



**The Association of System
Performance Professionals**

The **Computer Measurement Group**, commonly called **CMG**, is a not for profit, worldwide organization of data processing professionals committed to the measurement and management of computer systems. CMG members are primarily concerned with performance evaluation of existing systems to maximize performance (eg. response time, throughput, etc.) and with capacity management where planned enhancements to existing systems or the design of new systems are evaluated to find the necessary resources required to provide adequate performance at a reasonable cost.

This paper was originally published in the Proceedings of the Computer Measurement Group's 2009 International Conference.

For more information on CMG please visit <http://www.cmq.org>

Copyright 2009 by The Computer Measurement Group, Inc. All Rights Reserved

Published by The Computer Measurement Group, Inc., a non-profit Illinois membership corporation. Permission to reprint in whole or in any part may be granted for educational and scientific purposes upon written application to the Editor, CMG Headquarters, 151 Fries Mill Road, Suite 104, Turnersville, NJ 08012. Permission is hereby granted to CMG members to reproduce this publication in whole or in part solely for internal distribution with the member's organization provided the copyright notice above is set forth in full text on the title page of each item reproduced. The ideas and concepts set forth in this publication are solely those of the respective authors, and not of CMG, and CMG does not endorse, guarantee or otherwise certify any such ideas or concepts in any application or usage. Printed in the United States of America.

Leveraging the Cloud for Green IT:

Predicting the Energy, Cost and Performance of Cloud Computing

Amy Spellmann
Optimal Innovations
amy@optimalinnovations.com

Richard Gimarc
Hyperformix, Inc.
rgimarc@hyperformix.com

Mark Preston
RS Performance
mark.preston@rsperform.com

Cloud computing is maturing, becoming a viable alternative to classic on-premise IT. Cloud facilitates scalability, promising lower fixed and variable costs while supporting enterprise growth. The scalability benefits and cost savings can be achieved through on-demand infrastructure provisioning and reduced on-premise energy consumption. The benefits are compelling; however, a quantitative analysis is required. This paper describes and demonstrates a methodology for predicting performance, energy and cost for expanding on-premise IT into the Cloud.

1 Introduction

Cloud computing, one of the current buzzwords in information technology (IT), is a viable option for addressing scalability and cost constraints that characterize today's business systems and applications. Although Cloud solutions are still in their infancy, a number of organizations are finding that off-premise Cloud is an economical alternative to traditional on-premise IT sprawl. However, before jumping on the Cloud bandwagon, there are a number of key questions that need to be answered, such as:

- Have Cloud solutions evolved enough to meet the demands of your growing enterprise?
- Is Cloud a good fit for your business from a cost, energy and performance perspective?
- Are Cloud solutions greener?

Making the right decision requires a formal evaluation to ensure that the benefits of Cloud can be quantified, evaluated and realized.

There are numerous definitions of Cloud computing. "Cloud" has the same diversity and abuse of definition that "Virtualization" experienced in recent years. As a starting point, consider this definition of Cloud and associated services [SEAR2009]:

Cloud computing is a general term for anything that involves delivering hosted services over the Internet. These services are broadly divided into three categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). ... A cloud service has three distinct characteristics that differentiate it from traditional hosting. It is sold on demand, typically by the minute or the hour; it is elastic -- a user can have as much or as little of a service as they want at any given time; and the service is fully managed by the provider.

Cloud sounds very promising, but definitely requires close scrutiny. This paper defines and demonstrates a methodology to evaluate scaling into the Cloud. In our case study, we are tasked with finding a way to scale our infrastructure to support increasing workload volume. The major constraint we have to address is that our data center is reaching its power capacity, so we must minimize any increase in its overall energy footprint. Can Cloud meet the needs of the business while limiting the expansion of the data

center? The answer is “Yes”, but there are critical factors to fully explore, especially performance and cost.

In the following sections we describe how to quantify and evaluate the risks and benefits of Cloud computing:

- Business Motivation
- Cloud Considerations
- Methodology
- Case Study

Our primary goal is to demonstrate the methodology for quantitative evaluation of scalability, performance, cost and greenness of growing into the Cloud.

2 Business Motivation

Cloud provides “on demand” compute resources which promise a more flexible, dynamic, timely, cost-effective and “greener” solution than traditional on-premise computing. Today’s business requirements necessitate highly adaptable IT services. In our current lean economic environment, companies are looking for alternatives to expanding their on-premise infrastructures. Many are not willing (or able) to make large investments to support business growth. The Cloud is ideal for enabling scalability with low entry cost and the opportunity to limit capital expenditure budgets.

By offering elasticity, pay for what you actually use, Cloud supports optimal scalability of IT resources. Traditional on-premise computing tends to contribute to IT sprawl; each business unit wants its own IT resources and must have enough capacity to handle peak workloads. This approach ultimately results in a large number of under-utilized IT resources, especially servers and storage components. With Cloud, each business unit can arrange for the timely introduction of compute resources based on current demand. Nevertheless, Cloud services must still be closely managed and monitored to ensure efficient delivery of business services.

Additionally, the elastic pay-as-you-go cost model may prove attractive to the finance department. Operating expenses (OpEx) are often much easier to secure than large capital expenditures (CapEx), especially for small to medium sized businesses. IT hardware expenditures are classified as CapEx and usually budgeted for technology refresh every 3-5 years. OpEx generally includes electric bills, rent, salaries, etc. In the Cloud model, computing costs fall under the OpEx budget, smoothing the hardware investment over time versus a large lump sum. The risk with Cloud, however, is that accumulated cost over time may far exceed comparable CapEx investments for hardware and IT resources. Accordingly, these investments must be well understood and quantified prior to committing to the Cloud.

Data centers are reaching power limits while rising energy costs are becoming a larger portion of data center budgets. Energy demand of servers drives approximately 80% of total data center IT energy load [EPA2007] [KOOM2007A]. Migrating compute tiers into the Cloud prevents an increase in on-premise servers and their associated energy footprint. The primary reasons that energy consumption is important when considering Cloud computing are [KOOM2007B]:

- Costs of in-house computing energy are not well understood, even by those inside the organization
- Cloud providers have advantages due to load diversity and economies of scale

We will not attempt to quantify the greenness of the Cloud, but assume that energy consumption can be managed more effectively by the providers.

Government programs may also affect IT budgets when considering Cloud; the Environmental Protection Agency (EPA) has recently released Energy Star ratings for servers [ESTA2009] and the US Department of Energy is offering rebates for efficient data center operation [DOE2009]. Utility companies such as Pacific Gas & Electric and Austin Energy currently have data center efficiency rebate programs in place [PGE2009] [AUSE2009]. PG&E's customer NetApp received a whopping \$1.4 million rebate for data center energy improvements [NETA2008]. Further, the US House of Representatives Climate Bill promises future assessment of carbon penalties or taxes [WSJ2009]. Cloud potentially offers benefits from rebates and penalty reductions.

In spite of all the Cloud promises, industry best practices dictate the use of a rigorous methodology that includes planning, modeling and energy footprint projections to evaluate and quantify the risks and benefits of Cloud computing. This quantitative information uniquely supports the business decision to grow into the Cloud (or not).

3 Cloud Considerations

To evaluate the risks and benefits of Cloud computing, several factors must be analyzed. The table below lists the key attributes and considerations for both traditional on-premise and Cloud computing. Attributes highlighted in bold are covered in our case study (see Section 5).

| Attribute | Traditional On-Premise | Cloud Computing |
|--------------------------------------|--|--|
| Infrastructure Scalability | <ul style="list-style-type: none"> ▪ Add/upgrade/consolidate servers (CapEx) ▪ May require data center expansion (CapEx) | <ul style="list-style-type: none"> ▪ Utilize Cloud resources to scale (OpEx) ▪ Limits data center expansion (CapEx) |
| Deployment Timeline | <ul style="list-style-type: none"> ▪ Weeks to months to acquire & deploy | <ul style="list-style-type: none"> ▪ On demand deployment can be accomplished in minutes/hours |
| Infrastructure Management | <ul style="list-style-type: none"> ▪ Focus on physical and virtualized infrastructure | <ul style="list-style-type: none"> ▪ Focus on (virtual) Cloud resources |
| Business Continuity | <ul style="list-style-type: none"> ▪ On-premise IT staff responsible for redundancy, high availability, etc. | <ul style="list-style-type: none"> ▪ Partner with Cloud provider to ensure required redundancy, availability, etc. |
| Physical Resource Utilization | <ul style="list-style-type: none"> ▪ Infrastructure must be sized to handle peak loads ▪ Generally results in low utilization during non-peak time | <ul style="list-style-type: none"> ▪ Leverage on-demand resources to dynamically scale infrastructure ▪ Eliminates the need to overbuild |
| Network Infrastructure | <ul style="list-style-type: none"> ▪ Data center IT staff responsible for networks and network expansion (CapEx) | <ul style="list-style-type: none"> ▪ Utilize Cloud network infrastructure (OpEx) ▪ Additional on-premise network resources may be required for Cloud communication (CapEx) |
| Performance | <ul style="list-style-type: none"> ▪ Response time depends on workload volumes & supporting infrastructure | <ul style="list-style-type: none"> ▪ Response time depends on the responsiveness of the virtual resources ▪ Potential new network delay between on-premise and Cloud |
| Energy | <ul style="list-style-type: none"> ▪ Energy consumption driven by IT growth, infrastructure size & supporting facility ▪ Data center facility has a fixed limit on energy consumption (power draw) | <ul style="list-style-type: none"> ▪ Energy consumption managed by the provider ▪ Limits on-premise energy growth |
| IT Budget Categories | <ul style="list-style-type: none"> ▪ CapEx for infrastructure ▪ OpEx for facility operation & energy | <ul style="list-style-type: none"> ▪ Primarily OpEx for infrastructure |

Cloud promises flexibility in allocating the IT budget; costs can be structured as upfront expenditures (CapEx) or as expenses incurred in the course of ordinary business (OpEx). Some Cloud providers now support both cash outlay timing models in which an upfront CapEx expense offsets lower recurring OpEx expenses. Again, these variations in pricing models and service delivery by Cloud providers dictate the use of a rigorous analysis process.

In addition to the above attributes, there other factors which are out of scope for this paper that may require consideration, including:

- Security
- Software licensing
- Operations staffing
- Facilities costs (e.g., leasing, etc.)
- Maintenance

Planning for business growth into the Cloud requires consideration of these key attributes to ensure adequate infrastructure, scalability, energy and cost management.

4 Methodology

Our methodology begins with the identification of potential application tiers or components that can be moved to the Cloud. Next, we determine what Cloud services can be utilized. Then, predictive models are developed to evaluate infrastructure requirements, capacity, response time, energy and cost. The methodology includes the following steps:

1. Select candidate workloads and business processes
 - Choose a service, tier or platform to move to the Cloud
 - Verify the application is Cloud ready
2. Identify candidate Cloud providers
 - What services/resources are available (compute power, network, etc.)?
 - What are the Cloud provider's performance and capacity characteristics?
 - How does the Cloud provider charge for usage?
3. Model the application (on-premise and Cloud)
 - Determine the on-premise and Cloud resources required to support growth
 - Predict resource utilization, response time and throughput
 - Project the energy footprint
 - Evaluate pricing/TCO
4. Compare performance, energy and cost factors (on-premise vs. Cloud)
 - Does Cloud reduce the energy required to support application workloads?
 - Can Cloud provide the same or comparable performance as on-premise?
 - Is Cloud cost effective for short and/or long term?

The selection of candidate workloads and business processes must be business-driven. Which application tiers are suitable for moving to the Cloud both from an architectural feasibility as well as a business process standpoint?

Choosing a Cloud provider requires a survey of those that offer services that meet the needs of the business. Furthermore, the pricing structures must be analyzed to choose the most cost-effective solution for meeting performance, cost and energy requirements.

Modeling the application requires the development of analytic or simulation models to predict capacity, response time and throughput. In order to compare traditional on-premise to off-premise, Cloud resource capacity requirements must be determined. Cloud resources are generally priced according to use, so the variability of the workload and associated resource usage must be well understood to get an accurate

prediction. Many Cloud providers also charge for network bandwidth, so the traffic in and out of the Cloud will need to be included in these models.

The energy footprint (EFP) is developed using the process described in [SPEL2008]. Energy usage is based on the amount of power consumed over a period of time (e.g., an hour - kWh). The challenge is to get reliable power usage metrics for servers. Name plate power metrics are not the best source since they represent the maximum power that could possibly be consumed (but never is). Instead, we use server power metrics, active idle (IdleW) and maximum power (MaxW), from online vendor tools (e.g., Dell Data Center Capacity Planner [Dell2009]). Since power consumption has a proven linear relationship with CPU utilization, we scale the power based on measured CPU utilization. Once the server energy footprint is determined, we augment that value with the facility overhead by multiplying by PUE (Power Usage Efficiency) to get the fully loaded power consumption.

To calculate the server energy footprint per hour, we use the following equation [SPEL2008]:

$$\text{EFP}(\text{hour}) = \text{PUE} \times \text{IT_kWh}(\text{hour}), \text{ where PUE} = 2.0$$

IT_kWh, the energy used per hour by a set of k servers, is computed using:

$$\text{IT_kWh}(\text{hour}) = \sum [(\text{MaxW}_k - \text{IdleW}_k) \times \text{Util}_k + \text{IdleW}_k] \times (1 \text{ hour})$$

A PUE of 2.0 is an industry average endorsed by [EPA2007]. We will compute the EFP for the on-premise scenario and then compare this to growing into the Cloud. The energy footprint is used to calculate energy cost as part of the total cost of ownership (TCO).

TCO is simplified in our methodology to focus on an individual application; for a full-blown data center-level analysis see [UPTM2008]. In this paper we focus on the cost of hardware, Cloud services, and energy for IT and associated site overhead. We assume that software licensing is similar for on- versus off-premise and staffing in the data center does not change significantly when moving a single application. We do, however, endorse a more complete TCO analysis for a larger scope; for example, a full data center outsourcing, to include licensing, alternative software architectures, testing, maintenance, and geographic placement.

5 Case Study

The goal of our case study is to demonstrate how to quantify the effect of leveraging the Cloud for scalability and energy efficiency. The application we chose may not be ideal for moving to the Cloud, but it provides the context to illustrate the pertinent evaluation criteria and possible risks. Our study begins with a functioning on-premise system that is expected to double in terms of workload volume. Our primary question is:

Should we expand our on-premise infrastructure to handle the expected growth, or should we leverage the Cloud to support the increased workload volume?

Assume that we will continue to use the current on-premise infrastructure. The focus is on accommodating future workload growth. Factors that will be addressed in the case study are:

- What additional on-premise infrastructure would we need to support workload growth?
- What Cloud resources would we utilize to support the additional workload volume?
- How does the cost compare; on-premise versus Cloud?
- How is our energy bill affected by growing into the Cloud?
- Is there a performance penalty for augmenting our on-premise IT infrastructure with the Cloud?
- Are we able to leverage the on-demand nature of Cloud resources?

The case study is organized as follows:

1. Describe the baseline system (our starting point for the study)
2. Determine how we can satisfy business requirements
 - Identify tiers to grow into the Cloud
 - Analyze the daily workload volume fluctuation
 - Choose a Cloud provider/solution
3. Evaluate the required on-premise infrastructure
 - Model the projected growth to determine the additional infrastructure required
 - Determine the cost for additional infrastructure
 - Compute the energy footprint projection (EFP)
4. Model/predict the effect of growing into the Cloud
 - Predict the Cloud resources required to support the projected workload volume
 - Compute the cost of moving into the Cloud
5. Compare the performance, cost and energy of the on-premise versus Cloud scenarios

5.1 Baseline System

Our case study is based on a Web-based system that resembles the TPC-W benchmark. The benchmarked system is fully described in [DELL2002] [TPC2002]. TPC-W represents a transactional Web-based application that mimics the behavior of an online retail store. Hardware tiers in the system are Web, Cache, Image, and Database servers.

Instead of starting with the hardware in the benchmarked system, we will assume that the workload and application have grown and the servers have been upgraded to more current models. Figures 1 and 2 illustrate and describe the hardware components used as a starting point for our case study.

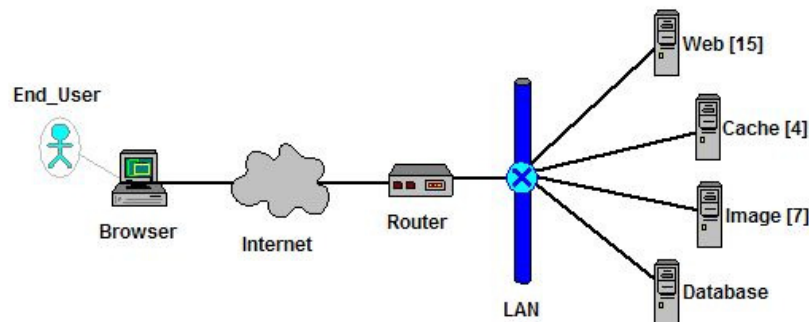


Figure 1. Baseline System Topology

| Quantity | Tier | Make/Model | TPP | Active Idle (Watts) | Max Power (Watts) |
|----------|----------|----------------------------|-----|---------------------|-------------------|
| 4 | Cache | Dell PowerEdge 1850 2.8GHz | 20 | 300 | 444 |
| 1 | Database | Dell PowerEdge 2950 1.8GHz | 81 | 254 | 333 |
| 7 | Image | Dell PowerEdge 1850 2.8GHz | 20 | 300 | 444 |
| 15 | Web | Dell PowerEdge 1850 2.8GHz | 20 | 300 | 444 |

Figure 2. Baseline Server Details

TPP represents the “Total Processing Power” for each server. TPP is a server configuration parameter in the modeling tool we used for our case study [HYPR2009]. TPP quantifies a server’s total compute power, taking into account multiple processor chips, cores, threads and their non-linear scaling factors.

The right-most columns in Figure 2 are the two power metrics for each server showing the number of watts consumed when the server is idle (Active Idle) and when it is running at maximum capacity (Max Power). Power metrics were obtained from [DELL2009].

The baseline CPU utilization of the servers is shown in Figure 3. Throughout this study we will observe a 70% utilization threshold. That is, we will attempt to maintain a maximum CPU utilization of approximately 70% for all servers.

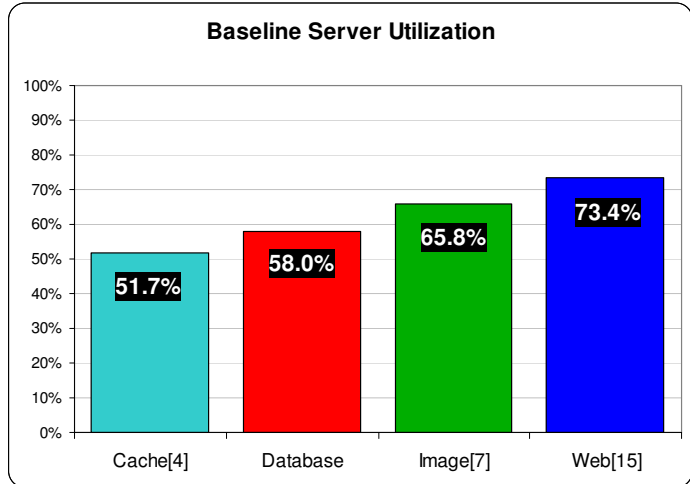


Figure 3. Baseline Server Utilization

5.2 Business Requirements

Our business is expecting significant growth over the next few months; we expect our peak workloads to double. Supporting this increase will require investment in IT infrastructure. We also have a physical constraint in the data center: we are at or near our existing power capacity. As a means of controlling our sprawling IT infrastructure, we need to evaluate supporting the expected growth by growing into the Cloud.

Since the security of our customer data is a high priority, the database will remain on-premise. We will evaluate the effect of adding the necessary Cloud resources to support a doubling of the workload for the non-database tiers. Since the database is staying on-premise, additional latency will be introduced between the database and the other tiers. This new latency will be incorporated into our analysis of response time.

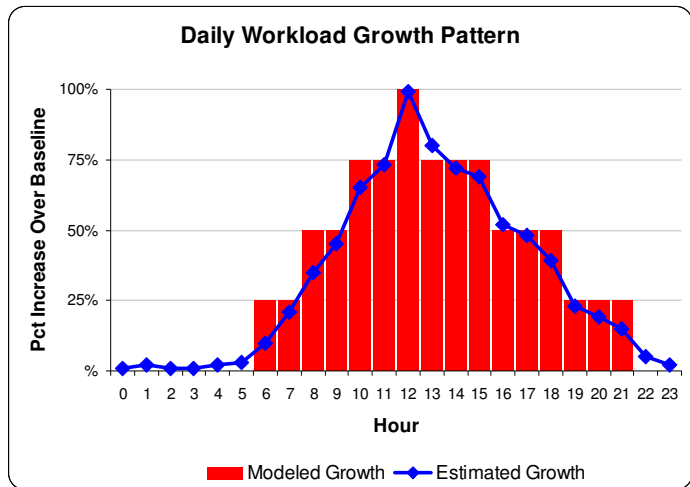


Figure 4. Daily Workload Growth

Next, we need to examine the daily and hourly workload volumes. Figure 4 illustrates how the daily workload volume fluctuates hour-by-hour over the course of a day. The line represents the historical trend of daily workload variability. To simplify our analysis, we use the bars to estimate the daily workload fluctuation. Instead of analyzing 24 different workload levels (one per hour), we will examine daily workload variability in 25% increments (only 4 levels to evaluate).

The final step in meeting our business requirements is to identify a Cloud provider. For this case study, we chose Amazon Web Services (AWS) [AMAZ2009] because of the readily available information on performance and pricing. AWS offers several different Web Services; Amazon Elastic Compute Cloud (EC2) fits our needs from a PaaS perspective. EC2 offers two types of compute instances: Standard and High-CPU. We will use Standard instances in our case study. EC2 also offers two different pricing models: On-Demand and Reserved Instances. The On-Demand feature allows us to minimize our upfront

investment, necessitated by our OpEx budget considerations. Over time, if we decided to continue to utilize EC2, Reserved Instance pricing may be more cost-effective.

5.3 On-Premise Capacity Planning

In this section we determine the on-premise hardware required to support the increasing workload volume. Figure 5 shows the model results of growing the baseline workload in increments of 25%.

These results show that all four tiers will need to be resized. We expanded the capacity of the Web, Cache and Image tiers by adding new servers. The Database was upgraded to a more powerful server, Dell's R900.

Figure 6 lists the details of the required on-premise hardware and the per-server attributes of interest for this case study. This infrastructure will serve as the on-premise comparison point for our analysis.

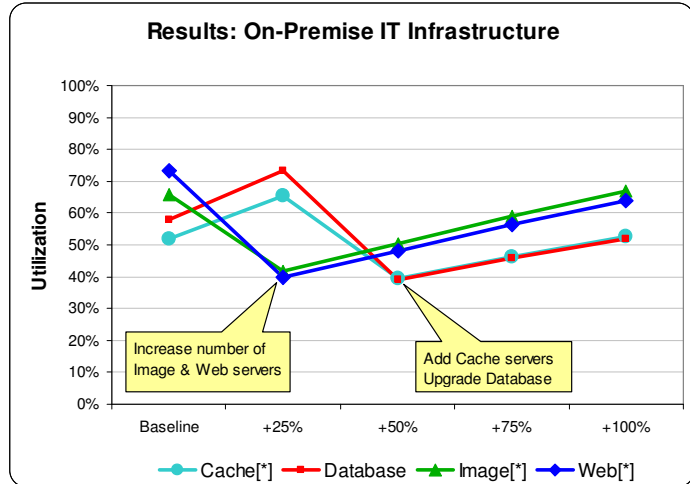


Figure 5. On-Premise Infrastructure Sizing

| Quantity | Change | Tier | Make/Model | TPP | Active Idle (Watts) | Max Power (Watts) |
|----------|--------|----------|----------------------------|-----|---------------------|-------------------|
| 8 | +4 | Cache | Dell PowerEdge 1850 2.8GHz | 20 | 300 | 444 |
| 1 | - | Database | Dell PowerEdge R900 2.4GHz | 182 | 536 | 809 |
| 14 | +7 | Image | Dell PowerEdge 1850 2.8GHz | 20 | 300 | 444 |
| 35 | +20 | Web | Dell PowerEdge 1850 2.8GHz | 20 | 300 | 444 |

Figure 6. On-Premise Sizing Details

Next, we evaluate the effect of growing into the Cloud. Instead of adding 31 new servers to our on-premise infrastructure, we will provision capacity from the Cloud. Our next step is to construct a defensible representation of the EC2 Standard Instances for our modeling tool.

5.4 Cloud Model Configurations

The three sizes of Standard instances are shown in Figure 7. The goal of this step is to create representative configurations for these instances that are suitable for use in our modeling tool.

Note the multiplicative scalability of the three Standard instances in terms of Compute Units; a Large is four times the power of a Small and an Extra Large is twice the power of a Large.

We created new server configurations in our modeling tool that are consistent with the EC2 specification (see Figure 8). We assume that a processor corresponds to a virtual core. These new configurations will be used to represent EC2 Standard servers for modeling.

| Standard Instance | EC2 Compute Units | Number of Virtual Cores | Description |
|-------------------|-------------------|-------------------------|---|
| Small | 1 | 1 | 1.7 GB of memory, 1 Amazon EC2 Compute Unit (1 virtual core with 1 EC2 Compute Unit), 160 GB of instance storage, 32-bit platform |
| Large | 4 | 2 | 7.5 GB of memory, 4 Amazon EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each), 850 GB of instance storage, 64-bit platform |
| Extra Large | 8 | 4 | 15 GB of memory, 8 Amazon EC2 Compute Units (4 virtual cores with 2 EC2 Compute Units each), 1690 GB of instance storage, 64-bit platform |

Figure 7. EC2 Standard Instances

| Standard Instance | EC2 Compute Units | Number of Processors | TPP |
|-------------------|-------------------|----------------------|-----|
| Small | 1 | 1 | 6 |
| Large | 4 | 2 | 24 |
| Extra Large | 8 | 4 | 48 |

Figure 8. Model Representation of EC2 Standard Instances

5.5 Cloud Resource Planning

The model is used to estimate the Cloud resources required to support the doubling of our baseline workload volume in increments of +25%. Our on-premise infrastructure will consist of the originally sized Web, Cache and Image tiers (15, 4, and 7 servers, respectively). For the Database, we will start with the upgraded R900 (since it is required to handle the increasing workload volume).

Figure 9 shows the hour-by-hour instance requirement (in terms of count). We applied our 70% utilization threshold to determine the number of required instances; that is, we added sufficient instances such that the CPU utilization never exceeded 70%.

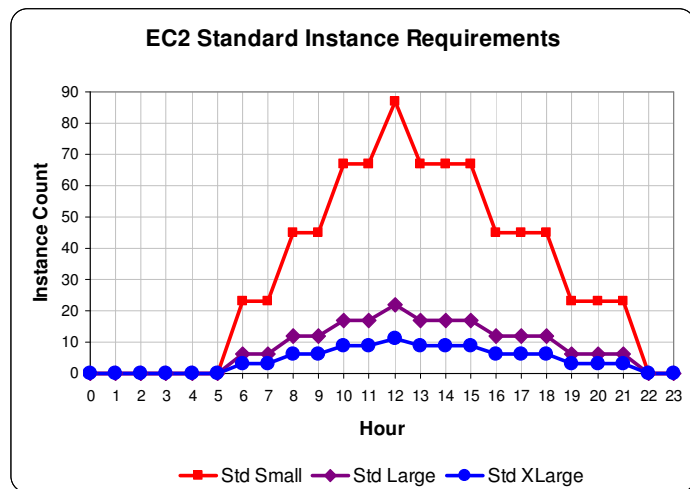


Figure 9. Standard Instance Requirements

A visual inspection of this chart confirms that we are using the proper number of instances. For example, consider the instances required for 12PM: 87 Small, 22 Large and 11 XLarge. These counts approximate the ratio of EC2 Compute Units for these instance types shown in Figure 7.

At this point, we have determined the Cloud infrastructure required to support the daily workload volume. The next step is to compute the cost of these additional Cloud resources.

5.6 Cloud Cost Estimation

We use our modeling results to estimate the cost for the three Standard instance sizes. For pricing, we compute the expected cost per month. The following table shows the monthly cost estimation for the EC2 Cloud instances.

| Standard Instance | Total Number of Hourly Instances per Day | Windows Usage per Hour | Total Monthly Cost |
|-------------------|--|------------------------|--------------------|
| Small | 762 | \$0.125 | \$2,897 |
| Large | 197 | \$0.50 | \$2,996 |
| Extra Large | 101 | \$1.00 | \$3,072 |

Figure 10. EC2 Standard Instance Pricing

Also included in AWS pricing is a Data Transfer fee. There are separate fees for transfers in and out of the Cloud. In-bound traffic has a fixed rate per GB of data transferred. Out-bound prices vary based on the amount of data transferred per month. For our application, the majority of this traffic is in and out of the Web tier in the Cloud; the traffic between the database and the Cloud is less than 5%. Figure 11 shows our analysis of Data Transfer fees. These fees are specified per month and are independent of the instance size.

| Fee Category | Price per GB | GB per Month | Total Monthly Cost |
|-------------------|--------------|--------------|--------------------|
| Transfer IN | \$0.10 | 885 | \$89 |
| Transfer OUT | | | |
| First 10 TB/Month | \$0.17 | 10,000 | \$1,700 |
| Next 40 TB/Month | \$0.13 | 6,816 | \$818 |
| Next 100 TB/Month | \$0.11 | - | - |
| Over 150 TB/Month | \$0.10 | - | - |
| TOTAL | | 17,701 | \$2,607 |

Figure 11. EC2 Data Transfer Pricing

Figure 12 summarizes the instance and data transfer costs per month. Instance costs for the three instance types are very similar since instance pricing is proportional to the number of Compute Units. Note that the smaller the instance, the more closely it is able to match the required compute power and avoid excess unused capacity.

| Standard Instance | Instance Cost per Month | Data Transfer Cost per Month | Total Cost per Month |
|-------------------|-------------------------|------------------------------|----------------------|
| Small | \$2,897 | \$2,607 | \$5,504 |
| Large | \$2,996 | \$2,607 | \$5,603 |
| Extra Large | \$3,072 | \$2,607 | \$5,679 |

Figure 12. Estimated EC2 Monthly Pricing

5.7 Energy Footprint Projections

Energy footprint projections for the on-premise and Cloud scenarios provide us with insight into the efficiency of these two solutions. Our EFP will show the kWh used per hour over the course of a day.

- Each solution includes the energy footprint of the baseline system
- For the on-premise scenario, we will add the EFPs for the 31 servers required to satisfy workload growth.
- For the Cloud solution, we assume that the on-premise EFP remains flat (same as the baseline)

The results of the hourly energy footprint projections for the on-premise scenario are shown in Figure 13. The largest consumer of energy is the Web tier, which has the largest number of servers (35). Note the gradual rise in power consumption that mirrors the daily workload fluctuation. Figure 14 shows the on-premise energy footprint when workload growth is handled by the Cloud. Here the energy footprint remains constant since our additional load is routed to the Cloud. At this point we have achieved our goal of not increasing our on-premise energy consumption by utilizing the Cloud.

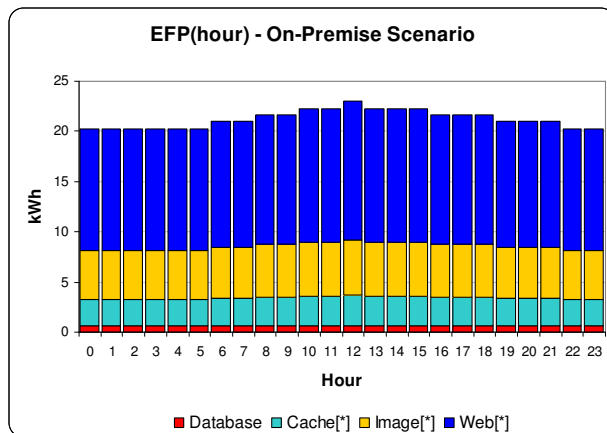


Figure 13. On-Premise Scenario - Hourly EFP

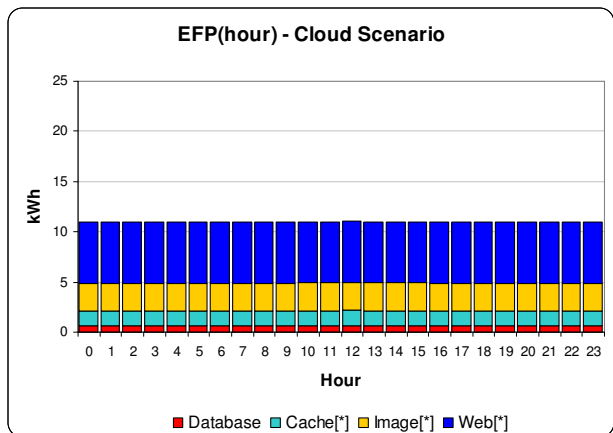


Figure 14. Cloud Scenario - Hourly EFP

6 Comparison

In this section we compare the following factors from the on-premise and Cloud scenarios:

- Energy – On-premise energy
- Performance – Modeled response time
- Cost – On-premise hardware and Cloud resources

Our analysis so far has shown that Cloud is a viable option for keeping on-premise energy consumption flat. However, performance and cost must also be evaluated to ensure they meet our business requirements. Before proceeding with Cloud, a complete TCO for both on-premise and Cloud should be performed.

6.1 Energy Comparison

Energy Footprint: Recall that one of our business goals was to limit data center power consumption. The energy footprint comparison shows that the Cloud solution helps address this goal. Figure 15 shows the hour-by-hour energy footprint for the on-premise infrastructure in both scenarios. The Standard Small Cloud energy footprint remains flat since we handle the growth in the Cloud (no increase in on-premise energy). The on-premise energy starts at a higher point since we added 31 servers to support increased workload.

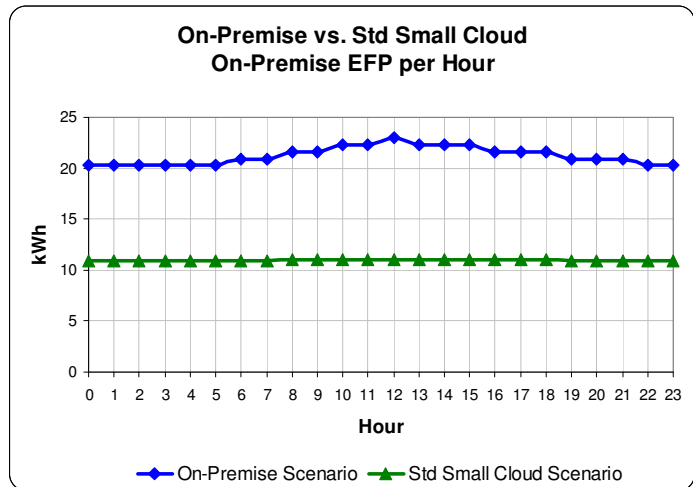


Figure 15. Comparison - EFP per Hour

Energy Cost: Figure 16 shows the cumulative monthly energy cost for the on-premise versus the Cloud scenario. We assume a price per kWh of \$0.10 based on the US commercial average for March 2009 [EIA2009]. Your results will vary according to the actual cost of electricity.

Since the Cloud scenario only includes the energy cost for a smaller on-premise infrastructure footprint, after two years it is less than half of the cost of having all infrastructure located on-premise.

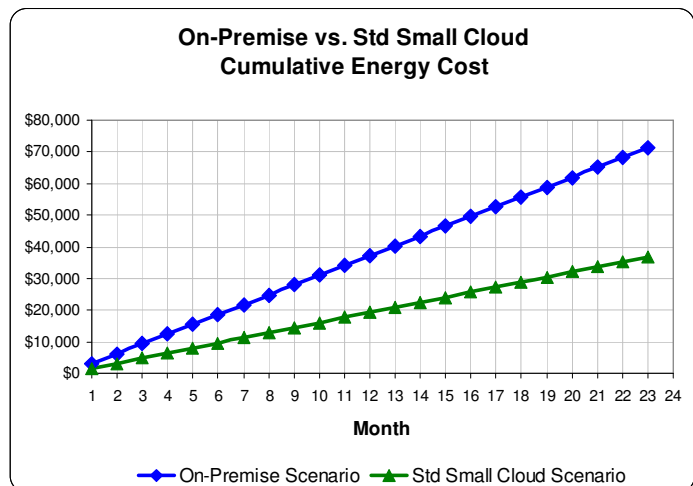


Figure 16. Comparison - Monthly Energy Cost

6.2 Performance Comparison

The response times reported in this section include CPU and network. For simplicity, we did not include database I/O response time since that was assumed to be a fixed component for both the on-premise and Cloud scenarios. Our intent is to focus on the system components that will be most affected by movement to the Cloud which are CPU and network delays.

Two separate response time analyses are performed: with and without network latency. The “without network latency” analysis provides insight into the use of Cloud instances rather than on-premise servers. The “with network latency” augments those initial results by including the time required for on-premise to Cloud communication.

Response Time without Network Latency: Figure 17 shows the model results: variability of response over the course of a day (where new Cloud instances are deployed to match the workload volume). This initial chart does not include any network latency. Variations in response times are due to:

- **Hours 0-5:** During the first 6 hours, all of the workload is handled on-premise, for both the on-premise and Cloud scenarios. The reason the on-premise scenario has a smaller response time is because the Web, Cache and Image tiers have already been scaled out to handle the peak daily workload volume.
- **Small Instances:** As the workload increases during the day, the response time using the Small instances increases dramatically. Although we have sufficient total compute capacity, the response time on the Small instances is larger than their on-premise counterpart since each individual Small instance is slower. Also, the speed of a virtual core on a Small instance is half the speed of a virtual core on the Large and XLarge instances.
- **Large & XLarge:** The effective processor speed of the Large and XLarge instances is faster than their on-premise counterparts. This is why we see a decrease in response time for the Large and XLarge instances; they are running on faster machines.

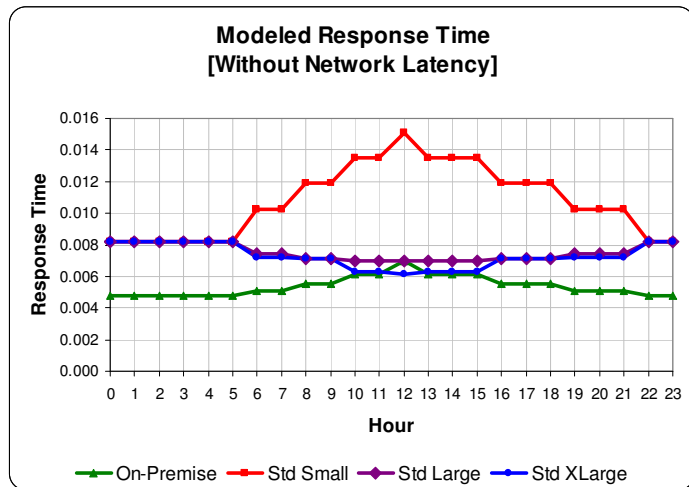


Figure 17. Modeled Response Time Without Network Latency

Response Time with Network Latency:

Figure 18 shows the modeled average transaction response time with network latency between the Cloud instances and the on-premise database server. On average, there are 1.5 database accesses per transaction. We assumed the on-premise servers are co-located, while the Cloud round-trip network latency adds 100 milliseconds (approximately the latency between Austin, TX and Washington, DC). Network latency (0.15 seconds on average) dominates the millisecond-level response time when workload grows into the Cloud. This additional latency is only applied to the transactions that are processed in the Cloud. Worst case, the database intensive transactions have approximately 5 database accesses, a 0.5 second increase in response time. More detailed modeling of the transactions would provide more granular predictions, but for an initial analysis this result provides what we need.

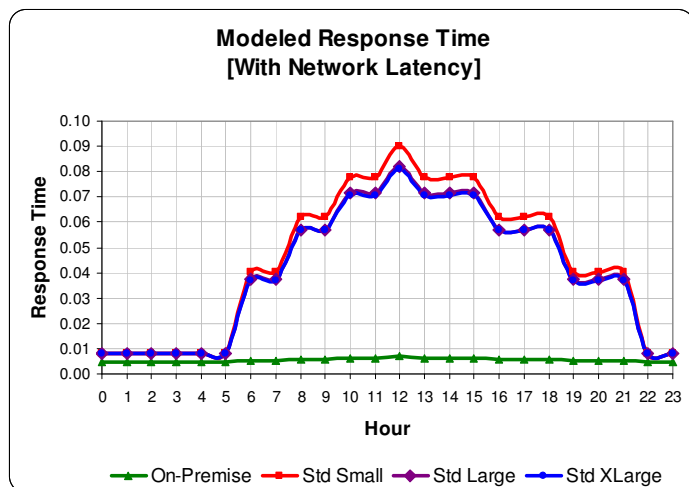


Figure 18. Modeled Response Time with Network Latency

6.3 Cost Comparison

The on-premise costs include the new hardware (31 new servers) plus the energy costs for the fully built-out on-premise server infrastructure. The Cloud cost includes on-premise energy costs for just the on-site systems plus the cost of the EC2 instances and data transfer. The instance-only view of the Cloud solution looks less expensive for two years, however, when the additional Cloud data transfer costs are included in the analysis, the Cloud solution is only cost effective for the first 12 months.

In this comparison we are only examining the EC2 Small Instances. Cloud costs for the Large and XLarge were shown to be marginally higher (less than 5%).

Total Cumulative Cost: Figure 19 shows the total cumulative monthly cost for the on-premise solution versus the Small Instance Cloud Solution. There are two cost curves for the Cloud; one that includes data transfer costs and one without.

- The on-premise solution starts off at a higher cost point. This is due to the purchase of the extra servers to handle the expected workload growth (31 new servers).
- The Cloud solution does not incur this up-front cost since it will be growing into the Cloud as needed.

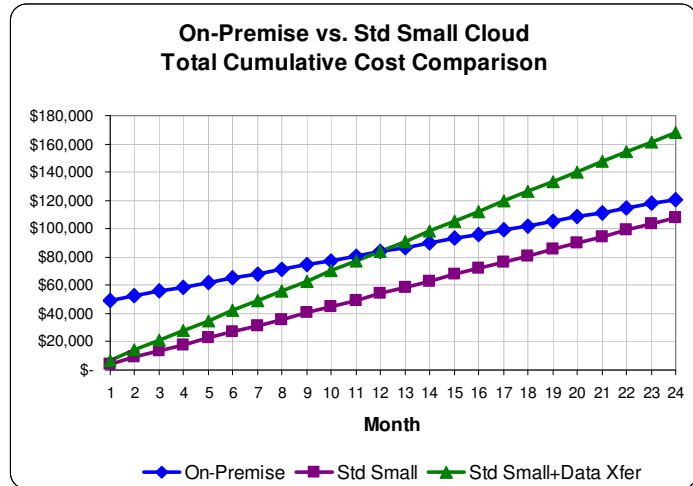
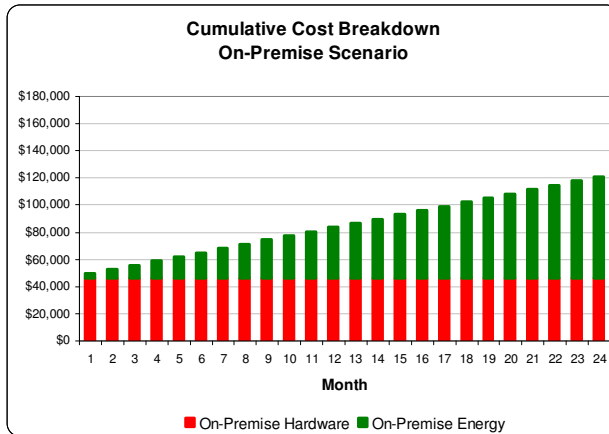


Figure 19. Total Cumulative Cost Comparison

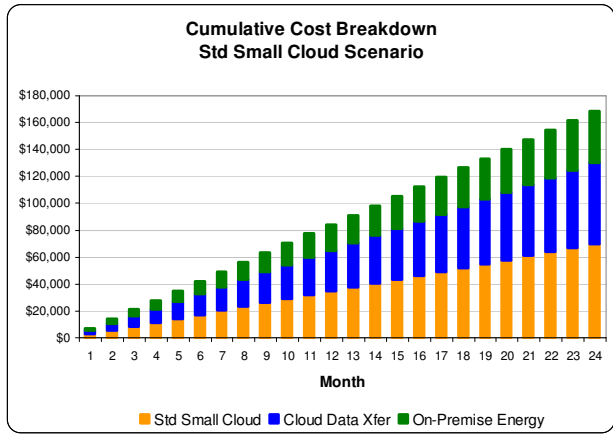
Further analysis reveals the following:

- When data transfer costs are included with instance costs, we see the breakeven point at 12 months.
- The on-premise solution may also incur additional network bandwidth costs. Further analysis and modeling is needed to understand exactly how much.
- The slope of the Cloud cost curves is steeper than the on-premise curve. This implies that the per-month expense of utilizing the Cloud is more expensive than the per-month cost of an on-premise solution.

On-Premise Cost Breakdown: The two major cost components for the on-premise solution are shown in Figure 20 (hardware and energy). We see that the initial hardware cost required to handle the increased workload volume is incurred at the beginning of the two year period and stays flat. However, the energy cost continues to increase at the rate of \$3,098 per month.



**Figure 20. On-Premise Scenario
Cumulative Monthly Cost**



**Figure 21. Std Small Cloud Scenario
Cumulative Monthly Cost**

Cloud Cost Breakdown: Figure 21 shows the breakdown of costs for the Standard Small Instance Cloud solution. There are three cost categories for the Cloud solution:

- Cloud instance cost
- Cloud data transfer cost
- On-premise energy cost (for the initial set of baseline servers)

Overall the total is increasing as all the categories increase linearly. The full-loaded monthly increase is approximately \$7,018. The on-premise energy is the smallest contributor, while the instances and data transfer are approximately 77% of the total.

7 Summary and Recommendations

The goal of the paper was to demonstrate a quantitative methodology for evaluating the impact of leveraging the Cloud. Results from the case study show that evaluating a move into the Cloud is a non-trivial task. As we saw, modeling the Cloud is a key ingredient to gaining a full understanding of the impact.

Recall the questions posed at the beginning of the case study.

1. What additional on-premise infrastructure would we need to support workload growth?

We developed a capacity plan for the on-premise infrastructure required to support the expected workload growth: upgrading the database and adding a total of 31 new Web/Image/Cache servers. See Figure 6.

2. What Cloud resources would we utilize to support the additional workload volume?

Modeling was used to evaluate the required dynamic Cloud resources in terms of Small (87 at peak hours), Large (22 at peak), and XLarge EC2 Standard Instances (11 at peak). See Figure 9.

3. How does the cost compare: on-premise versus Cloud?

Growing into the Cloud requires less up-front cost. However, the Cloud can end up costing more. You need to have a good understanding of the per-month Cloud costs. See Figure 19.

4. How is our energy bill affected by growing into the Cloud?

Growing into the Cloud enabled us to keep our on-premise energy bill flat – no increase. This is one of the key benefits of utilizing Cloud solutions. See Figures 13, 14 and 15.

5. Is there a performance penalty for augmenting our IT infrastructure with the Cloud?

As we saw in the response time analysis, there are two factors to consider: speed of the Cloud instances and additional network latency. The reduced speed of the Standard Small Instance caused an increase in response time for the Cloud-bound workload. Although there was sufficient compute capacity in the Cloud, the speed of the Small instance resulted in increased response time. It was also shown that network latency can have a significant affect on response time. Since we kept the database on-premise, we suffered an increase in response time due to the Cloud instance interactions with this server. See Figures 17 and 18.

6. Are we able to leverage the on-demand nature of Cloud resources?

We developed our Cloud plan based on the expected daily workload variability. New Cloud instances were provisioned when required, and then released. This dynamic provisioning enabled us to more closely match the infrastructure with the required workload demand. See Figures 4 and 9.

The recommendations from this paper are:

- Focus your analysis on the Cloud provider's pricing model(s). Evaluate data transfer cost and any other miscellaneous costs carefully. As we saw, the cost of Cloud data transfer was almost equal to the server instance costs.
- Don't ignore performance. Although the Cloud can provide the required capacity, you also need to evaluate the performance. Response times of the Cloud can be the critical decision factor, especially if they do not meet business requirements.
- Manage your energy footprint by leveraging the Cloud. Utilizing Cloud resources effectively caps your on-premise energy cost. The Cloud offers an alternative to building out your on-premise infrastructure to accommodate future growth. Data center power efficiency can be managed with the Cloud.
- Balance CapEx and OpEx based on your business requirements. Cloud allows you to shift large infrastructure investments into predictable recurring monthly costs.

Leveraging the Cloud supports Green IT and the requirements of a growing business. However, a quantitative analysis of the performance, cost and energy is critical for success. When considering growing into the Cloud, mitigate your risk by applying the methodology presented in this paper.

8 Acknowledgements

The authors express their thanks and appreciation to the following industry experts for their reviews and recommendations which improved the content and presentation of the material included in this paper:

- Nathan Banwart (Selexis)
- Jonathan Koomey (LBNL, Stanford University)
- John Pflueger (Dell, Green Grid)
- Jeremy Rodriguez (VMware)
- Annie Shum (2004 CMG Michelson Award Winner)

9 References

[AMAZ2009] Amazon Elastic Compute Cloud (Amazon EC2), <http://www.amazon.com/ec2>

[AUSE2009] Austin Energy Power Saver Program, <http://www.austinenergy.com/go/datacenter>

[DELL2002] TPC Benchmark W Full Disclosure Report, DELL PowerEdge 6650/1.6GHz with PowerEdge 1650/1.4GHz, May 31, 2002, www.tpc.org

[DELL2009] Dell Data Center Capacity Planner, www.dell.com/calc

- [EIA2009] Energy Information Administration, "Average Retail Price of Electricity to Ultimate Customers by End-Use, by State", May 2009, http://www.eia.doe.gov/cneaf/electricity/epm/table5_6_a.html
- [DOE2009] US Department of Energy, Energy Efficiency and Renewable Energy, Industrial Technology Program, <http://www1.eere.energy.gov/industry/datacenters/>, 2009
- [EPA2007] "Report to Congress on Server and Data Center Energy Efficiency", U.S. Environmental Protection Agency, August 2007
- [ESTA2009] Energy Star Server Specification 1.0, <http://www.energystar.gov>, May 15, 2009
- [HYPR2009] Hyperformix Capacity Manager, www.hyperformix.com, 2009
- [KOOM2007A] Jonathan Koomey, "Estimating Total Power Consumption by Servers in the World", 2007
- [KOOM2007B] Jonathan Koomey, "A Simple Model for Determining True Total Cost of Ownership for Data Centers", White Paper, The Uptime Institute, 2007
- [NETA2008] NetApp, "NetApp Receives \$1.4 Million Rebate from PG&E for Data Center Energy Efficiency", <http://www.netapp.com/us/company/news/news-rel-20081208.html>, December 8, 2008
- [PGE2009] PG&E, "Energy Savings & Rebates", <http://www.pge.com/mybusiness/energysavingsrebates/incentivesbyindustry/hightech/>, 2009
- [SEAR2009] SearchCloudComputing.com Definitions, "Cloud Computing", http://searchcloudcomputing.techtarget.com/sDefinition/0,,sid201_gci1287881,00.html
- [SPEL2008] Amy Spellmann, Richard Gimarc and Charles Gimarc, "Green Capacity Planning: Theory and Practice", CMG2008 International Conference
- [TPC2002] TPC BENCHMARK W (Web Commerce) Specification, Version 1.8, Feb 2002, <http://www.tpc.org>
- [UPTM2008] Jonathan Koomey, Kenneth Brill, Pitt Turner, John Stanley, and Bruce Taylor, "A Simple Model for Determining True Total Cost of Ownership for Data Centers", Version 2.1, March 2008, Uptime Institute
- [WSJ2009] The Wall Street Journal, "House Passes Climate Bill", <http://online.wsj.com/article/SB124610499176664899.html>, June 28, 2009