Capacity Planning:
A Revolutionary Approach for Tomorrow's Digital Infrastructure

Amy Spellmann
The 451 Group
amy.spellmann@the451group.com

Richard Gimarc
CA Technologies
richard.gimarc@ca.com

Capacity planning has been a well-established practice for over 30 years. During that time, the tools, techniques and processes have been defined and refined. However, our traditional approach cannot keep pace with today's rapidly changing environments; we need to revolutionize the practice of capacity planning.

This paper will examine the current scope and focus of capacity planning and propose an innovative methodology to evaluate, predict and plan for the all-inclusive Digital Infrastructure. It is no longer sufficient to utilize yesterday's outmoded approach when planning for tomorrow's applications, systems and facilities infrastructures. We need to revolutionize the practice of capacity planning. This paper identifies the goals and challenges of Digital Infrastructure capacity planning and defines a new approach that adapts to tomorrow's extraordinarily dynamic, diverse and expanding environments.

1 The World Has Changed

Capacity planning has evolved over the past 30 years, adapting to the changing landscape of IT systems and application architectures. However, the evolutionary process can no longer keep up with the rapidly increasing complexity, size and scope of today's IT enterprises. The exponential transformation in technology, applications and infrastructures is demanding today's capacity planning practice to include the entire Digital Infrastructure. Our use of the term Digital Infrastructure is important and intentional since it describes the breadth of today's capacity planning world view: The 451 Group defines Digital Infrastructure as encompassing the enterprise-wide computing environment and the data center; including business requirements, technology infrastructure, and facilities infrastructure [SPEL2012]. The practice of capacity planning must be revolutionized to embrace a methodology that keeps pace with the rate of technological evolution, necessitating the inclusion of all components in the Digital Infrastructure.

Traditional brick and mortar, mainframe and distributed systems no longer support the needs of today's dynamic business environment. In today's marketplace, businesses must be visionary and aggressive in providing IT services. Keeping pace with clever competitors and innovative entrepreneurs has driven the need for accelerated delivery of IT services to support business requirements more cost effectively than ever before.

To make matters more complicated,

- IT services can now be constructed from a plethora of technologies and architectures which must adjust to business requirements on demand; to the extreme of automated and dynamic resource allocation.
- The sheer magnitude of compute devices has skyrocketed; from millions to billions.
- Application architectures comprised of the latest technology stacks, database and storage options are widespread.
- Ownership across the Digital Infrastructure now varies to the extreme; interconnected components can be delivered internally or in the cloud (e.g., SaaS, PaaS, IaaS, or XaaS); from modular data centers to combinations of hosting options in various geographic locations.

The factors above dramatically affect capacity, performance, and cost for IT Services. The revolutionary approach to capacity planning introduced in this paper must address any possible combination of technologies and delivery options.

1.1 Magnitude of Change

In Mani Chandy's 1985 A. A. Michelson Award acceptance speech [CHAN1985] he predicted that "*the primary difference between problems of the year 2000 and those of today will be one of size".* Chandy's prediction was looking forward from 1985 to 2000, a mere 15 years. When you look at today's world you can see the significance and effect of his prediction. Today's capacity planners are responsible for environments that are close to 4 orders of magnitude larger than 1985.

To elaborate further, let's review briefly, "What has changed?"

1. Complexity has increased along with the numbers
    - Increasing number of components and the way they interact
    - Heterogeneous components; we are not planning for millions of identical "things" but unique ones
    - Mobile application architectures are moving more of the processing load from the end user device back into the data center (reversal of the historical trend where desktops/laptops kept most data and processing local)
2. Ownership of components varies drastically, more and easier choices are available
    - Many combinations of physical locations versus on premise
    - Compute can happen anywhere: cloud, hybrid, public, private
    - Commodity computing (hardware, software, applications)
    - Converged infrastructure
3. Global presence and inter-connectedness are requirements for most businesses today
    - Multiple data centers, world-wide
    - "The sun never sets"
4. Facilities design choices and costs are now aligned to the all-inclusive Digital Infrastructure
    - Data center capacity: space, power, cooling options
    - Compute per kWh has doubled almost every year from the 1940's through 2010
    - Adding facilities capacity costs $M; the tendency is to overbuild

Taking all of these trends into consideration, we see the problem of capacity planning changing by at least an order of magnitude. As Dijkstra was quoted by Chandy [CHAN1985], "*When the size of a problem changes by an order of magnitude, the problem itself changes.*" We are looking at a new problem! Thus we require a revolutionary solution.

1.2 A Brief History of Capacity Planning

Capacity planning started in the 1970s. At that time, capacity planners were responsible for a handful of "servers" (a.k.a. mainframes). A large amount of data was collected and analyzed (mostly SMF and RMF). Rudimentary modeling and forecasting was done to predict future infrastructure requirements.

Between 1970 and today a number of things have changed:

- Capacity planners are now responsible for a wide variety of platforms.
- IT architectures have changed from a centralized platform to physically distributed servers.
- There is an ever expanding set of measurement data sources.
- Many tools are available for data collection, analysis and predictive modeling.

Figure 1 shows an overview of the capacity planner's changing landscape.

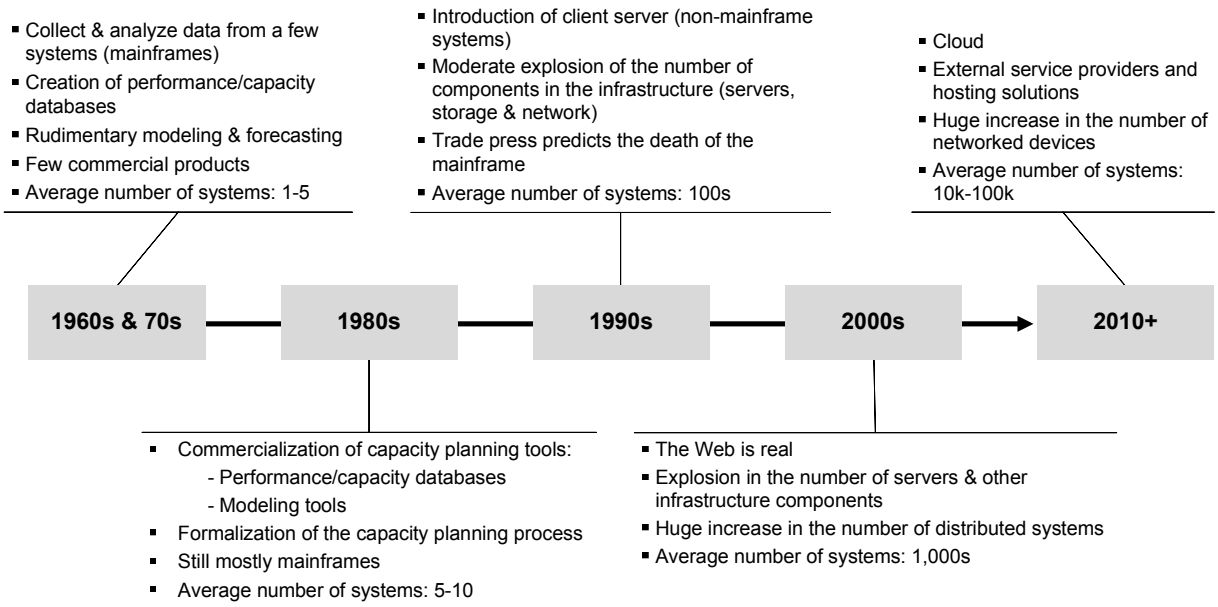

Figure 1. Capacity Planning Historical Timeline

What can we learn from history? Capacity planners are now responsible for ever increasing, diverse, dispersed and interconnected environments. The single attribute that has most affected today's Digital Infrastructure is size; we are now dealing with hundreds of applications and thousands of servers. Capacity planners must find a way to deliver their services in a more comprehensive manner that spans the breadth of the Digital Infrastructure, from high-level business drivers to the data centers that host the IT equipment.

1.3 Why Traditional Methods No Longer Suffice

With the large amount of change in magnitude and complexity comes the realization that traditional capacity planning practices are insufficient and not sustainable. In general, all our methods have been "siloed" across the Digital Infrastructure: servers, storage, network and facilities. Each silo focuses on their narrow niche and ignores the holistic view of the enterprise infrastructure. In many companies, silos even have secondary silos (e.g., Windows, Unix, zOS).

The table below shows the common capacity planning areas (silos): Server, Storage, Network, Power/Space/Cooling (facilities) and Cost. The table lists the typical metrics and methods used for capacity planning and their limitations.

	Metrics & Methods	Limitations
Server (CPU & Memory)	- Platform level CPU utilization based trending, forecasting and modeling - Modeling used to predict and evaluate future infrastructure requirements	- Server-centric view of the enterprise; single OS instance per server prevailed in the past, but virtualization changes the problem - Memory usage has become a popular sizing method over CPU - Difficult to apply to cloud instances
Storage	- I/Os per second, space used (GB) and storage bandwidth. (MB/sec) - Trending and forecasting of storage space utilized/free	- Generally viewed in isolation - New technologies require more granular tracking of application resources
Network	- Isolated network traffic, latency, bandwidth utilization - Traffic modeling used to predict latency, network utilization and response time	- Partial view of the world. - Generally ignores compute bottlenecks - Service provider unknowns and geographic limitations - Increased complexity, difficult to track individual applications
Power, Space, Cooling	- Space (sq. ft.), power (kWh) and cooling (BTUs) spreadsheet analysis of electrical trends on an annual basis	- Performed independently of IT capacity planning - Relies primarily on current snapshot, not predictive
Cost	- IT costs (mostly CAPEX) - Facilities OPEX and CAPEX - Trending based on historical growth, practices	- Not comprehensive; siloed - Forces decision making based on total IT and facilities costs which are not correlated

Current methods do not utilize information across silos. Cross-silo communication is required to develop a comprehensive capacity plan for the Digital Infrastructure. A new paradigm is required.

2 The New Paradigm

We previously defined capacity planning as follows [SPEL2008]:

> Capacity Planning is the process of predicting when future business demand will exceed the availability of IT equipment, energy and space in the data center and then determining the most cost-effective way to meet SLAs and delay saturation.

Based on today's changing landscape we propose the following refinement:

> Capacity Planning is the process of predicting the *impact* of business demand on the availability and scalability of IT equipment, space, power and cooling in the data center and then determining the most cost-effective way to *optimize* service delivery and meet SLAs.

Why did we change our definition? The quick answer is that in today's environment we need to refocus on aligning our Digital Infrastructure to evolving business demands rather than just avoiding saturation. Both definitions share common characteristics:

- Create an environment that will meet and satisfy business demand
- Avoid saturation

We refined our 2008 definition to address *optimizing* the Digital Infrastructure to satisfy business demand. The previous definition was more concerned with avoiding saturation. However, today's capacity planner must look for ways to optimize the Digital Infrastructure based on increases or decreases in business demand.

2.1 Guiding Principles

The reason traditional methods no longer meet the needs of the business is that the g*oals and scope* for capacity planning have changed. We are now driven by a new set of guiding principles:

- Cost effective IT service delivery
- Reduced infrastructure footprint
- More scalable management requirements
- Optimized infrastructure for business needs
- Reduced power, space, cooling requirements for facilities
- Holistic planning across the entire Digital Infrastructure
- Long term tracking of success/efficiency factors across the Digital Infrastructure

These guiding principles are aligned with our refined definition of capacity planning. The scope of today's capacity planners has changed; they are now responsible for the entire Digital Infrastructure rather than just the IT equipment. Business demand and requirements are still the driving force. However, today's capacity planners must be able to translate those requirements into a form that is more amenable and adaptable to diverse technology and hosting solutions. Furthermore, capacity planners must have a way to track and demonstrate their long-term success.

2.2 What's New and Revolutionary?

In this paper we are introducing a new structured method for capacity planning. Our methodology is based on the *Capacity Planning Stack* that incorporates all components in the Digital Infrastructure, organizing them into a cohesive and comprehensive planning paradigm. This new methodology has the following characteristics:

- End-to-end view encompasses all components of the Digital Infrastructure
    - Organized in a multi-level hierarchy
    - Each level corresponds to a portion of the Digital Infrastructure
    - The hierarchy supports capacity planning workflow from the business to the data center (facilities)
- Organized workflow between the levels of the Capacity Planning Stack
    - Well defined dependencies and workflow between stack levels (demand and feedback)
    - Efficiency metrics at each level used for long term tracking (measures of success)
    - Useful work products generated at each level of the stack
    - Conceptually transparent and straightforward
- Integrated and inclusive capacity plan
    - Business view of the costs of the supporting Digital Infrastructure
    - Cost allocation across all components of the Digital Infrastructure

3 The Capacity Planning Stack

We propose to view capacity planning in terms of a *stack*. Within IT we are already familiar with a number of stacks. For example, consider the following examples:

- **Technology Stack** - A set of software that provides the infrastructure for a computer. The stacks differ whether installed in a client or a server. [DICT2013]
- **Solution Stack** - An ordered collection of software that makes it possible to complete a particular task. [TECH2013]
- **OSI Model Stack** - The Open Systems Interconnection (OSI) divides the complex task of computer-to-computer communications into a series of stages known as layers. Layers in the OSI are ordered from lowest level to highest. Together, these layers comprise the OSI stack. The stack contains seven layers: application, presentation, session, transport, network, data link and physical. [WIKI2013]

The Capacity Planning Stack consists of an ordered set of hierarchical tasks that must be performed to develop a complete, viable and defensible capacity plan for the Digital Infrastructure. Our initial view of the Stack is illustrated in Figure 2.

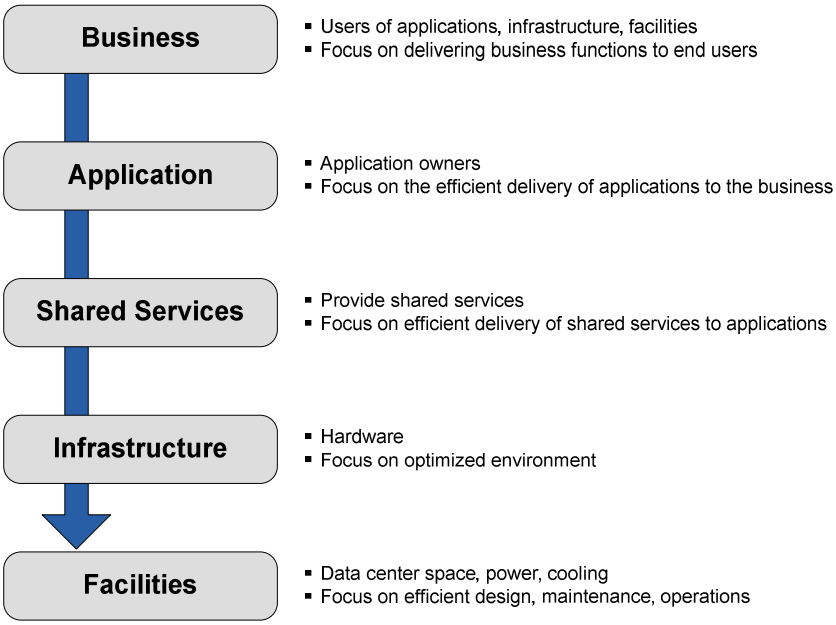

Figure 2. The Capacity Planning Stack

Business Capacity planning always starts with the businesses it supports. It is assumed that each business is supported by one or more applications. The business is responsible for providing usage estimates to the supporting application owners. Usage estimates are usually expressed in terms of the business transaction volumes.

Application Applications are the business' interface to the Digital Infrastructure. Applications are designed and developed to provide the IT services that support the business. Application-level capacity planners are focused on the efficient delivery of applications to the business. These planners are tasked with translating the higher level business requirements into IT resource requirements. This translation is done per application. Examples include back-office and customer facing applications which can be home-grown, legacy, COTS (commercial off the shelf), and SaaS. Outputs from the application capacity planners serve as input to two lower levels: shared service requirements and infrastructure requirements.

Shared Service Shared services include the hardware and software components that support one or more application. Examples include message queuing systems, message brokers, databases, Web server farms, or cloud services. The shared service planner's responsibility is to size their environment in order to support their upstream application users. Application capacity planners provide the shared services planners with their expected demand on the shared components of the infrastructure. The shared services planners forward their infrastructure requirements to the next level in the Stack (infrastructure).

Infrastructure The infrastructure level contains the typical physical and virtual components included in traditional capacity planning; servers, memory, storage and network. Infrastructure can be geographically dispersed, in the cloud or hosted on premise. The infrastructure capacity planners utilize input from the application and shared services layers to

determine the most cost effective way to optimize service delivery and meet SLAs. Infrastructure space and power requirements for on premise physical components are sent to the facility level.

Facilities Facilities planners are responsible for ensuring that their data center can support the required IT infrastructure from a space, power and cooling perspective. Their focus is on the efficient design, maintenance and operation of the data center. The primary task of the facilities level is to plan for future IT infrastructure support, ensuring that the data center can provide adequate resources as IT evolves. The challenge for facilities planners is adapting their timeline, which spans years, to the contrasting IT monthly (or less) horizon.

3.1 Workflow – Capacity Planning Stack

The previous section introduced the Capacity Planning Stack. This section will provide more details about the workflow between the Stack levels and the work products produced as part of the capacity planning process.

A refined diagram of the Capacity Planning Stack is shown in Figure 3. The following additions have been made to the Stack diagram.

- **Demand flows down the Stack.** Business owners provide the application planners with their expected business volumes. The application planners translate business demand into resource requirements and pass them downstream to the infrastructure level. Additionally, if appropriate, the application plan determines the shared service requirements, in terms of transaction volumes, and provides this requirement to the shared service level (which in turn determines the associated resource requirements and passes them down to the infrastructure level. This level-to-level communication process continues throughout the Stack. The last demand flow is from the infrastructure to the facilities level; this step is required to ensure that the data center can support the entire breadth and depth of the Digital Infrastructure.
- **Feedback flows up the Stack.** A feedback loop communicates requirement results back up the Stack to ensure alignment to higher level plans, assist in future planning and potentially refine upstream estimates or designs. For example, the capacity plan developed by the application planners determines an infrastructure solution and associated costs. This information is passed back to the business level for evaluation per the business plan. If, for example, the costs exceed business budget, there may ensue negotiations between the levels. Another example to consider is at the bottom of the Stack. Suppose the infrastructure planners determine that they need 100 more mid-range servers. What happens if the facilities planners estimate that there is insufficient power or cooling capacity available in the data center to support the additional servers? Again, the feedback mechanism provides a means to convey this message back up the Stack. Optimization or a change in delivery option at any higher level can potentially alleviate a large facility expense.
- **Efficiency metrics at each Stack level.** Each level in the Stack maintains their own set of efficiency metrics to track long term trends. Efficiency metrics can be viewed as a “measure of success” or “report card” for each Stack level. For example, the application planners can generate a productivity measure for their application that describes the number of transactions processed per unit of resource (similar to miles per gallon for an automobile) and/or report performance of transactions against SLA’s. Facilities planners would use PUE as one of their efficiency metrics [TGG2007].

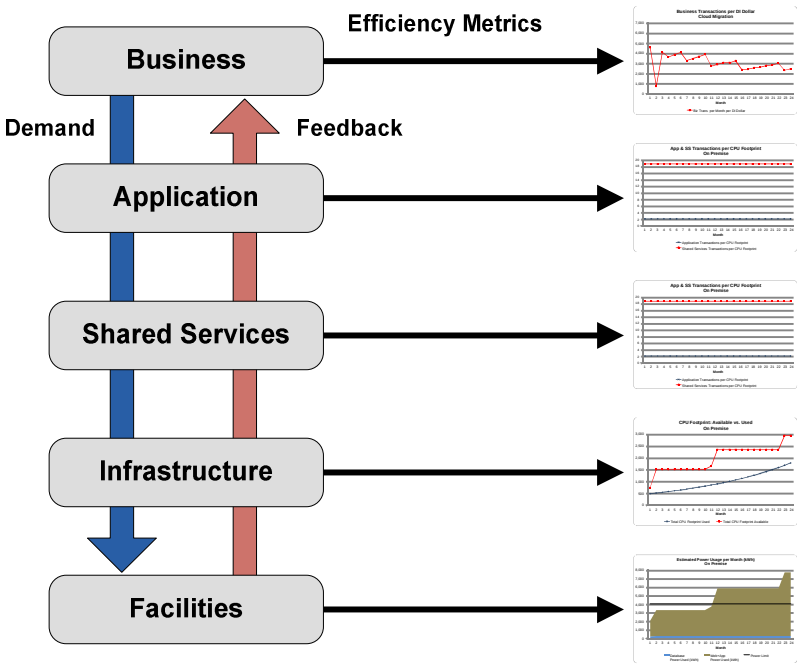

Figure 3. Refined Capacity Planning Stack

3.2 Demand, Feedback & Efficiency Metrics

In this section we take a closer look at the demand, feedback and efficiency metrics that can be used at each level of the Capacity Planning Stack. Figure 3 contains a summary of the interactions and work products produced by each level of the Stack. The following section will provide examples for each level in the Stack.

	Demand Factors (↓)	Feedback (↑)	Efficiency Metrics (→)
Business	- Business volumetrics & priorities - Performance requirements & SLAs	- Total cost - Total time to satisfy requirements - Expected performance	- Business transactions per Digital Infrastructure dollar - Total cost (cumulative from all lower levels)
Application	- Map Business volumetrics to application architecture - Estimates volume of Shared Service and/or Infrastructure requests - Estimates required Application resource footprint and instances - Determines performance requirements per transaction	- Cumulative cost from all lower levels - Application requirements (software licenses and hardware). - Expected performance. - Time to deploy, - Staffing requirements.	- Transactions/minute per resource footprint - Cost per transaction ($) - Cumulative from lower levels - Performance (e.g., response time) to demonstrate SLA achievement
Shared Services	- Map Shared Service requests to Infrastructure requests - Estimates required Shared Service resource footprint and instances - Determines performance requirements	- Cumulative cost from all lower levels - Shared Services requirements (software licenses and hardware). - Expected performance. - Time to deploy - Staffing requirements.	- Transactions/minute per resource footprint - Cost per transaction ($) - Cumulative from lower levels - Performance (e.g., response time) to demonstrate SLA achievement
Infrastructure	- Translate Application & Shared Service resource footprint and instance requirements to Infrastructure requirements - Determine physical hardware requirements - Initiate procurement process - Evaluates expected performance, headroom and SLAs	- Cumulative cost for infrastructure and facilities - Infrastructure requirements (e.g., servers, storage, network) - Time to procure & deploy	- Count of IT components (servers, storage, network) - Processing capacity per IT component category - Headroom for each IT component category - Cumulative cost of Infrastructure and Facilities

	Demand Factors (↓)	Feedback (↑)	Efficiency Metrics (→)
Facilities	- Estimate required space, power & cooling - Uptime SLA requirements	- Cost for facilities - Data center facilities requirements - Time to satisfy and/or build out	- Power, cooling, space per IT Load - PUE - Facilities headroom - Total Cost (OPEX)

Business. All capacity planning starts with the business; this does not change with the Stack. The Business develops projections for future workload (either increasing or decreasing). These projections (Demand) are passed to the Application level where they are translated into more application-centric resource and demand metrics.

The Application level provides feedback to the Business. Since the Business is at the top of the Stack, the feedback it receives is an aggregation of all lower levels. This feedback enables to Business to get a comprehensive view of what is required in the Digital Infrastructure to support their projected demand.

The Capacity Planning Stack introduces the concept of an efficiency metric. The motivation is to provide each level in the Stack with a way to track and measure their long-term success. A sample Business efficiency metric is shown in Figure 4. This chart shows the number of business transaction that can be processed per Digital Infrastructure dollar over the 24-month planning horizon. The three dips in the chart correspond to new hardware purchases (new servers were required to satisfy the projected business demand). A couple of comments about this type of efficiency metric:

- If business volume is not changing, the line should be flat.
- If business volume is increasing, you should see an increase in the line until you reach the point where a hardware upgrade is required. In that case, the Digital Infrastructure dollars spent for the new hardware will cause a dip in the efficiency line.

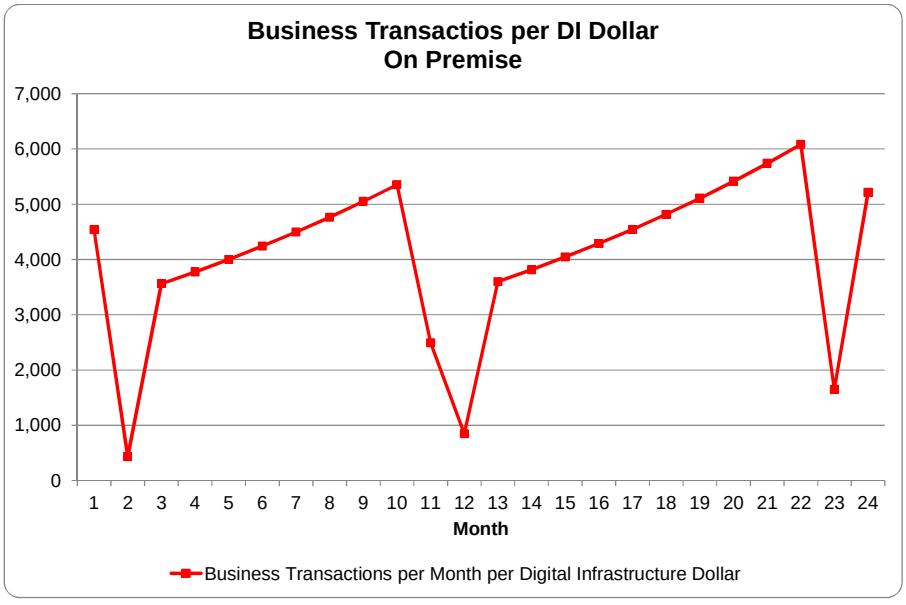

Figure 4. Efficiency Metric - Business Level

Application. The Application level is responsible for translating Business volumetrics into application-centric planning metrics. These planning metrics describe the resources required to satisfy the Business demand.

Planning metrics should be described in a way that is portable across server platforms and architectures. As an example, consider CA's Resource Score [ZINK2013]. Their Resource Score (Rx) is a platform independent vector that describes CPU, memory, storage and network resource requirements for an application. If the Application planners can describe their resource requirements in terms of something similar to a Resource Score, then the Infrastructure team can use that sizing information to determine the appropriate platform for the application (e.g., physical server, VM, cloud instance). For this paper we will use the terms *resource footprint* to describe an application's infrastructure resource requirements.

There are two demand factors that are generated by the Application planners and passed to lower levels in the Stack:

- Shared Services demand – This is similar to the demand factors passed from the Business to the Application. The difference is that these describe the Application's expected demand on Shared Services such as message queuing systems or shared databases.
- Infrastructure demand – The Application level will deliver their resource and instance requirements to the Infrastructure level. It will be their responsibility to evaluate hosting options.

The efficiency metrics produced by the Application level enable long term trending of the resource and/or cost of their application implementation. An example is shown below in Figure 5. The chart shows the number of application transactions that can be processed per CPU footprint over a 24-month planning horizon. In this example, the Application's resource demand per transaction is not changing which implies that its resource footprint per transaction remains constant over the planning horizon. If an optimization effort were undertaken to reduce the resource demand, you would expect to see the line increase (since you can do more work per CPU footprint). This efficiency chart also shows that you can process more Shared Services transactions than Application transactions per unit of CPU.

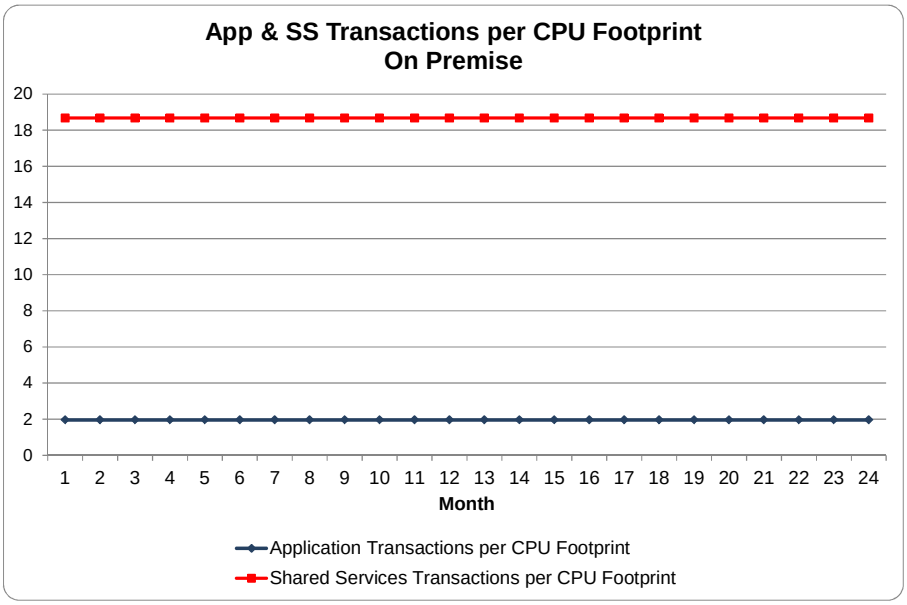

Figure 5. Efficiency Metrics - Application & Shared Services Levels

Shared Service. The Shared Service level is very similar to the Application level. The primary difference is that Shared Services receives their demand from the Application in terms of the number of requests that must be satisfied by their components (e.g., message queuing systems or shared databases).

The Shared Service planning and efficiency metrics are similar to the Application level.

The demand factors generated by Shared Services are forwarded to the Infrastructure level. After Shared Services does their job, the Infrastructure level will have a complete set of demand factors from the higher level Application and Shared Services levels.

Infrastructure. The Infrastructure level resembles many of today's capacity planning groups. They determine the IT infrastructure required to support Business demand. The primary difference with the Capacity Planning Stack is that all demand input to the Infrastructure goes through the intermediate Application and Shared Services steps. The Stack is enforcing capacity planning based on application demand.

The Infrastructure level translates the resource/instance demands from the Application and Shared Services levels into actual servers. The target servers may be physical or virtual and hosted locally on premise or in the cloud. Hosting decisions are made at the Infrastructure level based on business requirements. The Infrastructure level creates demand input to the Facilities level that describes the change in Facilities resources (e.g., power, space, cooling) required to support the changing Infrastructure.

Figure 6 shows an example of an efficiency metric for the Infrastructure level. This chart shows the total CPU footprint versus what is projected to be used over the 24-month planning period. The top line shows the available capacity in the data center; the lower line shows predicted usage. Note that this view of resource availability versus usage is consistent with the recommendations from Cockcroft [COCK2006].

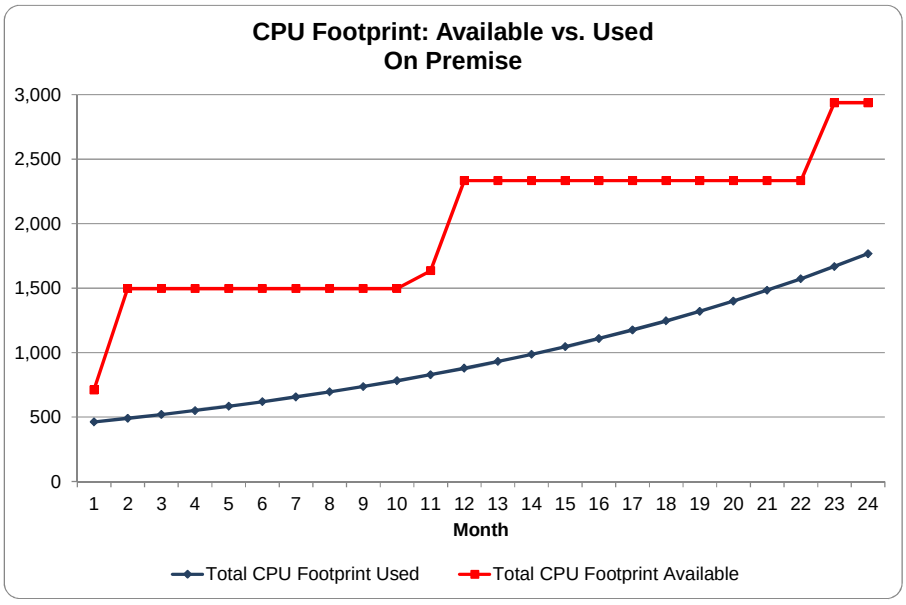

Figure 6. Efficiency Metric - Infrastructure Level

Facilities. The lowest level in the Capacity Planning Stack is Facilities. Facilities receives demand from the Infrastructure level that projects, for example, future power, space and cooling requirements for IT components. It is the responsibility of Facilities to ensure that they can handle the demand.

Figure 7 shows an example of an efficiency metric for Facilities. This chart shows the expected power usage projected over a 24-month period. The black line across the chart shows the data center's capacity. This chart clearly shows that the data center will not be able to satisfy the Infrastructure requirements. Something must change at month 12:

- Increase the power capacity for the data center
- Construct a new data center
- Migrate workload to the cloud

This example illustrates an event that would trigger feedback up through the Stack; something must be done to reduce the power demand.

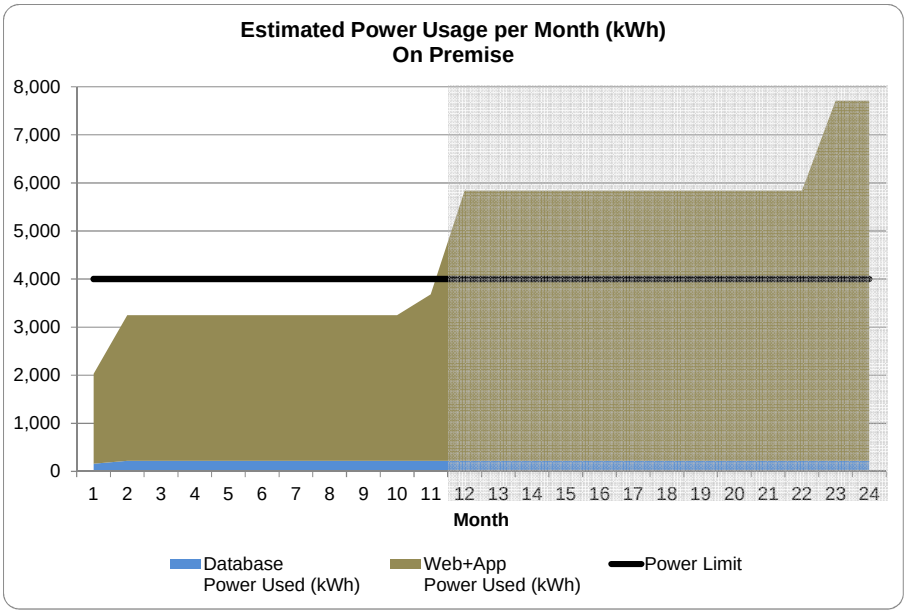

Figure 7. Efficiency Metric - Facilities

4 Case Study

This section walks through a case study that illustrates the use of the Capacity Planning Stack. Assume that we have a Business that is projecting a 6% monthly growth in their workload volume. The capacity planner's job is to determine the resource and infrastructure requirements for the next 24 months.

Figure 8 provides a preview of the workflow. The capacity planning exercise will start with the Business requirements (6% growth per month). Demand will be translated at each layer and flow downward through the Stack.

At the Facilities level a problem is noted; the current data center does not have sufficient power to support the required Infrastructure.

Feedback will be directed back to the Infrastructure level. An alternative hosting solution will be developed and evaluated (migrate into the cloud) that satisfies the data center's power constraint.

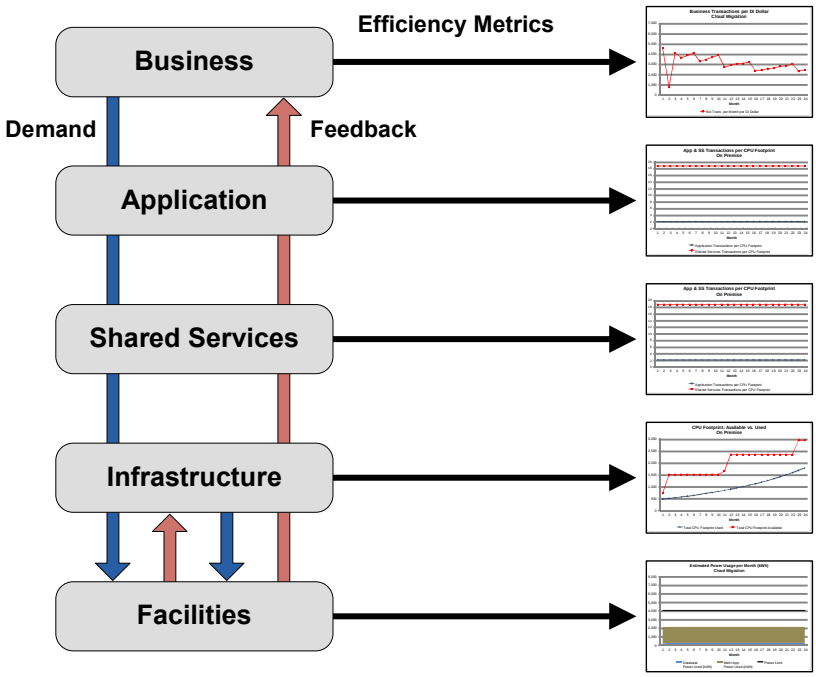

Figure 8. Case Study Workflow

4.1 On Premise: Application & Shared Services

The Application planners receive the 6% monthly demand from the Business. They translate the Business' workload volumetrics into Application resource requirements. The following table shows the current state of their 3-tier application.

	Application		Shared Services
Tier	Web Tier	App Tier	DB Tier
OS Instance Count	10	3	1
Total CPU Footprint	325	57	84

The table also shows the requirements for the Shared Services level (since this application is using a shared database instance). The next step is to project this initial state forward 24 months with a 6% monthly growth. The following two charts (Figures 9 and 10) show the instance count and CPU footprint requirements for both the Application and Shared Services levels. This information is now passed to the Infrastructure planners where they will determine the appropriate hosting solution.

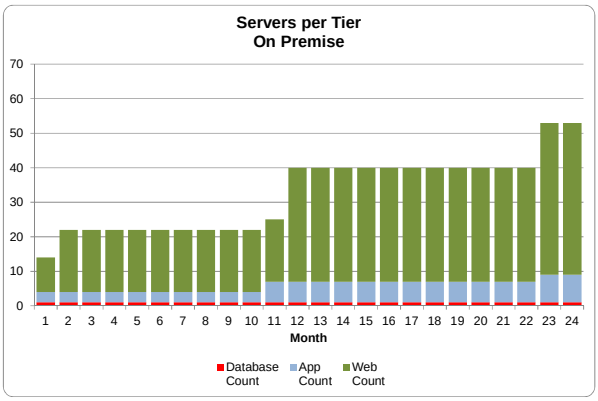

Figure 9. Instance Requirements

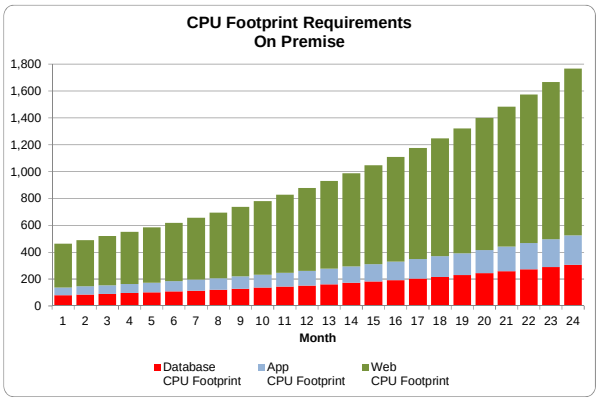

Figure 10. CPU Footprint Requirements

4.2 On Premise: Infrastructure

The Infrastructure planners receive the demand from the Application and Shared Services planners. They produce a similar 24-month projection except that their projection is based on actual infrastructure components (e.g., servers). The application is currently hosted in their data center on physical servers; they decide to simply scale out the physical environment (remain on premise) to satisfy the demand.

The Infrastructure group uses predictive modeling to evaluate their requirements for the next 24 months. Figure 11 shows the results from their modeling scenarios. Each curve represents the utilization of a tier. The required infrastructure was built out to satisfy the 24-month planning horizon while maintaining a 70% utilization threshold. The drops in the utilization curves represent where a tier was upgraded; the database server was upgraded at month 2 and the Web and App tiers were each scaled out twice. Figure 12 shows the resulting capacity curves (i.e., the CPU footprint); the top curve shows the available capacity and the lower curve shows the predicted usage.

These results satisfied the demand received from the Application and Shared Services planners. The next step is to forward this Infrastructure demand to Facilities.

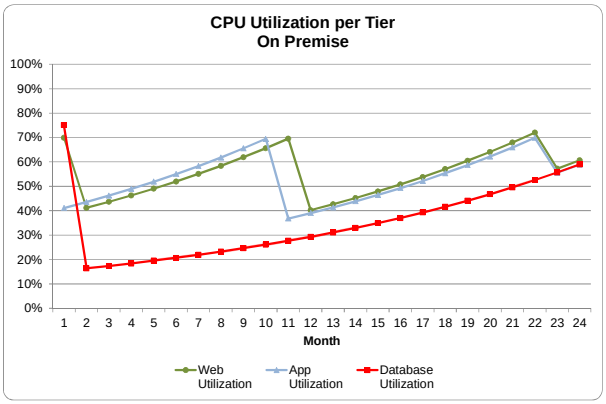

Figure 11. Server Utilization Planning Results

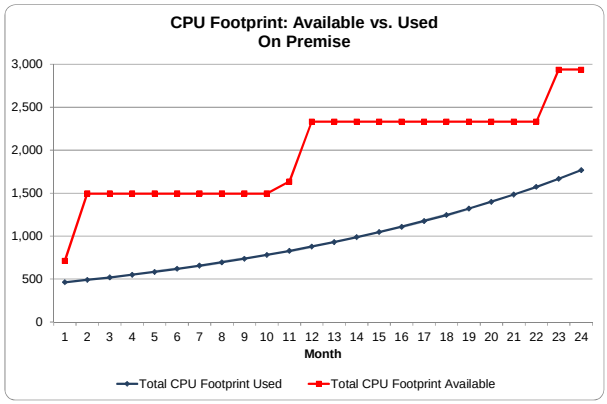

Figure 12. Server Headroom

4.3 On Premise: Facilities

Facilities planners are tasked with determining if their current power, cooling and space capacity will handle the requested infrastructure. Figure 13 shows the results of their analysis. It is clear from this chart that the proposed Infrastructure cannot be satisfied in the data center. Power capacity will be exceeded at month 12.

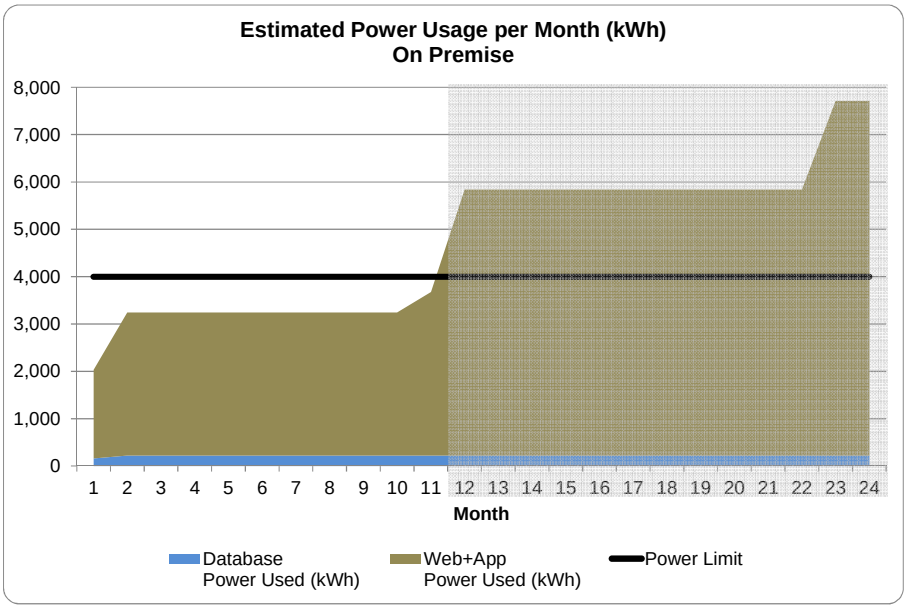

Figure 13. Facilities - Power Capacity Exceeded

At this point Facilities has two options:

- Commence a data center upgrade project to increase their power capacity.
- Provide feedback to the Infrastructure level to determine if they can develop an alternative hosting solution that avoids exceeding the power constraints in the data center.

The next section will examine the Infrastructure feedback option.

4.4 Cloud: Infrastructure

The Infrastructure planners are now tasked with developing a hosting solution that fits within Facilities' capacity constraints. They have three options:

- Return to the Business (feedback up the Stack) to see if their workload growth estimates were overly optimistic; maybe they can be adjusted downward.
- Provide feedback to the Application planners and discuss optimization alternatives. Perhaps they can tune their application to use fewer resources (which in turn would reduce the Infrastructure requirements).
- Explore alternate hosting options. For example, migrate some of the increasing workload volume into the cloud.

The Infrastructure planners decided to evaluate the use of a hybrid solution for this application:

- Maintain the infrastructure currently in the data center.
- Utilize the cloud for additional Web and App tier servers.

Migrating into the cloud would bound the growth in the data center and avoid the power capacity problem. The results from this new hosting approach are show in Figures 14 and 15. Figure 14 shows the server count for the three tiers for the on premise servers and the new cloud instances required to satisfy workload growth. You can clearly see that the count of servers in the data center remains constant; we are growing into the cloud. Figure 15 show a view of the CPU footprint for the entire application. The height of this chart is similar to Figure 10; the only difference is that now some of the servers are in the cloud.

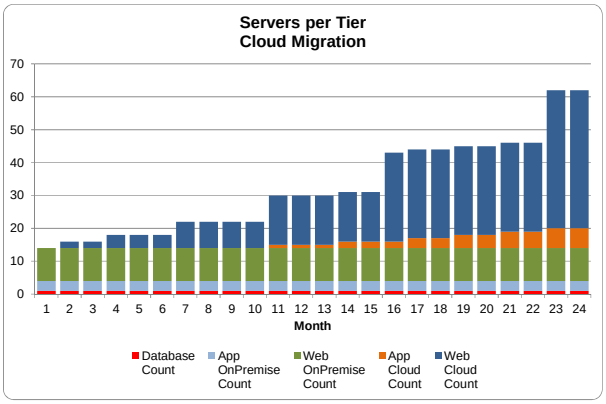

Figure 14. Servers per Tier

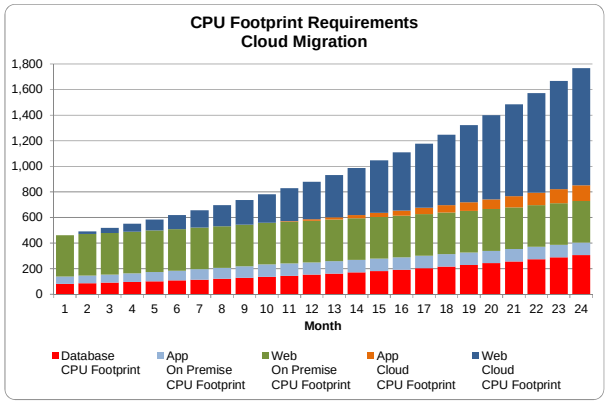

Figure 15. CPU Footprint Requirements

At this point the Infrastructure planners forward their predicted demand to the Facilities planners.

4.5 Cloud: Facilities

The new cloud-based approach satisfies the power constraints in the data center. Figure 16 shows the Facilities results from the new hybrid cloud solution; power usage within the data center has been capped; it stays flat.

At this point, positive feedback flows up through the Stack from Facilities to the Business. This feedback describes how demand was satisfied.

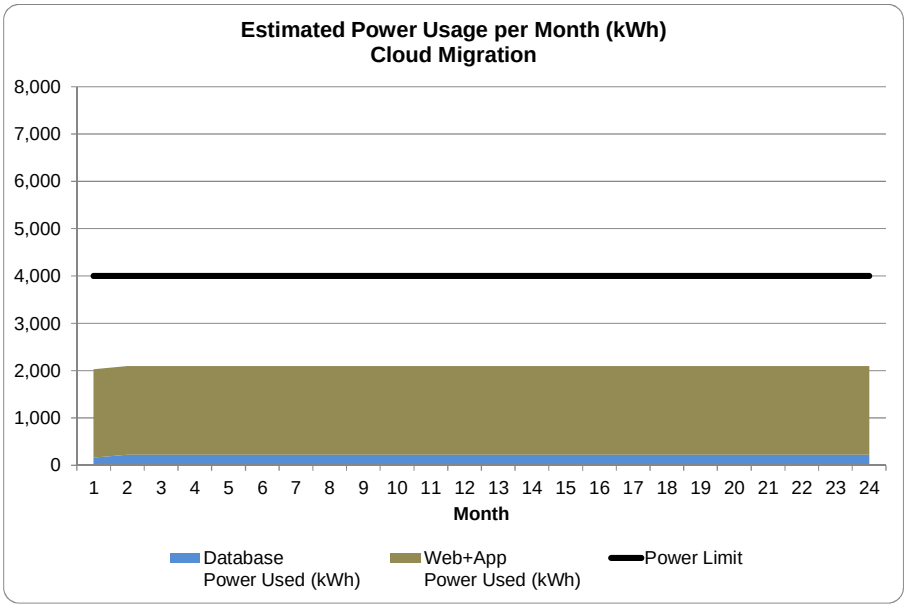

Figure 16. Facilities Efficiency Metric

4.6 Case Study Summary

Our case study illustrated the use of the Capacity Planning Stack. Demand started at the Business and flowed down through the five Stack levels. The feedback mechanism was used to identify and resolve a capacity problem. After successful navigation through the Facilities level, feedback passed back up to the Stack to provide information for efficiency tracking. This section will compare and contrast some of the work products produced from the Capacity Planning Stack.

CPU Footprint and Capacity: The capacity and usage charts for the two scenarios are shown in Figures 17 and 18.

- The usage curves were developed by the Application and Shared Services planners. These curves described the CPU footprint that the Infrastructure planners were required to satisfy. The usage curves for each scenario are identical; this case study did not alter Business demand or the requirements at the Application or Shared Services levels, the Infrastructure planners simply chose a different hosting solution.
- There is a noticeable difference in the two curves labeled "CPU Footprint Available":
    - The On Premise curves show the four upgrade points where capacity increased; this is where new servers were added to the infrastructure.
    - The Cloud curves show more frequent increases in capacity. Since the cloud supports on-demand provisioning, it made more sense to acquire cloud instances as needed.
- The headroom available in both hosting solutions is about the same. Headroom is the area between the "Used" and "Available" curves. The Infrastructure planners utilized a 30% headroom requirement.

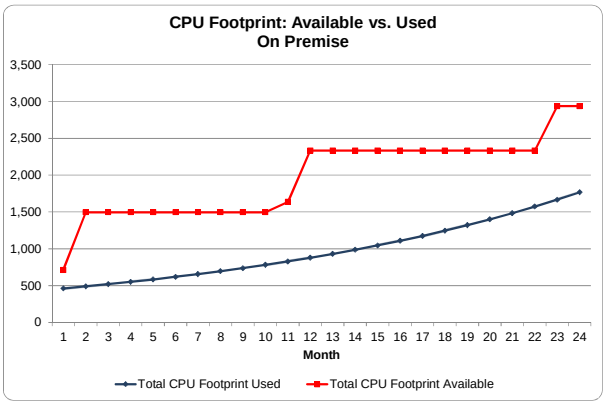

Figure 17. On Premise CPU Footprint

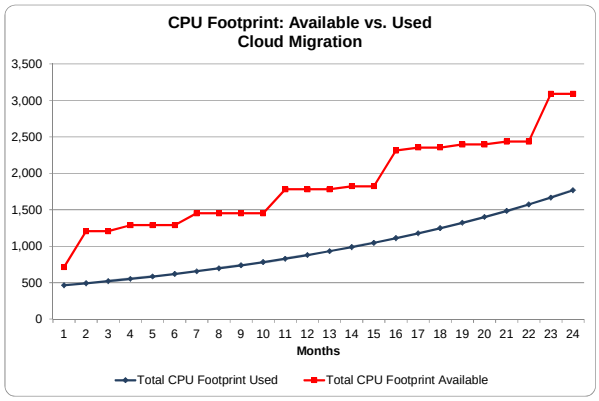

Figure 18. Cloud CPU Footprint

Monthly Cost. Figures 19 and 20 show the estimated monthly cost for each solution. Factors included in the costing model were:

- Infrastructure - New hardware purchase (CAPEX)
- Infrastructure - Cloud instance rental (OPEX)
- Facilities - Power and cooling (OPEX)
- Application & Shared Services - Software licenses (OPEX)

Consider the following when comparing these two charts side by side:

- The On Premise solution has larger CAPEX costs; the purchase of new servers at four points during the planning scenario.
- OPEX costs are significantly larger for the Cloud solution since you are now effectively renting servers in a remote location (i.e., the cloud). In addition, since you are gradually scaling out the infrastructure in the Cloud, the monthly OPEX costs are increasing month to month.
- *There is one factor missing from the On Premise chart – data center upgrade (potentially millions of dollars).* If the On Premise solution was selected, then a data center upgrade would be required to expand the power capacity constraint identified earlier.

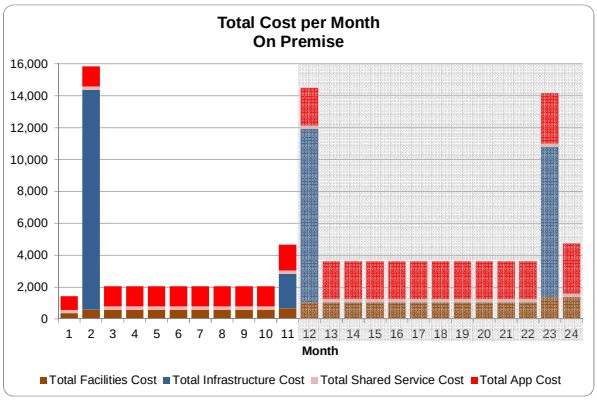

Figure 19. On Premise Monthly Cost

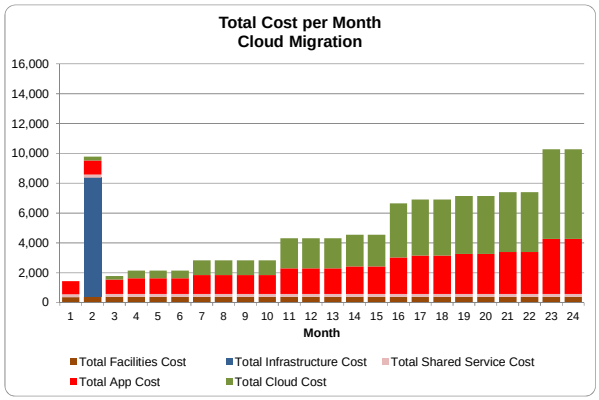

Figure 20. Cloud Migration Monthly Cost

Business Transactions per Digital Infrastructure Dollar. The side by side solution comparison of the cost per Business transaction is shown in Figures 21 and 22. The two metrics required to produce this efficiency metric are:

- Number of Business transactions per month (this is the Business' primary demand factor).
- Total cost for the Digital Infrastructure supporting the Business (aggregation of cost feedback up through all levels of the Capacity Planning Stack).

These two charts show the number of business transactions that can be processed per Digital Infrastructure dollar over the 24-month planning horizon.

Observations from a side by side comparison in Figures 21 and 22:

- Both solutions start at the same efficiency point for the first month.
- The On Premise chart does not include the extra cost required for a data center upgrade (this is the gray area of the chart).
- Each dip in the curves corresponds to an increase in the server/instance count.
- The On Premise curve dips when new servers are purchased, but the ongoing costs only include power (electricity) costs for IT and facilities.
- Efficiency for the Cloud solution is trending slowly downward. As the Business volume increases by 6% per month, the number of cloud instances required to handle the load must also increase. The cause for decreasing efficiency is recurring monthly cost of cloud instances.
- The Cloud solution traded decreasing Business efficiency for the cost of a data center upgrade.

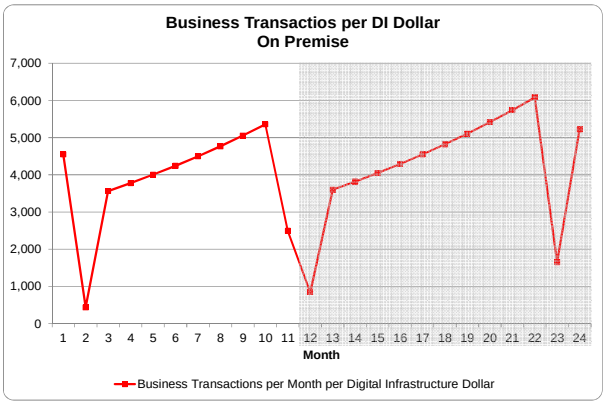

Figure 21. On Premise - Business Efficiency

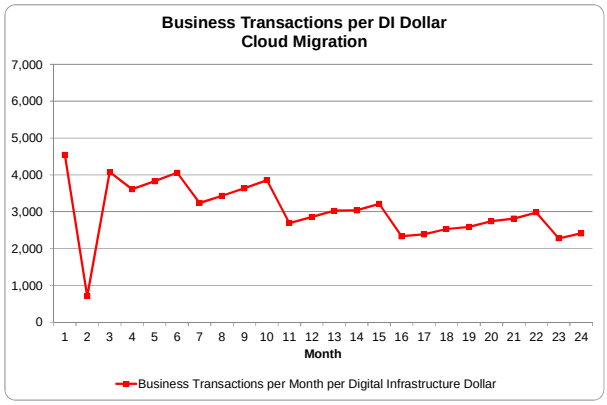

Figure 22. Cloud - Business Efficiency

5 Why is this Approach More Suited to Today's Digital Infrastructure

In this paper we introduced a new structured method for capacity planning. Our methodology is based on the *Capacity Planning Stack* that incorporates all components in the Digital Infrastructure, organizing them into a cohesive and comprehensive planning paradigm.

The Capacity Planning Stack organizes the Digital Infrastructure components in a multi-level hierarchy that supports capacity planning workflow from the Business to Facilities (i.e., the data center). The Stack hierarchy describes dependencies and communication between levels (demand and feedback). In addition, each Stack level has a set of efficiency metrics that can be used for long term tracking (measures of success). The entirety of the Stack supports a straightforward and transparent communication mechanism for a complete assessment of the Digital Infrastructure.

The case study illustrated how the Capacity Planning Stack can be used to translate Business requirements into lower level requirements for the Application, Shared Services, Infrastructure and Facilities. The case study also demonstrated the feedback mechanism that can be used when requirements cannot be satisfied. And finally, the cumulative feedback after all requirements have been satisfied provides the guidance required for ongoing monitoring and efficiency reporting at each level of the Stack.

The Capacity Planning Stack offers a new way to organize and think about today's capacity planning tasks. Two notable characteristics of the Stack are:

- Separation of Application/Shared Services planning from Infrastructure planning. There are a number of capacity planning practices today that have this separation. This approach is recommended so that each planning organization can focus on what they do best. In addition, the separation enables the Infrastructure planners to utilize new, innovative and cost effective hosting solutions where appropriate.
- Including Facilities in the planning process. It is not unusual to hear about data centers that are reaching capacity limits (as we saw in the case study). Conversely, there are a number of data centers that are overbuilt. The Stack's demand/feedback loop between Infrastructure and Facilities planners facilitates working together to create a Digital Infrastructure that is sized appropriately for the Business (not too big, and not too small).

The practice of capacity planning must adapt to today's changing Digital Infrastructure. Traditional approaches are no longer sufficient. The methodology described in this paper is a step towards revolutionizing how we approach, discuss and implement a capacity planning practice.

6 References

[CHAN1985] Mani Chandy, Acceptance speech for the A. A. Michelson Award at the CMG 1985 International Conference, December 1985.

[COCK2006] Adrian Cockcroft, "*Utilization is Virtually Useless as a Metric!*", CMG 2006 International Conference, December 2006.

[DICT2013] The Free Dictionary, http://encyclopedia2.thefreedictionary.com/technology+stack.

[SPEL2008] Amy Spellmann, Richard Gimarc, Charles Gimarc, "*Green Capacity Planning: Theory and Practice*", CMG 2008 International Conference, December 2008.

[SPEL2012] Amy Spellmann, Shawn Novak, Richard Gimarc, Michael Musso, "*Digital Infrastructure Planning for Private Cloud*", Journal of Computer Resource Management, Issue 131, Winter 2012 (a publication of the Computer Measurement Group).

[TECH2013] WhatIs.com, http://whatis.techtarget.com/definition/solution-stack.

[TGG2007] The Green Grid, "*The Green Grid Data Center Power Efficiency Metrics: PUE and DCiE*", 2007, www.thegreengrid.org.

[WIKI2013] Wikipedia, http://en.wikipedia.org/wiki/OSI_model.

[ZINK2013] Ken Zink, "*Efficiency, Optimization and Predictive Reliability - CA Technologies Capacity Management Resource Scoring Explained*", CA Technologies, 2013, www.ca.com.