IT on a Budget: Living Within Your Means

Chris Molloy

In today's economic environment, IT organizations continue to be asked to do more with less with declining IT budgets. So what happens when your data center is full, you don't have any money to buy a new data center or new equipment, but still have growing IT requirements? You determine your growth rate and install technology that has higher capacity per unit of power, space, and cooling that exceeds the growth rate. This paper describes a method to understand IT growth and determine the effects of technology (e.g. virtualization, data de-duplication) within major IT constraints such as limited capital for new hardware or data center upgrades. It includes a supply and demand model for determining how new technologies improve the supply of data center resources in order to meet the anticipated growth demand.

1 Introduction

In August 2007, the US Environmental Protection Agency (EPA) reported to the US Congress [1] that energy consumption by IT equipment in 2006 was 1.5% of the energy produced in the US. This figure had doubled between 2000 and 2006, and was expected to double again by 2011 even taking into account current trends for higher efficiency. In addition to the growing amount of energy needed, the unit cost of that energy is also going up by double digit percentages on an annual basis. This increase in energy usage is caused by the additional IT equipment required based on business requirements. This increase can also require additional data center space construction. The combined effect of the factors above are stretching IT budgets, which have tended to be flat for many organizations, or at least flat for existing IT services.

As a result of the IT growth many companies have run out of data center space, or can anticipate when the data center will be out of space. Building a new data center is very expensive and capital intensive, characteristics which are forcing companies to evaluate alternatives in light of the current economic climate. A typical commercial grade data center of 50,000 square feet with 5 megawatts of power (average density of 100 watts per square feet) costs $10 million per megawatt or $50 million. The cost of the IT equipment that goes into the data center is typically equal to or higher than the initial build price of the data center. To be conservative, this results in $100 million to build a data center with IT equipment. The predicted energy costs over a 20 year period are roughly 4 times the initial build cost, or $200 million. Most companies are not looking forward to making that type of investment every couple years to handle their growth.

Many companies have implemented one of the industry recognized IT models for running their organization. The most popular of these models is Information Technology Information Library (ITIL) which is now at version 3. The scope of capacity management in version 3 changed to include not only traditional IT capacity planning (e.g. CPU, memory, disk, and network) but also include data center facility capacity planning (e.g. power, space, and cooling). The increase in the ITIL scope is indicative of the integration of IT and data center facilities required in the future. For example, the facilities organization can build new data centers for IT growth, but have to understand the IT equipment characteristics (e.g. increase power usage per server) in order to build the data center to handle changing IT requirements. Previous CMG papers [2,3] discuss how this should be done and the effect of "green IT" in the industry.

Fortunately, the EPA report to Congress described many technologies that can be implemented in order to delay or prevent having to build new data centers. Through the use of technology data centers can mitigate the consumption of additional energy and increase in IT equipment costs. They classified the technologies into three categories: improved operation, best practice, and state of the art. The EPA predicted that implementing best practices in existing data centers and consolidating many servers to one server could reduce energy usage by 20%. This reduction could then be reapplied to grow IT content in existing data centers without having to build new ones.

Such was the purpose of the study described in the rest of this paper. The company involved had a very large IT footprint in many data centers. We understood that continued data center growth was not affordable, so we looked to alternatives. We started making our data centers more efficient, both with facilities equipment such as chillers and power distribution units (the facilities side) and on with the IT equipment such as servers, storage, and network equipment (the IT side). We had already started virtualizing servers, and found it to be a very effective

way to liberate resources to then be redeployed. What was missing was a way to tie how much virtualization we needed to perform on the IT side in order to prevent the consumption of additional data center resources on the facilities side. We referred to this as living within our means by putting IT on a budget to stay within their current power, space, and cooling footprint. We decided to model what amount of virtualization would be needed to prevent having to build new data center space in 2010 and 2011.

2 Model Methodology – Basic Questions

In order to perform a rough order of magnitude sizing of the rate of technology adoption needed to contain projected growth, we created a spreadsheet of inputs and outputs. Each of the metrics used in the spreadsheet was assigned a fixed value.
The basic questions we were looking to answer from the model were as follows:

1. Based on historical data, for each dollar of signings, how much new data center resource is consumed?
2. How much data center resource will we have after spending all our 2009 capital?
3. What do we expect our signings to be for 2010 and 2011?
4. How much data center resource will we consume in 2010 and 2011?
5. How much of a shortage of data center resources do we predict for year end 2011?
6. Based on the average number of square feet per server, how many servers do we need to virtualize to mitigate the shortage?

We chose server virtualization as the single technology in the model because it was listed in the EPA report as the single largest IT technology that would impact data center resource requirements. This was based on the fact that most distributed servers were sized based on peak workload which often resulted in servers running at less than 20% on average. Server virtualization addresses this problem by isolating many server images on a single physical server, and sharing surplus resources between the images thereby reducing surplus resources. Additional reduction in resources can be obtained by optimizing the pool of virtual servers and moving images between the physical servers based on load.

While server virtualization was the technology selected for the model, there are other technologies—both on the data center facilities side and the IT side—to improve the data center resource supply. On the facilities side, this may be as simple as improving cooling air flow by filling cable cutouts, installing blanking plates on IT racks of equipment, ensuring perforated tiles are properly located, or installing IT equipment in a hot aisle, cold aisle configuration. On the IT side, technology improvements could include the sunset of old applications, enabling energy management features of IT equipment, compressing data, tiered data, or improving the efficiency of applications to require less IT resources per transaction. For example, there is a white paper that describes the implementation of these technologies at a single data center [2]. This paper describes the benefit of each technology was added to the model to reduce the amount of server virtualization needed to contain the growth.

3 Model Input – Facts

There were several pieces of information that we were able to include in the model that were based on several years of measurement, or where we had a sufficiently large number of data points to have a high level of confidence on what value should be used in the model.
These values are as follows:

1. PREVIOUS BUSINESS GROWTH RATE – This value represents the amount of new IT business (in dollars) we received over the last several years.
2. AMOUNT OF DATA CENTER SPACE CONSUMED – This value represents the amount of facilities resources that was used by the business growth indicated in the previous value. As the mix of server, storage, and network devices changes year to year coupled with improvements in IT capacity per unit of facilities resource, we considered computing resources per revenue a variable value that would change over time.
3. UTILIZATION RATE – This value represents the monthly average physical utilization of non-virtualized servers and servers which have virtual workloads running on them. We found that the servers with virtual workloads were running at twice the average utilization as their non-virtualized counterparts, but still had room for improvement as most capacity planners in the virtual environment were being conservative and not sizing the new servers for high amounts of utilization.
4. VIRTUALIZATION CONVERSION RUN RATE – This value represents the amount of virtualization we were able to perform for the last several years. We averaged about 20,000 images per year. This was used as a sanity check that the number of images we needed to convert against the current resources

available to convert the workload from physical to virtual images.

5. NUMBER OF IMAGES PER VIRTUALIZED SERVER – This value represents the average number of images that can reside on a virtualized server. This number was computing by taking the number of existing virtual images and dividing that by the number of physical servers they resided on. While servers were of different size (both physically and number of processors) and platform (both UNIX and Intel) we have about 100,000 images which tended to normalize the data. We used a value of six for the model.
6. PERCENTAGE OF IDLE SERVERS – This value represents the amount of servers that are considered idle as indicated by a monthly server utilization of 3% or less. Some of these are idle for a specific reason (e.g. active/passive implementation of high availability cluster) while others are at end of life and need to be discontinued. In 2009, we reported on about a third of our images, roughly 100,000 images reporting daily. About 30% of these images reported average monthly utilization in the idle category. Having servers with low utilization is not uncommon in the industry, as servers were sized for peak utilization verses average utilization. "One of the most troubling aspects about data centers is that in a lot of these cases, we're finding that server utilization is actually around seven percent," Federal Chief Information Officer Vivek Kundra said [3]. With numbers such as these one can imagine the opportunity of even just eliminating half the server capacity within an account by doubling the server utilization. In the US government case, this would mean going from 7% utilization to 14% utilization which is still extremely low and provides room for growth.

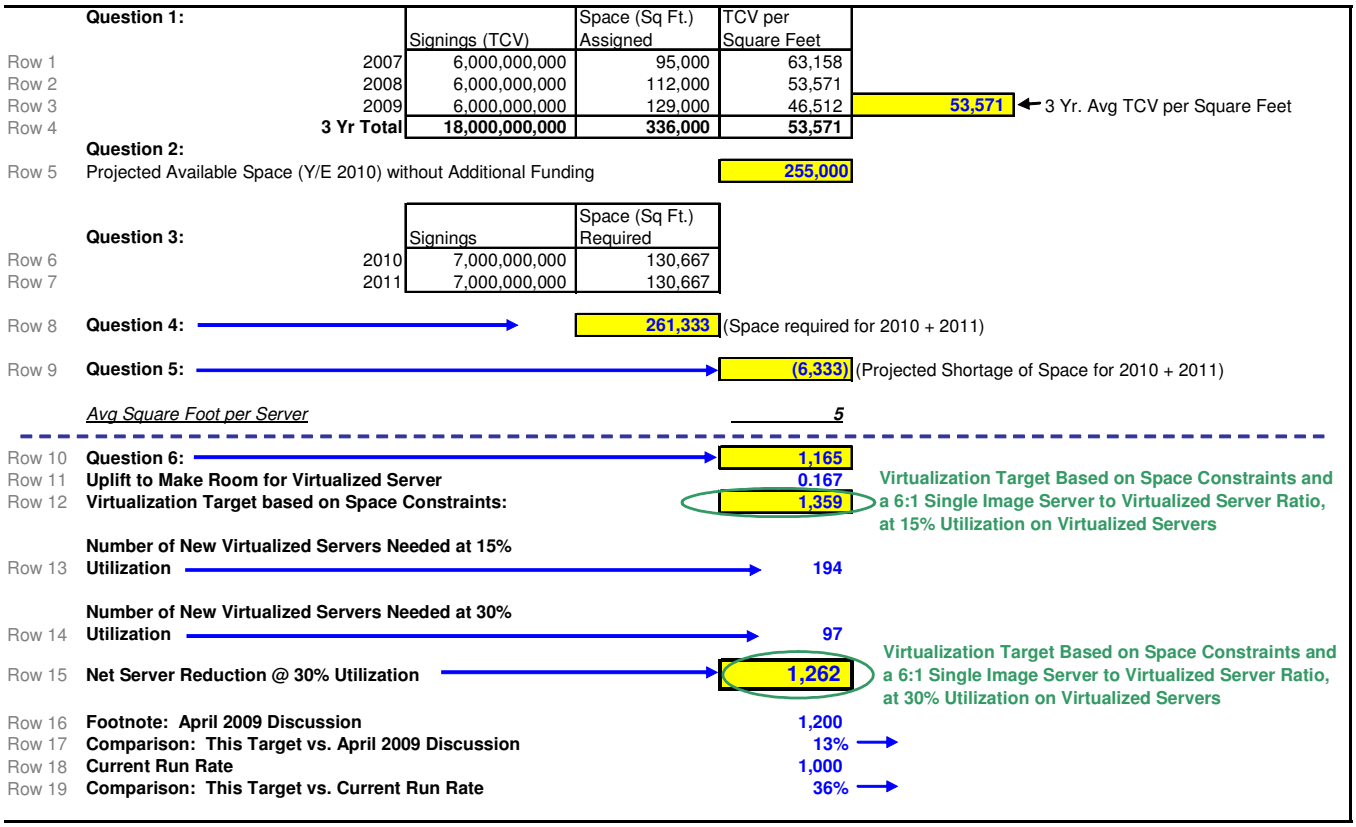
4 Model Input – Assumptions

In addition to the historical values we were able to determine (or compute), there were several values that we did not have data for. In those cases we made the following assumptions:

1. FUTURE BUSINESS GROWTH RATE – This value represents that we expect to grow new IT business (in dollars). We assumed that we were going to grow at the growth rate of the company (which was 6%) as we are a large portion of the companies business and this growth rate was published to our investors.
2. FACILITEIS GROWTH RATE – This value represents the amount of facilities resources required resulting from converting the business revenue to space requirements. In order to calculate this value we multiplied the dollar growth projected by the historical space per revenue metric.
3. NUMBER OF FULLY DEPRECIATED SERVERS – This value represents how many servers we could virtualize without having to be concerned with the financial terminal value of the server (book value). We calculated this value based on the number of physical machines that didn't run virtual images yet, and divided that by five (the number of years in the depreciation schedule).
4. NET BOOK VALUE OF SERVERS VIRTUALIZED – This value represents the number of existing physical servers that had no net book value. In looking at the business case for technology adoption, we did not want to write off assets financially. The approach we used here was to assume that the average purchase price of a server was $10,000, it was depreciated over five years, and that the target servers were over four years old (additional servers that would fully depreciate by the time we ran this year long project). We assumed that we may have to virtualize servers with one year left of depreciation, so we set the net book value of those servers to $2000.
5. AMOUNT OF SPACT THAT A SERVER CONSUMES – This value represents the conversion ratio from physical servers to facilities resources needed. We simplified the model by using two dimensions (length and width) verses three (height). We did this by estimating how many servers are in a rack and how much space (including service clearance) a rack of IT equipment took up in the data center. After sampling several configurations, we decided to use the value of 5.4 square feet per server.

5 Model Results – Model Iterations

The input data was put into the spreadsheet, with the following results:



	Question 1:		Signings (TCV)	Space (Sq Ft.) Assigned	TCV per Square Feet	
Row 1		2007	6,000,000,000	95,000	63,158	
Row 2		2008	6,000,000,000	112,000	53,571	
Row 3		2009	6,000,000,000	129,000	46,512	53,571 ← 3 Yr. Avg TCV per Square Feet
Row 4		**3 Yr Total**	**18,000,000,000**	**336,000**	**53,571**	
	Question 2:					
Row 5	Projected Available Space (Y/E 2010) without Additional Funding				255,000	

	Question 3:		Signings	Space (Sq Ft.) Required
Row 6		2010	7,000,000,000	130,667
Row 7		2011	7,000,000,000	130,667

Row 8	**Question 4:** →	261,333	(Space required for 2010 + 2011)
Row 9	**Question 5:** →	(6,333)	(Projected Shortage of Space for 2010 + 2011)
	Avg Square Foot per Server	*5*	
Row 10	**Question 6:** →	1,165	
Row 11	**Uplift to Make Room for Virtualized Server**	0.167	
Row 12	**Virtualization Target based on Space Constraints:**	1,359	**Virtualization Target Based on Space Constraints and a 6:1 Single Image Server to Virtualized Server Ratio, at 15% Utilization on Virtualized Servers**
Row 13	**Number of New Virtualized Servers Needed at 15% Utilization** →	194	
Row 14	**Number of New Virtualized Servers Needed at 30% Utilization** →	97	
Row 15	**Net Server Reduction @ 30% Utilization** →	1,262	**Virtualization Target Based on Space Constraints and a 6:1 Single Image Server to Virtualized Server Ratio, at 30% Utilization on Virtualized Servers**
Row 16	**Footnote: April 2009 Discussion**	1,200	
Row 17	**Comparison: This Target vs. April 2009 Discussion**	13% →	
Row 18	**Current Run Rate**	1,000	
Row 19	**Comparison: This Target vs. Current Run Rate**	36% →	

The answers to the questions in section 2 are indicated in the appropriate row in the table. The revenue numbers in the example have been modified for purposes of this paper. We are able to state how much space we would have available in 2010 without any additional funding (no new data center builds) if there were no growth (space on hand at year end). We then computed the space required by our anticipated signings (question 3) and added 2010 and 2011 requirements to get space required by year end 2011 (question 4). Subtracting new space requirements from space available left us with a projected space shortage of 6,333 square feet. There are various ways that we could make that many square feet available such as build or rent new data center space. For purposes of this paper the single technology improvement we are considering is server virtualization.

Assuming that a server consumes 5.4 square feet of space, we would need to virtualize 1,165 servers to make 6,333 square feet available. As the virtual server itself takes space, this was normalized to 1,359. This figure assumed that servers would continue to be virtualized at 15% target average utilization. At 6:1 images per server this results in 194 servers being needed. If utilization was targeted at 30%, 97 servers would be needed, bringing the net servers reduced to 1,262. This amount was in line with the 1,200 initial estimates that were being discussed as it was about 20% higher than the 1,000 server run rate that we were currently converting. The thought was that we could improve our labor productivity by 20%, and be able to perform the 1,262 conversions with the existing resources being funded for conversions.

We also compared the conversion target to the terminal financial value of the servers (NBV) and the amount of idle servers. In both cases, there were sufficient servers with no NBV or are idle such that we did not need to reserve money to write off the older equipment or have low enough risk to sunset or virtualize servers with minimal to no workload.

6 Recommendations

This resulted in the following recommendations for 2010:

1. Use four socket verses 16 socket (dual core) servers with maximum amount of memory as some of the existing virtual servers ran out of memory before they ran out of processing power. The reduction in processing power while keeping the maximum amount of memory increases the amount of memory that each processor will be able to have allocated to it.
2. Target a minimum of six images per server.
3. Target the fully depreciated servers first.
4. Target the locations that are space constrained.
5. Sunset idle servers instead of virtualizing them.
6. Plan new servers with a minimum of 30% average monthly utilization.

These recommendations were given to the team as part of setting their target of 1,262 physical servers to convert to virtual images.

7 2010 Measurements

As we started 2010, we set the following targets which we measured on a monthly basis:

1. Convert 1,262 images
2. Target a 1% increase in the overall utilization of our server base, which would be a combination of our actions to virtualize existing servers, sunset idle servers, and improve utilization of existing virtualized environments through technology like Dynamic Resource Scheduling (DRS).
3. Reduce data center capital build budget by 50% as new data centers would not be needed for existing customer growth. The remaining capital would be used for new customer growth.
4. Reduce distributed hardware capital budget by 30% as improved utilization would not require as much hardware. This was actually implemented as a reduction in our overall IT capital budget by 15% as we did not measure distributed hardware directly. We analyzed 2009 spend to determine that distributed hardware was approximately half of the total hardware budget.
5. Increase the servers we received capacity management data on to 80% of our servers. This was done to increase the accuracy of our model and to aid in identifying the idle servers for sunset or the low utilization servers to virtualize (lower risk as less workload was running on them for the higher gain of floor space).

These metrics are reported on a monthly basis. As of May 2010, we were meeting these assigned targets with the exception of the pooled utilization target. While we continue to virtualize servers at double the physical utilization rate, we have not yet removed idle servers or increased existing virtual server pools in order to raise the overall utilization average. The base of existing equipment is large enough so that virtualizing existing images is insufficient to raise the utilization to the target by itself.

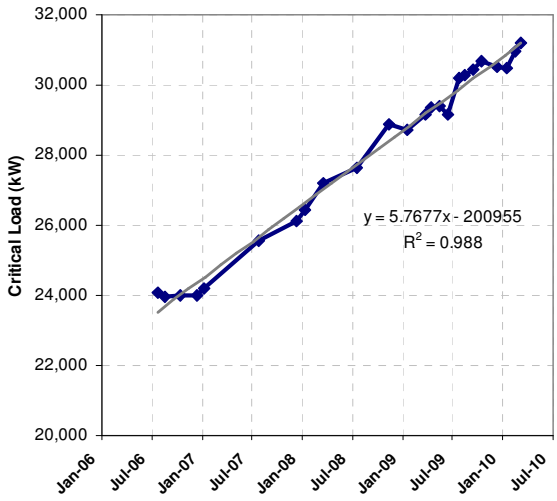
8 Model Improvements – Simulation

As we proceeded into 2010, we decided to add more sophistication to the model. We switched the approach from using a spread sheet to creating a Monte Carlo simulation of the model. This also allowed us to change the input variables from fixed values to a distribution which would provide us some sensitivity analysis.

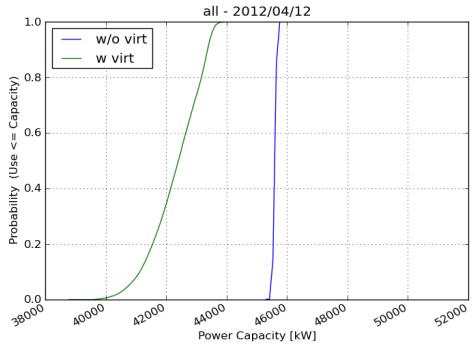
We also decided on changing the units of the constrained facilities resource. In building a data center, the data center space accounts for 5-10% of the initial cost of the build. The largest part of the initial cost of the build (over 50%) is for the power and cooling equipment. We therefore changed our units of resources to power instead of space.

As part of looking at previous revenue growth, we took a closer look at the requests we were getting for data center resources and determine how much power was being consumed over time. At the total data center level, it appeared that power usage was growing in a linear pattern. Assuming that this was the case, we performed a least squares linear regression to determine our power load growth.

The following chart represents the regression analysis:



This analysis indicated that power usage was increasing at the rate of two megawatts a year.
We ran the simulation with and without the affects of virtualization to understand for the next four years how the power usage would increase. Here is the chart from year two:



The scope for this analysis is about a dozen data centers which contain over a million square feet of data center space. Our current capacity for the data centers in scope is about 46,000 kilowatts. The majority of our portfolio operates at about 40 watts per square foot. Recent data centers are being built with a modular approach to continue to increase the power density without disruption of service. The IBM data center in Research Triangle Park was designed for 150 watts per square foot. The above chart shows that we will be at full capacity by end of year two if we do not virtualize at the projected levels. If we do virtualize we will consume about 90% of the power resources, with a variance of about 5% in both directions. This means that if we continue to virtualize at the run rate approximately 20% improved over our existing rate that we will not run out of data center power until year three. We had already demonstrated we would fit within our space footprint. The last thing a data center requires is cooling. Fortunately the data centers were sized for the cooling to match the power. As there will be enough power, there will be enough cooling.

9 Future Improvements to the Model

We are currently working on adding the following improvements to the model:

1. Effects of injecting other technologies (e.g Like dynamic monitoring of air flow and temperature as input to load balancers) other than server virtualization
2. Performing the analysis at a data center level to determine what growth is needed at each data center verses at the aggregated level. This would allow us to determine separate growth rates at each data center and create a custom mitigation plan for each one.

3. Incorporating the virtualization adoption curve for each data center so we know how much adoption we have left to do.
4. Validate the additional utilization data we are collecting which indicates servers have the same utilization independent of having the tools on them.
5. Modeling power, space, and cooling within the same model to determine where the constraints occur first with which resource.

10 Summary

This paper started with a discussion about how IT demand was going to continue to increase at levels that will consume the current data center resources. As building a data center is very expensive, companies are looking to see if there are other technologies on both the facilities and IT side of the organization to prevent from having to build a new data center. This paper analyzed one such technology, server virtualization to determine what amount of virtualization would be required to contain two years worth of anticipated data center growth. The amount of conversion was at a run rate 20% higher than we had been converting images in the past. We expect to make productivity improvements to our conversion process to eliminate the additional labor resources required to covert enough images and prevent having to build new data centers as the only solution to IT growth. We also validated that we would not have financially stranded assets and that there were sufficient low risk (idle or low utilization) servers to take action on.

As a result of the level of sever virtualization, we are able to reduce data center build capital, hardware capital, better utilize our data centers, and reduce software licensing while still having sufficient data center resources until year end 2011. This means that we don't have to have new data center space built before January 2012, giving us time to consider what data center build options we want to pursue, and still provide us the time to build the new data center (which typically takes 12-18 months from ground breaking).

While we started our model as a simple spread sheet, we increased in complexity to a simulation model. This confirmed that the rate of conversion was sufficient to prevent from having to build additional data centers. Since virtualization isn't the only technology we were injecting to improve data center efficiency or reduce IT demand, the other improvements would result in having additional room for growth past the two year horizon. For example, we could understand the effects of changing the power density from 40 watts per square foot to 150 watts per square foot, or changing the efficiency of the platform. This allows us to consider thinking of capacity planning based on energy as the primary metric, and not based on server utilization alone.

11 References

[1] EPA. 2007. Report to Congress on Server and Data Center Energy Efficiency. August.

[2] Computer Measurements Group (CMG) 2008. Green Data Center: A Case Study by Chris Molloy, session 351.

[3] Data Center Knowledge 2010. Kundra: Fed Data Centers 7 Percent Utilized by Rich Miller. The article is available at Http://www.datacenterknowledge.com/archives/2010/04/09/Kundra-fed-data-centers-7-percent-utilized/ April 9, 2010.