Conserving Energy by Migrating to Cloud Data Center

T.Murugesan[1]
ECE Dept
Paavai College of Engineering,
Taminadu, India
murugesan67@yahoo.com

M.Malathi[2]
CSE Dept
Indra Ganesan College of Engineering
Tamilnadu,India
malathi.con@gmail.com

Abstract **- Energy consumption from data centers doubled between 2000 and 2005 from 0.5 percent to 1 percent of world total electricity consumption. That figure, which currently stands at around 1.5 percent, is expected to rise further. The issue of surging worldwide IT-related energy consumption is both a bottom-line concern to the companies involved and, increasingly, an environmental worry. Cloud-computing companies hope to offer a solution by focusing on energy efficiency within massive data centers.In this paper, an attempt is made to study the costs associated with traditional data centers and how cloud data centers provide ways for decreasing the costs while providing areas of opportunity for increasing the energy efficiency.**

Key words - amortization, energy, cost, virtualization, consolidation.

I Introduction

A data center is a facility used to house computer systems and associated components. The main purpose of a data center is running the applications that handle the core business and operational data of the organization. It generally includes backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression) and security devices. An organization of any size will have a substantial investment in its data center .That includes buying and maintaining the hardware and software, providing the facilities in which the hardware is housed, and hiring the personnel who keep the data center running. An organization can streamline its data center by taking advantages of cloud technologies internally or by offloading the workload into the public. In section 2, various costs and solutions for reducing the costs are mentioned. In section 3, types of data centers are included. In section 4, a conclusion is drawn.

II Costs incurred by Data Centers

Generally costs associated with data centers are server costs, infrastructure costs, costs associated with power, network and operational costs.

Table 1: Guide to where costs go in the data center.

Amortized costs	Cost Component	Sub-Components
~40%	Servers	CPU, memory, storage systems
~25%	Infrastructure	Power distribution and cooling
~15%	Power draw	Electrical utility costs
~15%	Network	Links, transit, equipment
~5%	Operational	Administration, managerial

A Server Cost

The highest data center costs go to servers. For example, assuming 50,000 servers, a relatively aggressive price of $3000 per server, a 5% cost of money, and a 3 year amortization, the amortized cost of servers comes to $52.5 million dollars per year.[1].With prices this high, achieving high utilization, i.e. useful work accomplished per dollar invested, is an important goal. Unfortunately, utilization of these resources in the traditional data center can turn out to be remarkably low; e.g., 10%. The reasons for this are,

1) Uneven Application fit

A server integrates CPU, memory, network and storage components. It is often the case that the application fit in the server does not fully utilize one or more of these components.

2) Uncertainty in demand forecasts

Demands can spike quickly, especially for new services, far beyond what conventional forecasts would predict.

3) Long provisioning time scales

Purchases, whether for upgrades or new builds, tend to be large, with components bought in bulk. Infrastructure is typically meant to last very long time periods may be fifteen years. But Servers are meant to last as long as 3-5years.

4) Risk Management

If successful, a service creator might reason, demand could ramp up beyond the capacity of the resources allocated to the service. Inability to meet demand brings failure just when success is at hand.

5) Hoarding

It is easy to buy in from a service team for provisioning new resources, and less easy for returning them. Inefficiencies of this type multiply across service instances.

B Reducing These Costs by migrating to Clouds

Cloud computing has these essential characteristics which gives solutions to these issues.

1) On-Demand self –Service.

This feature enables users to use cloud computing resources as needed without human interaction between the users. And the cloud service provider.

2) Broad Network Access

Cloud computing providers deliver applications via the internet, which are accessed from a web browser, while the business software and data are stored on servers at a remote location.

3) Device and location independence

Resource pooling [2]enable users to access systems using a web browser regardless of their location or what device they are using (e.g., PC, mobile phone).

4) Rapid Elasticity

Rapid elasticity is the ability of the cloud to expand or reduce allocated resources via dynamic ("on-demand") provisioning of resources, without users having to engineer for peak loads. [3]

5) Measured service.

The services allotted can be monitored and billed based on the usage of the session.

C Cloud's role in increasing energy efficiency

1) **Standardization/consolidation** Consolidation means to reduce the number of data centers a large organization may have. It helps to reduce the number of hardware, software platforms, tools and processes within a data center. Organizations replace aging data center equipment with newer ones that provide increased capacity and performance. Computing, networking and management platforms are standardized so they are easier to manage. Cost is claimed to be reduced and in a public cloud delivery model capital expenditure is converted to operational expenditure [4]. Pricing on a utility computing basis is fine-grained with usage-based options and fewer IT skills are required for implementation (in-house).

2) **Rationalize hardware**. Underutilized or old systems should be taken out, and workloads should be shifted to more-efficient hardware. Rationalization and consolidation programs can lead to a 5% to 20% reduction in the number of servers deployed.

When considering energy, the optimum savings in consolidation might not occur at very high resource utilization rates. Another metric that has to be considered is energy per unit service provided. In their paper addressing this subject, Srikhantaiah, Kansal and Zhao state that the following factors should be considered to optimize energy usage. Design an effective consolidation strategy that takes into account the impact of consolidating applications on the key observable characteristics of execution, including resource utilization, performance and energy consumption. Consolidation methods must carefully decide which workloads must be combined on a common physical server. There exists an optimal performance and energy point which changes with acceptable degradation in performance and application mix. The optimal performance and energy point should be tracked as workloads change to improve energy efficient consolidation

3) **Virtualization** Server virtualization has evolved in providing flexibility in the utilization of resources and as an enabler of cloud computing. There is a trend to use IT virtualization technologies to replace or consolidate multiple data center equipment, such as servers. Virtualization helps to lower capital and operational expenses [5]and reduce energy consumption. Virtualization provides a level of freedom of choice for a customer and efficiencies and cost savings for cloud providers. As users become accustomed to employing cloud computing they will embrace the capability to have large numbers of servers available on demand and conversely to rapidly reduce the number of servers they are using when they are not required. Additional benefits can be obtained by developing autonomic computing and integrating change management into virtual machine model.

4) **Excess capacity** In the context of cloud computing has two interpretations. One refers to cloud-consuming organizations having to meet peak service demands on an intermittent basis such as for a holiday season. The other denotes an organization offering cloud computing to generate revenue from its excess server storage or database connectivity.

By using cloud computing to meet peak demands an organization can invoke the "pay as you go" model and not have to incur the costs of unused capacity in local hardware, including power , depreciation, personnel and software. This model is best executed if the excess capacity requirements are practicable and can be integrated into an organization's IT resource planning.

Conversely excess capacity in the cloud can be considered an advantage for a cloud provider and used to service many organizations, thus making the cloud environment more efficient by spreading costs over multiple consumers. Thus fewer resources are required to serve a variety of clients, which results in the conservation of energy and capital.

5) **Agility** Agility inside a single data center means that any server can be dynamically assigned to any service anywhere in the data center, while maintaining proper security and performance isolation between services. Unfortunately, conventional data center network designs work against agility - by their nature fragmenting both network and server capacity, and limiting the dynamic growing and shrinking of server pools.

6) **Application programming interface (API)** Accessibility to software that enables machines to interact with cloud software in the same way the user interface

facilitates interaction between humans and computers. Cloud computing systems typically use REST-based APIs.

7) **Multi-tenancy** Multi tenancy enables sharing of resources and costs across a large pool of users thus allowing for Centralization of infrastructure in locations with lower costs (such as real estate, electricity, etc.)Peak-load capacity increases (users need not engineer for highest possible load-levels).Utilization and efficiency improvements for systems that are often only 10-20% utilized.

8) **Reliability** Reliability is improved if multiple redundant sites are used, which makes well-designed cloud computing suitable for business continuity and disaster recovery.[6]

9) **Performance** Performance is monitored and consistent and loosely coupled architectures are constructed using web services as the system interface.

10) **Security** Security could improve due to centralization of data, increased security-focused resources, etc., but concerns can persist about loss of control over certain sensitive data, and the lack of security for stored kernels[7] .Security is often as good as or better than under traditional systems, in part because providers are able to devote resources to solving security issues that many customers cannot afford. However, the complexity of security is greatly increased when data is distributed over a wider area or greater number of devices and in multi-tenant systems that are being shared by unrelated users. In addition, user access to security audit logs may be difficult or impossible.

D Infrastructure Cost

A key opportunity for reducing the cost of data centers is to eliminate expensive infrastructure, such as generators and UPS systems, by allowing entire data center applications be distributed across multiple locations, consolidation of resources and virtualization. By infrastructure, we mean facilities dedicated to consistent power delivery and to evacuating heat. Drawing power from the utility leads to capital investments in large scale generators, transformers, and Uninterruptible Power Supply (UPS) systems. With typical infrastructure cost of $200M, 5% cost of money, and 15 year amortization, the cost of infrastructure comes to $18.4 million/year.

Reducing These Costs

Modern data centers try to use economizer cooling, where they use outside air to keep the data center cool [9]. They do not use chillers/air conditioners, which creates potential energy savings in the millions. What if we were to deploy networks including larger numbers of smaller data centers? Among appropriate groups of these data centers, the target is 1:N resilience at data center level, that is, the failure unit becomes an entire data center. With resilience at the data center level, layers of redundancy within each data center can be stripped out. Two rules of thumb emerge. First one being on is better than off. Given the steep fixed costs for a server installed in a data center and the server's three year lifetime, it is always better for the server to be on and engaged in revenue producing activity. The challenge is achieving again agility which is the feature of cloud computing, so that any server can be applied to any problem at hand. This enables the creation of large pools of free servers with statistical multiplexing benefits, and it eliminates the structural and risk management reasons for over-construction that lead to low server utilization. Secondly build resilience at systems level.

E Power costs

In 2007 the entire information and communication technologies or ICT sector was estimated to be responsible for roughly 2% of global carbon emissions. With data centers accounting for 14% of the ICT footprint.[10]. Given a business as usual scenario greenhouse gas emissions from data centers is projected to more than double from 2007 levels by 2020[11].

Power used by support equipment, often referred to as overhead load, mainly consists of cooling systems, power delivery, and other facility infrastructure like lighting. Google publishes quarterly actual efficiency performance from data centers in operation [14].The U.S. Environmental Protection Agency has an Energy Star rating for standalone or large data centers. To qualify for the eco label, a data center must be within the top quartile of energy efficiency of all reported facilities [15].

To track where the power goes, we postulate application of state-of-the art practice based on currently well understood techniques and implementation based on good quality but widely available equipment[16] provides a metric to describe data center Power Usage Efficiency (PUE)[12] as PUE = (Total Facility Power)/(IT Equipment Power). A state of the art facility will typically attain a PUE of 1.7[12], which is far below the average of the world's facilities but better than the best.

Reducing These Costs

Again consolidation of server's virtualization and multitenant model play important role in reducing the cost. Decreasing the power draw of each server is clearly has the largest impact on the power cost of a data center, and it would additionally benefit infrastructure cost by decreasing the need for infrastructure equipment. Those improvements are most likely to come from hardware innovation, including use of high efficiency power supplies and voltage regulation modules. Barroso and Hlzle introduced the term energy proportionality to refer to the desirable property that a server

running at N% load should consume N% power [17]. Creating servers that are closer to implementing energy proportionality would improve efficiency. One area of innovation that is impacted by networking is the idea of running the data center hotter literally reducing the amount of cooling to save money on cooling equipment and the power it consumes. Tools and techniques include raising the temperature of the data center to 75 degrees Fahrenheit, using outside air when possible as an alternative to air conditioning, setting up hot aisle/cold aisle configurations and deploying server-based energy management software tools to run workloads the most energy-efficient way.

Siting is one of the factors that affect the energy consumption and environmental effects of a data center. In areas where climate favors cooling and lots of renewable electricity is available the environmental effects will be more moderate. Thus countries with favorable conditions, such as[18] Finland[19] Sweden[20] and Switzerland[21] are trying to attract cloud computing data centers.

F **Networking Costs'** First is evaluating the location of the facility. Labor rates and the cost of energy need to be weighed against security risks. Next, the size of the site has to be measured against whether it can accommodate growth. A site must provide at least five years of capacity, including physical, electrical and networking to make the project worthwhile. Finally, in determining whether a data center overhaul is necessary, companies need to look carefully at the technical problems of integrating new components into an existing facility. They should also determine whether it is possible to keep the data center running while renovations are underway.Multiple applications run inside a single data center, typically with each application hosted on its own set of (potentially virtual) server machines. A single data center network supports two types of traffic (a) traffic flowing between external end systems and internal servers, and (b) traffic flowing between internal servers. A given application typically involves both of these traffic types. The conventional approach has the following problems that inhibit agility.

1) Static Network Assignment

To support internal traffic within the data center, individual applications are mapped to specific physical switches and routers, relying heavily on VLANs and layer-3 based VLAN spanning [19] to cover the servers dedicated to the application. While the extensive use of VLANs and direct physical mapping of services to switches and routers provides a degree of performance and security isolation, these practices lead to two problems that ossify the assignment and work against agility:

(a) VLANs are often policy-overloaded, integrating traffic management, security, and performance isolation, and (b) VLAN spanning, and use of large server pools in general, concentrates traffic on links high in the tree, where links and routers are highly overbooked.

2) Fragmentation of resources

Popular load balancing techniques, such as destination NAT (or half-NAT) and direct server return, require that all DIPs in a VIP's pool be in the same layer 2 domain [23]. This constraint means that if an application grows and requires more front-end servers, it cannot use available servers in other layer 2 domains - ultimately resulting in fragmentation and under-utilization of resources.

1) Poor server to server connectivity

The hierarchical nature of the network means that communication between servers in different layer 2 domains must go through the layer 3 portion of the network. Layer 3 ports are significantly more expensive than layer 2 ports, owing in part to the cost of supporting large buffers, and in part to marketplace factors. As a result, these links are typically oversubscribed. The result is that the bandwidth available between servers in different parts of the DC can be quite limited. Managing the scarce bandwidth could be viewed as a global optimization problem. Servers from all applications must be placed with great care to ensure the sum of their traffic does not saturate any of the network links. Unfortunately, achieving this level of coordination between (changing) applications is untenable in practice.

2) Proprietary hardware that scales up, not out

Conventional load balancers are used in pairs in a 1+1 resiliency configuration. When the load becomes too great for the load balancers, operators replace the existing load balancers with a new pair having more capacity, which is an unscaleable and expensive strategy. In order to achieve agility within a data center, we argue the network should have the following properties.

3) Location-independent Addressing

Services should use location-independent addresses that decouple the server's location in the DC from its address. This enables any server to become part of any server pool while simplifying configuration management.

4) Uniform Bandwidth and Latency

If the available bandwidth between two servers is not dependent on where they are located, then the servers for a given service can be distributed arbitrarily in the data center without fear of running into bandwidth choke points. Uniform bandwidth, combined with uniform latency between any two servers would allow services to achieve same performance regardless of the location of their servers.

5) Security and Performance Isolation

If any server can become part of any service, then it is important that services are sufficiently isolated from each other that one service cannot impact the performance and availability of another.

6) Optimal Placement and Sizing

Gartner also lists several tips for refurbishing a data center. First is evaluating the location of the facility. Labor rates and the cost of energy need to be weighed against security risks.

Next, the size of the site has to be measured against whether it can accommodate growth. A refurbished site must provide at least five years of capacity, including physical, electrical and networking, to make the project worthwhile.

Finally, in determining whether a data center overhaul is necessary, companies need to look carefully at the technical problems of integrating new components into an existing facility. They should also determine whether it is possible to keep the data center running while renovations are underway. Financial savings often follow consolidation of multiple sites into a small number of larger sites [22]. Placement and sizing of data centers presents a challenging optimization problem, involving several factors.

The first factor is the importance of geographic diversity. Placing data enters, whether mega or micro, in geographically separated areas has a number of benefits [1]. First, it helps with decreasing the latency between a data center and the user (assuming users can be directed towards nearby DCs). Second, it helps with redundancy, as not all areas are likely to lose power, experience an earthquake, or suffer riots at the same time.

The second factor is the size of the data center. As described earlier, cloud services need some number of mega data centers to house large computations. The size of a mega data center is typically determined by extracting the maximum benefit from the economies of scale available at the time the data center is designed. This is an exercise in jointly optimizing server cost and power availability, and today leads to designs with 100,000s of servers spread over 100,000s of square feet drawing 10 to 20MW of power. The third factor is network cost. One would like to place data centers as close to the users as possible while minimizing the cost and latency of transferring data between various data centers.

G) Operational staff costs

In a well-run enterprise, a typical ratio of IT staff members to servers is 1:100. Automation is partial W. [24] and human error is the cause of a large fraction of performance impacting problems [25] .In cloud service data centers, automation is a mandatory requirement of scale, and it is accordingly a foundational principle of design . In a well run data center, a typical ratio of staff members to servers is 1:1000.

1) Automating

Data center automation involves automating tasks such as provisioning, configuration, patching, release management and compliance. As enterprises suffer from few skilled IT workers automating tasks make data centers run more efficiently [26].

IV CONCLUSIONS

Cloud computing may raise privacy and security concerns, but this growing practice--offloading computation and storage to remote data centers run by companies such as Google, Microsoft, and Yahoo could have one clear advantage that is better energy efficiency, thanks to custom data centers, now rising across the country. Data center costs are concentrated in servers, infrastructure, power requirements, networking, and operational costs in that order. In this paper several approaches to significantly improve data center efficiency by migrating to clouds has been identified.

References

[1] The Cost of a Cloud Research Problems in Data Center Networks"by Albert Greenberg, James Hamilton, David A. Maltz, Parveen Patel Microsoft Research, Redmond, WA, USA
This article is an editorial note submitted to CCR

[2] Ritter, Ted. Nemertes Research, "Securing the Data-Center Transformation Aligning Security and Data-Center Dynamics,"

[3] A document from the Uptime Institute describing the different tiers (click through the download page) "Data Center Site Infrastructure Tier Standard: Topology" (PDF). Uptime Institute. 2010-02-13. Retrieved 2010-02-13.

[4] Miller, Rich. "Gartner: Virtualization Disrupts Server Vendors," Data Center Knowledge, December 2, 2008

[5] Sims, David. "Carousel's Expert Walks Through Major Benefits of Virtualization," TMC Net, July 6, 2010

[6] King, Rachael (2008-08-04). "Cloud Computing: Small Companies Take Flight". Businessweek. http://www.businessweek.com/technology/content/aug2008/tc2008083_619516.htm. Retrieved 2010-08-22.

[7] "Encrypted Storage and Key Management for the cloud". Cryptoclarity.com. 2009-07

[8] Microsoft office live. http://office.live.com

[9] "Economizer Fundamentals: Smart Approaches to Energy-Efficient Free - Cooling for Data Centers" (PDF).

[10] "Smart 2020: Enabling the low carbon economy in the information age". The Climate Group for the Global e-Sustainability Initiative. Retrieved 2008-05-11.

[11] "Smart 2020: Enabling the low carbon economy in the information age". The Climate Group for the Global e-Sustainability Initiative. Retrieved 2008-05-11.

[12] "Report to Congress on Server and Data Center Energy Efficiency". U.S. Environmental Protection Agency ENERGY STAR Program.

[13] "Data Center Energy Forecast". Silicon Valley Leadership Group.

[14] "Google Efficiency Update". Data Center Knowledge. Retrieved 2010-06-08.

[15] "Introducing EPA ENERGY STAR® for Data Centers" (Web site). Jack Pouchet. 2010-09-27. Retrieved 2010-09-27.

[16] The Green Grid. URL http://www.thegreengrid.org.

[17] L. A. Barroso and U. Hlzle. The case for energy-proportional computing. IEEE Computer, 40, 2007

[18] Canada Called Prime Real Estate for Massive Data Computers - Globe & Mail Retrieved June 29, 2011

[19] Finland - First Choice for Siting Your Cloud Computing Data Center.. Retrieved 4 August 2010

[20] Stockholm sets sights on data center customers. Accessed 4 August 2010

[21] Swiss Carbon-Neutral Servers Hit the Cloud.. Retrieved 4 August 2010.

[22] http://www.informationweek.com/news/hardware/data_centers/221901180?pgno=1

[23] KW J. Hamilton. Architecture for modular data centers. In *Third* Biemnial Conference on Innovative Data Systems, 2007.

[24] Enck et al. Configuration Management at Massive Scale: System Design and Experience. *IEEE JSAC - Network Infrastructure Configuration*, 2008

[25] Z. Kerravala. Configuration management delivers business resiliency. The Yankee Group, Nov 2002.

[26] Miller, Rich. "Complexity: Growing Data Center Challenge," Data Center Knowledge, May 16, 2007

[27] http://www.google.com/corporate/green/datacenters/