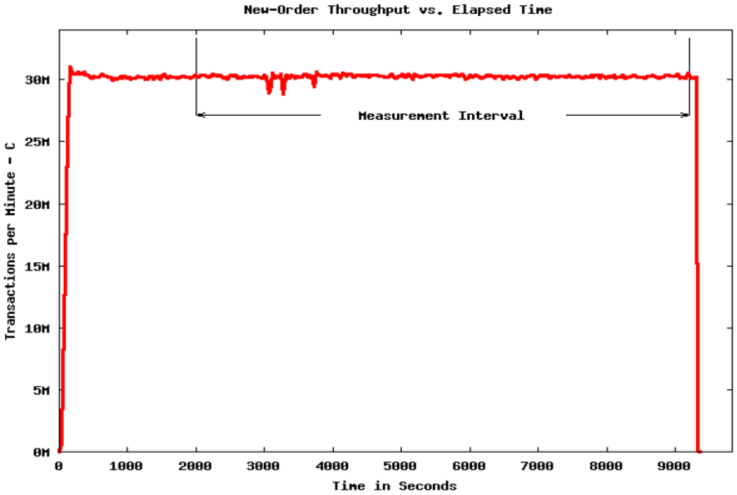

Figure 3: Throughput versus Time: TPC-C Publication 107111201

evaluates key performance factors such as user interface, communications, disk I/Os, data storage, and backup and recovery using a mixture of read-only and update-intensive transactions.

The typical TPC-C system is designed in three tiers: Tier 1, Driver System, Tier 2 Client and Tier 3 Database Server (see Figure 2). The Driver System emulates the user load.

The TPC-C performance reported in a benchmark publication is the transaction throughput of new orders during steady state condition [tmpC]. The performance is measured during the measurement interval, which must begin after the system reaches steady state, be long enough to generate reproducible throughput results that would be representative of the performance that would be achieves during a sustained eight hour period and extend uninterrupted for a minimum of 120 minutes. Another important metric of TPC-C benchmarks is price-performance. The price-performance metric [$/tpmC] is calculated by dividing the three-year cost of ownership of all components by the tpmC. See TPCs pricing specification [22] for how the three-year TCO is calculated.

Figure 3 graphs the number of new order transaction [tpmC] during the measurement interval as achieves by the current[2] performance leader.

[2] as of 1/4/2011 (see http://www.tpc.org/tpcc/results/tpcc_perf_results.asp)

3.2 The TPC-E Benchmark

TPC-E [30], approved in February 2007, is the next generation OLTP benchmark in TPCs benchmark suite. It implements a database-centric benchmark that:

1. Provides comparable performance. That is, performance from different vendors can be compared.
2. Implements an easy to understand business model.
3. Reduces the cost and complexity of running the benchmark, compared to the TPC-C benchmark.
4. Implements a complex database schema
5. Encourages database uses that are representative of client environments.
6. Leads to realistic benchmark implementations. That is, the software and hardware configuration used in the benchmark are similar to those actual end-users would use.

The TPC-E benchmark simulates the OLTP workload of a brokerage firm. Like TPC-C, the focus of TPC-E is the performance measurement of the central database that executes transactions related to the firms customer accounts. Although the underlying business model of TPC-E is a brokerage firm, the database schema, data population, transactions, and implementation rules have been designed to be broadly representative of modern OLTP systems.

TPC-E is similar to TPC-C in the following ways. The TPC-E configuration is a 3-tier model similar to TPC-C, with similar configuration and run rules (see 2). The primary metrics for TPC-E are tpsE, $/tpsE and availability date, which correspond to TPC-Cs tpmC, $/tpmC, and availability date. As in TPC-C the performance of TPC-E is measured during the measurement interval, which must begin after the system reaches steady state. TPC-C and TPC-E use a continuous scaling model and portions of the database scale in a linear fashion while the transaction profile is held constant.

Figure 4 graphs the throughput versus elapsed wall clock time of the current[3] performance leader of TPC-E [29].

3.3 Power Consumption Models for TPC-C and TPC-E

The general power estimation model, described in Section 2, can be applied to TPC-C and TPC-E benchmark publications because their measurement intervals are taken during steady state performance and because all components are fully utilized. The typical business objective of any TPC benchmark publication is to demonstrate performance and price-performance. Hence all TPC publications maintain optimal component ratios. This is because no vendor can afford to over-configure one part of the system because all parts that are used in a benchmark publication need to be disclosed and priced. And price-performance is widely being used by system vendors to showcase their advantages over those of their competitors. For instance, if a vendor over-configures a database server with 50 % more CPUs, those CPUs need to be priced, and, since the number of CPUs is disclosed, the result will be used by competitors to show that they can achieve the same performance with fewer CPUs. Lastly, some database vendors tie their pricing model to the number of CPUs, while some tie it to the number of disks. This inconsistency makes it even more unattractive to publish unbalanced TPC performance results.

In order to apply the power estimation model of Section 2 to TPC-C and TPC-E systems, we need to assign the different parts of TPC-Cs and TPC-Es three tier architectures (see 2) to the components of our power estimation model (see 1). Users of todays transaction systems are most often connected through the Internet rather than through closed circuit systems. Hence, in most cases, the deployment of a transaction system does not

[3] as of 1/4/2011 (see http://www.tpc.org/tpce/results/tpce_perf_results.asp)

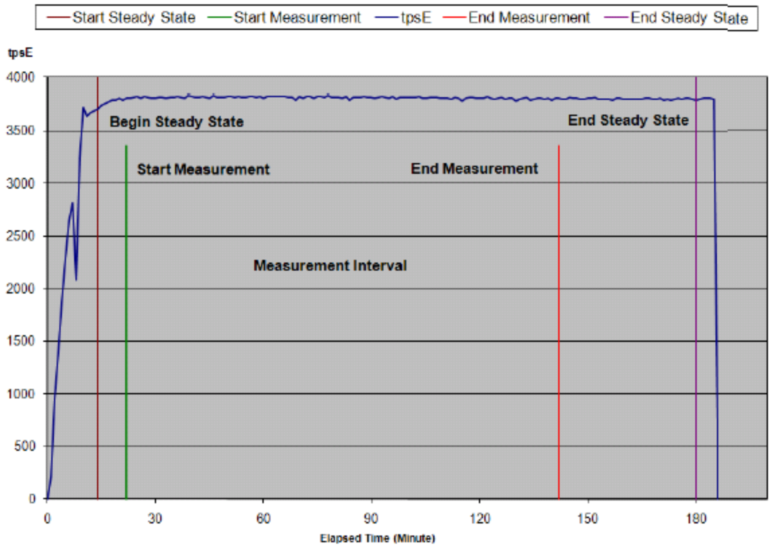

Figure 4: Throughput versus Time: TPC-E Publication 110092401

include the driver systems (Tier 1). In benchmark publications, the load imposed by users of the Tier 1 systems are emulated with far viewer systems than used in real life. Consequently, we only need to map Tier 2 and Tier 3 systems to our model. In TPC terminology, Tier 2 and Tier 3 systems are referred to as the System Under Test (SUT). This is true for TPC-C and TPC-C.

Systems that implement Tier 2 and Tier 3 can be mapped directly to the compute sub-systems of our model. The storage sub-system, which is usually connected to the database server, is mapped to the storage sub-system of our model.

4 Energy Consumption of Decision Support Systems

Generally, Decision Support (DS) workloads can be subcategorized into three distinct types of operations: Initial load, incremental load, and queries. They can be run in single- and multi-user modes. The single-user mode stresses a systems ability to parallelize operations such that the answer for a given request can be obtained in the least amount of time, as desired for overnight batch job processing. The multi-user mode stresses the systems ability to schedule concurrent requests from multiple users to increase overall system throughput. Furthermore DS workloads differ in the degree to which the queries are known in advanced (ad-hoc vs. reporting queries). The TPC currently has one decision support benchmark, called TPC-H. It is an ad-hoc benchmark, which executes each of the operations, initial load, incremental load, and queries in both single and multi-user mode.

4.1 The TPC-H Benchmark

TPC-H uses a 3rd normal form schema of eight base tables, which are populated with uniform data, i.e. without any data skew. Each TPC-H result is obtained on a database with a specific size, indicated by the scale factor (SF). The scale factor, specified in GByte, equals the raw data outside the database. The TPC rules prohibit comparing benchmark results between scale factors. The primary performance metric is the composite performance metric (QphH). It equally weights the contribution of the single-user and the multi-user runs by calculating the geometric mean of the single- and multi-user performances. With SF being the scale factor, T(qi) the elapsed

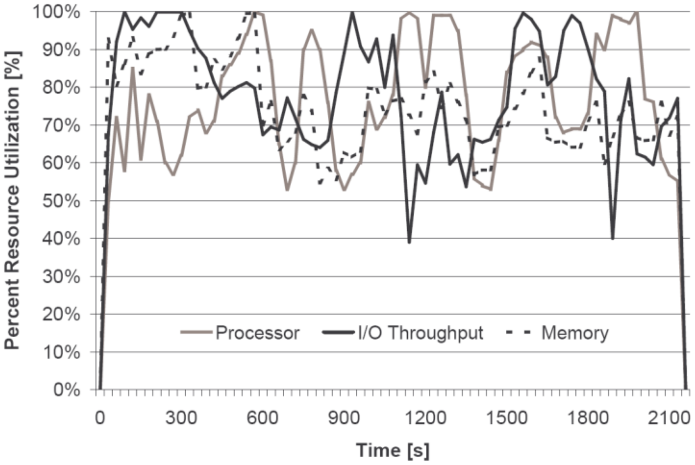

Figure 5: Oscillating Nature of the Resource Utilization During Decision Support Queries

time of Query i and T(ui) the elapsed time of Update Operation I, S the number of emulated users and Ts the elapsed time of the multi user run, the single user performance P$_{SingleUser}$ and multi-user performances P$_{MultiUser}$ are defined as follows:

$$P_{SingleUser} = \frac{3600 * SF}{\sqrt[24]{\Pi_{i=1}^{22} T(q_i) * \Pi_{i=1}^{2} T(u_i)}}$$

$$P_{MultiUser} = \frac{22 * S * 3600 * SF}{T_s}$$

$$P_{Composite} = \sqrt{P_{SingleUser} * P_{MultiUser}}$$

TPC-H 3rd Normal Form allows query execution of various execution paths. They are often dominated by large hash or sort-merge joins, but conventional index driven joins are also common. Large aggregations, which often include large sort operations, are widespread. This diversity imposes challenges both on hardware and software systems. High sequential I/O-throughput of large I/O operations is critical to excel in large hash-join operations. At the same time, index driven queries stress the I/O sub-systems ability to perform small random I/O operations. Due to the complex nature of the TPC-H queries, not all system resources, e.g. I/O, CPU, and Memory, are exhausted at any given point during query execution of the single user test. For instance, a hash join is typically CPU bound during the build phase of its hash table and, usually, I/O bound during its probe phase. Consequently, a system consumes more power in the storage sub-system during some time of the single user test and more CPU power during other times of the single user test. However, the magnitude of oscillating resources is alleviated during the multi-user mode because operations across users are usually not synchronized and, therefore, resource consumptions of concurrent queries do not align. For instance when multiple user issue queries performing hash-join operations one query might be in the build phase while another query is in the probe phase. Figure 5 shows the resource consumption of processor (CPU), I/O throughput and memory during a multi-user execution of typical decision support queries of 8 concurrent users. The CPU, I/O and memory resources are normalized by percent of their theoretical maximum value that can be achieved in this configuration. The graphs show that on average each resource is utilized about 80 percent.

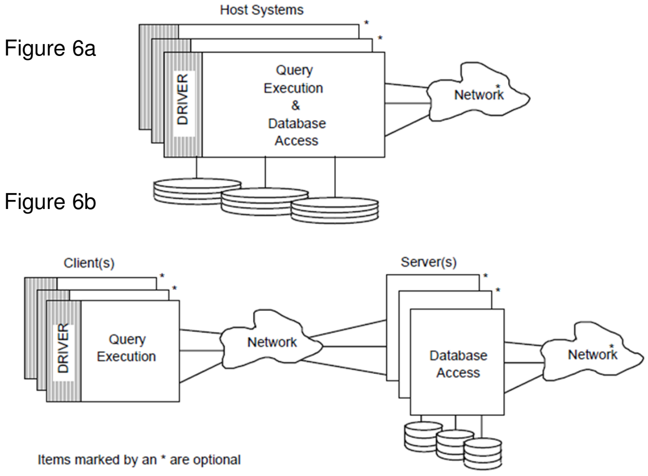

Figure 6: Typical TPC-H Configuration (source: TPC-H Specification 2.13.0)

4.2 Typical TPC-H Systems

The TPC-H specification allows two types of configurations, a host-based configuration in which the query execution and database access happens on the same system and a client/server configuration in which the query execution and database access happen on two systems that are connected via a network.

Figure 6 illustrates these two options. Figure 6a depicts the host-based solution and Figure 6b depicts the client/server solution. In both figures the driver is shown in the shaded area. The System Under Test (SUT) consists of the following components: 1

- The host and server systems (hardware and software)
- Client processing units
- Communication systems (hardware and software)
- Data storage

4.3 Power Consumption Model TPC-H

TPC-H fulfills both requirements of the power estimation model. The multi-user test of TPC-H is in a steady state. As shown in Figure 4 resources in a typical TPC-H system are used about 80 percent during the entire measurement interval. Additionally, as with TPC-C and TPC-E benchmark publications, the business objective of a TPC-H benchmark publication is to demonstrate performance and price-performance. Therefore, TPC-H systems are very well balanced.

Our generalized power estimation model can be used for both, the host-based and the client/server solutions of TPC H. For the host-based solution we only need to account for one compute systems, or multiple compute subsystems in case of clustered solutions, and one storage subsystems. In case of a client/server based solution, we have to distinguish between compute notes for the clients and server plus the storage subsystem.

Result	[26]	[25]	[24]
Measured tpmC	1,807,347	290,040	1,193,472
Measured Watts per ktpmC	2.46	4.22	5.93
Measured Power [W]	4,441.0	1,223.9	7,077.0
Estimated Power [W]	4,356.3	1,137.4	7,412.7
Total Difference [W]	-84.7	-86.5	335.7
Relative Difference [%]	-1.9	-7.1	4.7

Table 6: Power Consumption: Measured vs. Estimated values. Watts per ktpmC is a metric required by TPC-Energy. It shows how much energy is needed to run 1000 TPC-C transactions.

Result	[28]	[27]
Measured tpsE	2,001.12	1,400.14
Measured Watts per tpsE	5.84	6.72
Measured Power [W]	11683	9403
Estimated Power [W]	12065.7	9414.5
Total Difference [W]	382.7	11.5
Relative Difference [%]	3.3	0.1

Table 7: Power Consumption: Measured vs Estimated Values

5 Verification of Analytical Power Consumption Models With Fully Configured TPC-C/E and H Systems

The analytical power consumption models are verified against all TPC benchmark publications with energy metrics as of 1/21/2011: three TPC-C results, [26], [25], and [24], two TPC-E results, [28] and [27], and four TPC-H, [32],[33],[31] and [34]. Tables 6, 7 and 8 summarize the measured and reported power consumption of the entire SUT of the above nine TPC publications. Details of each benchmark configuration are summarized below.

The benchmark configuration of result [26] achieves 1,807,347 tpmC with one database server, populated with four Intel Xeon X7560 processors, 64 16 GByte memory modules, four internal disk drives, 18 disk enclosures with a total of 132 disk drives and 256 SSDs. The middle tier consists of four servers, each with two Intel Xeon X5670 processors, twelve 2 GByte memory modules and two internal disk drives. The benchmark configuration of result [25] achieves 290,040 tpmC with one database server, populated with one Intel Xeon X5650 processors, 16 16 GByte memory modules, 16 internal SSDs, three disk enclosures with a total of 25 disk drives and eight SSDs. The middle tier consists of three servers, each with one Intel Xeon E5506 processors, two 1 GByte memory modules and one internal disk drive. The benchmark configuration of result [24] achieves 1,193,472 tpmC with one database server, populated with two AMD Opteron 6167 SE processors, 16 16GByte and 32 8 GByte memory modules, two internal disk drives, 13 disk enclosures with a total of 166 disk drives and 120 SSDs. The middle tier consists of 24 servers, each with one Intel Xeon E5530 processors, two 1 GByte memory modules and 1 internal disk drive.

Each result uses a combination of rotational storage devices and SSD devices in their storage sub-system. Interestingly, although result [25] uses the least number of rotational devices compared to the others, it is not the most energy efficient result. Result [26], which is the most energy effective system with 2.46 Watts per ktpmC, uses 132 rotational devices, more than 5 times the number of rotational devices of result [25].

The power consumptions calculated with the power estimation model for TPC-C are within -7.1 % to 4.7 % of the reported power numbers.

The benchmark configuration of result [28] achieves 2,001.12 ptsE with one database server populated with four Intel Xeon X7560 processors, 64 16 GByte memory modules, four internal disk drives, 40 disk enclosures with a total of 990 disk drives. The middle tier consists of four servers, each with one Intel Xeon X5420

Result	[32] SF=100	[33] SF=100	[31] SF=300	[34] SF=300
QphH	73,975	71,438	107,561	121,346
Watts per QphH	5.93	6.48	9.58	10.33
Measured Power [W]	438	463	1031	1254
Estimated Power [W]	476.9	514.72	992.7	1141.8
Total Difference [W]	38.9	51.7	-38.3	-112.2
Relative Difference [%]	8.9	11.2	-3.7	-8.9

Table 8: TPC-H Power Consumption: Measured vs Estimated Values

processors, 12 x 2 GByte memory modules and four internal disk drives. The benchmark configuration of result [27] achieves 1,400.14 tpsE with one database server, populated with two AMD Opteron 6167 SE processors, 32 8 GByte memory modules, two internal disk drives, 28 disk enclosures with a total of 700 disk drives. The middle tier consists of four servers, each with one Intel Xeon E5420 processors, two 1 GByte memory modules and four disk drives. Result [28], which uses about 20 percent more energy (2280 Watts) than Result [27], is more energy efficient, because it achieves about 30 percent higher performance. Hence, its Watts per transaction metric is 0.88 Watts less. Overall the estimated power is very close compared to the measured power. Result [28] is 3.3 percent off (382.7 Watt), while Result [27] is 0.1 percent off (11.5).

For TPC-H we have the most number of published results available, namely four. The benchmark configuration of result [32] achieves 74,975 QphH with one database server populated with two Intel Xeon X5680 processors, 12 x 16 GByte memory modules, two internal disk drives and four SSDs. The benchmark configuration of result [33] consists of a database server populated with two AMD Opteron 6167 SE processors, 24 8 GByte memory modules three internal disk drives and four SSDs. It achieves 71,438 QphH. The benchmark configuration of result [31] achieves 107,561 QphH with one database server populated with four AMD Opteron 6167 SE processors, 16 16 GByte and 32 8 GByte memory modules, ten internal disk drives, and four PCI based flash devices. The benchmark configuration of result [34] achieves 121,346 QphH with one database server populated with four Intel X5670 processors, 16 16GByte and 48 8 GByte memory modules and eight internal disk drives. Even for TPC-H the power estimations are very close to the measured values. The relative difference ranges between -8.9 and 11.2 percent (-112.2 Watts to 38.9 Watts). The relative error margins for TPC-H seem to be higher compared to those of TPC-C and TPC-E. However, the TPC-H systems are far small compared to those of TPC-C and TPC-E. In most cases they are 10 times larger. Looking at the absolute difference rather than the relative difference the estimates of the TPC-H systems are small compared to those of the TPC-C and TPC-E systems. The estimates of TPC-C and TPC-E systems are between -86.5 and 382.7 Watts off while the estimates of the TPC-H systems are only -112.2 to 51.7 Watts off.

6 Conclusion

Historically vendors have optimized computer systems for performance and cost of ownership to be competitive in the market place. The TPC benchmarks have played a vital role by providing architecture and platform neutral metric and methodologies to measure these aspects. Recently announced TPC-Energy benchmark specification enables measurement and reporting of energy impact of performance and cost. While this new specification is expected to take several years to mature with publications across vendors and platform, we believe that analytical power estimation models based on nameplate data, presented in this paper are a useful tool for estimating and sizing TPC configurations and similar enterprise database systems.

As shown in 6 and 7 the power consumption estimates for OLTP benchmarks (TPC-C and TPC-E) are within 10 % of actual published numbers. For larger configurations, i.e. larger than 4kW peak power consumption, estimates are within 5 %. Power consumption estimates for TPC-H benchmarks are within 10 % of their measured power consumption, accept for one result (result [33]), which is 11.2 % larger than its measured power consumption (8). The authors believe that estimates within 10 % of actual power consumption meet most estimation requirements. The authors plan to keep a close eye of future benchmark publications and enhance the estimation model.

Acknowledgement

The authors would like to acknowledge Satinder Sethi and and Ray Glasstone for their invaluable comments and feedback and Heinrich Poess for his infinite support.

References

[1] 60 and 120 GB Solid State Drive http://h18000.www1.hp.com/ products/quickspecs/13415_div/13415_div.pdf

[2] A. Fanara, E. Haines, A. Howard. The State of Energy and Performance Benchmarking for Enterprise Servers. pp 52-66 TPCTC 2009

[3] A. Fanara, E. Haines, A. Howard: The State of Energy and Performance Benchmarking for Enterprise Servers. TPCTC 2009: 52-66

[4] AMD 6100 processors series specification: http://products.amd.com/en-us/opteroncpuresult.aspx?f1=AMD+Opteron%E2%84%A2+6100+Series+Processor&f2=&f3=Yes&f4=&f5=512&f6=&f7=D1&f8=45nm+SOI &f9=&f10=6400&f11=&f12=Active&

[5] E. Young, P. Cao, M. Nikolaiev. First TPC-Energy Benchmark: Lessons Learned in Practice TPCTC (LNCS Vol. 6417) 2010 136-152

[6] ENERGY STAR Program Requirements for Computer Servers (2009) available at: http://www.energystar.gov/ia/partners/product_specs/ program_reqs/computer_server_prog_req.pdf

[7] F. Xiaobo, W. Weber, and L. A. Barroso. 2007. Power Provisioning for a Warehouse-sized Computer. Proceedings of the 34th International Symposium on Computer Architecture in San Diego, CA. Association for Computing Machinery, ISCA '07. http://labs.google.com/papers/power_provisioning.pdf

[8] Fusion IO drive http://kb.fusionio.com/KB/a13/iodrive-and-iodrive-duo-power-consumption.aspx

[9] Hogan, T. "Overview of TPC Benchmark E: The Next Generation of OLTP Benchmarks" TPCTC (LNCS Vol. 5895) 2009 84-98

[10] Intel Xeon Processor Family specification: http://ark.intel.com/ProductCollection.aspx?familyID=594

[11] J. Mitchell-Jackson, J. G. Koomey, B. Nordman, and M. Blazek. Data center power requirements: measurements from Silicon Valley. Energy (Energy) ISSN 0360-5442, 28(4):837 850, 2003.

[12] M. Kunz. Energy consumption of electronic network components (English Version). Zurich: Bundesamt fr Energiewirtschaft Forschungsprogramm Elektrizitt, Basler & Hofman, November 26, 1997.

[13] M. Poess, R. O. Nambiar (2008, September). Energy cost, the key challenge of today's data centers: a power consumption analysis of TPC-C results. PVLDB 1(2). pp. 1229-1240

[14] M. Poess, R. O. Nambiar (2010, January). A power consumption analysis of decision support systems. WOSP/SIPEW 2010 pp. 147-152

[15] M. Poess, R. O. Nambiar (2010, September).Transaction Processing vs: Moores Law: A Trend Analysis. TPCTC (LNCS Vol. 6417) 2010 110-120

[16] M. Poess, R. O. Nambiar, K. Vaid, J. M. Stephens, K. Huppler, E. Haines (2009, March). Energy benchmarks: a detailed analysis. e-Energy. pp. 131-140

[17] Overview of the TPC Benchmark C: The Order-Entry Benchmark: http://www.tpc.org/tpcc/detail.asp

[18] Poess, M. and Floyd, C., "New TPC Benchmarks for Decision Support and Web Commerce". ACM SIGMOD RECORD, Vol 29, No 4 (Dec 2000)

[19] Specifications of disk drive from Hitachi: http://www.hitachigst.com/tech/techlib.nsf/techdocs/

[20] Specifications of disk drives form Seagate http://www.seagate.com/staticfiles/support/disc/manuals/, http://www.seagate.com/docs/pdf/datasheet/

[21] Suzanne Rivoire, Mehul Shah, Parthasarathy Ranganathan, Christos Kozyrakis, Justin Meza: Modeling and Metrology Challenges for Enterprise Power Management, IEEE Computer, December 2007

[22] TPC Pricing v1.5: http://www.tpc.org/pricing/spec/Price_V1.5.0.pdf

[23] TPC-C Benchmark Revision 5.9: http://www.tpc.org/tpcc/default.asp

[24] TPC-C Result 110062201 http://www.tpc.org/tpcc/results/ tpcc_result_detail.asp?id=110062201

[25] TPC-C Result 110081701 http://www.tpc.org/tpcc/results/ tpcc_result_detail.asp?id=110081701

[26] TPC-C Result 110083001 http://www.tpc.org/tpcc/results/ tpcc_result_detail.asp?id=110083001

[27] TPC-E Result 11006210 http://www.tpc.org/tpce/ results/tpce_result_detail.asp?id=110062103

[28] TPC-E Result 110062202 http://www.tpc.org/tpce/results/ tpce_result_detail.asp?id=110062202

[29] TPC-E Result 110092401 http://www.tpc.org/results/FDR/tpce/ fujitsu.RX900S1.100924.01.fdr.pdf

[30] TPC-E Version 2.12.0 http://www.tpc.org/tpce/spec/v1.12.0/TPCE-v1.12.0.pdf

[31] TPC-H Result 110062104 http://www.tpc.org/tpch/results/ tpch_result_detail.asp?id=110062104

[32] TPC-H Result 110070201 http://www.tpc.org/tpch/results/ tpch_result_detail.asp?id=110070201

[33] TPC-H Result 110071401 http://www.tpc.org/tpch/results/ tpch_result_detail.asp?id=110071401

[34] TPC-H Result 110091501 http://www.tpc.org/tpch/results/ tpch_result_detail.asp?id=110091501

[35] Transaction Processing Performance Council. http://www.tpc.org/information/about/abouttpc.asp

[36] United States Environmental Protection Agency. http://www.epa.gov/aboutepa/index.html