Presented at the Computer Measurement Group Conference, November 2015

The Languages of Capacity Planning: Business, Infrastructure & Facilities

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

Richard Gimarc
CA Technologies
richard.gimarc@ca.com

Capacity planning for today's Digital Infrastructure demands collaboration between the business, infrastructure and facilities. In general, those silos have a dysfunctional relationship. What's needed is a more synergistic relationship that supports capacity planning in today's cloud and converged IT service delivery environments.

One of the major challenges is language; each silo has its own terminology and vocabulary. The business uses words such as customers, revenue, cost and reputation. Application planners talk about performance, response time and transaction volumes. IT's focus is on utilization, availability and hardware procurement. And finally, facilities' view is in terms of power, space and cooling.

How does an organization translate and blend these different views, metrics and languages into a coherent description of IT service delivery? This paper describes a communication plan and common language that promotes capacity planning across today's Digital Infrastructure.

1 Introduction

Today's capacity planner must work with all areas of an enterprise's Digital Infrastructure which includes the business owners, the application teams, the infrastructure group and facilities. In a sense, this is good news. As the central point of contact for planning and coordination, the capacity planner is well-positioned to ensure that there is sufficient capacity across the breadth and depth of the Digital Infrastructure to satisfy business demand in a cost-effective manner. However, there is a challenge; how do you speak to those different areas using terminology and language that they understand? It is a simple matter to hear that the business' new Shopping application is intended to support 100,000 customers. The capacity planner's challenge is to translate that demand into the metrics and information required to develop a comprehensive plan. How do you translate 100,000 customers into application transactions, virtual machines, database instances and ultimately rack space and power requirements in the data center?

The Capacity Planning Stack (a.k.a. the Stack) was introduced in [SPEL2013] as a way to simplify, structure and focus the practice of capacity planning for today's Digital Infrastructure. The Stack organizes the task of Digital Infrastructure capacity planning into a multi-level hierarchy (see Figure 1). The hierarchy starts at the Business and progresses through the enterprise-wide computing environment. The case was made that the Stack supports a methodology for capacity planning that provides better coverage for today's Digital Infrastructure than the tried-and-true traditional methods that have evolved over the past 30+ years.

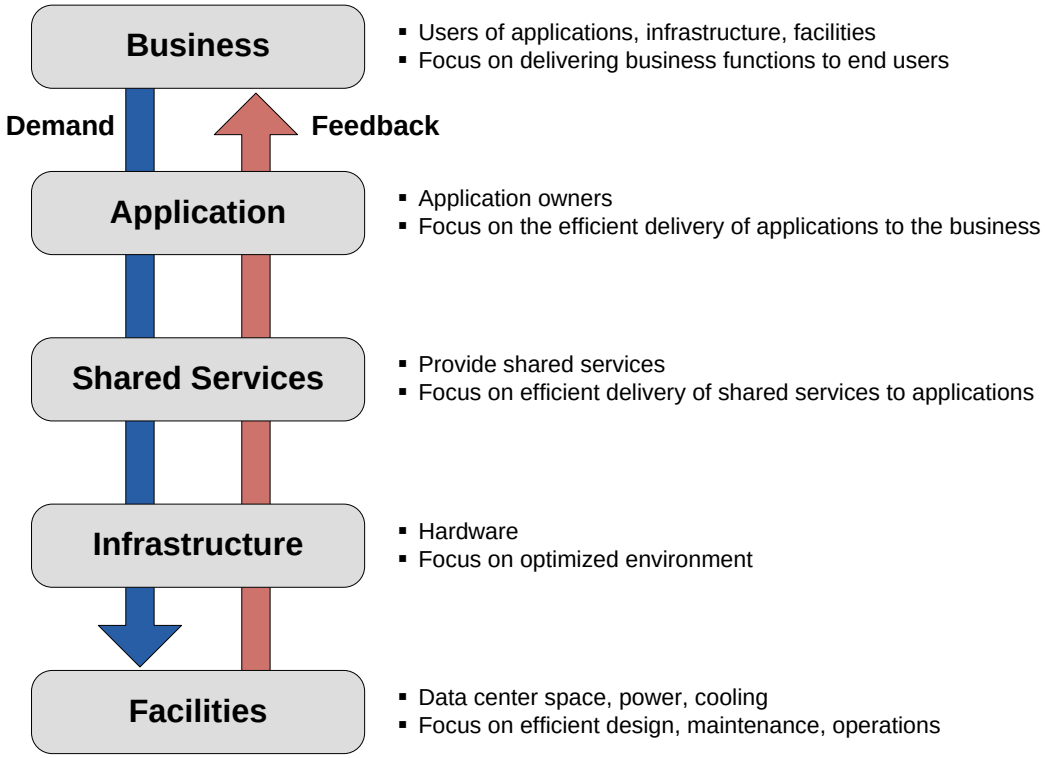

Figure 1. The Capacity Planning Stack

In this paper we take a closer look at the communication between the levels of the Stack as well as how the Stack can be applied to a variety of IT delivery models: owned and operated, hosted, cloud-based or any combination of execution venues. While communication of IT service demand and feedback may vary across delivery models, a*ny execution venue can be planned and managed using the Stack.* As IT evolves to a model of ITaaS (IT-as-a-Service) the Stack remains the critical framework for organizing and guiding the capacity planning process.

2 Stack Refresher

Before jumping into the details of using the Stack to support different execution venues, this section will take a quick view of the following basic characteristics and capabilities of the Stack:

- Stack Workflow
- Stack Metric Taxonomy
- Natural Forecasting Units
- Planning Horizon
- Delivery vs. Planning
- Budget and Cost

2.1 **Workflow - Capacity Planning Stack**

The Stack supports a capacity planning workflow of demand (down) and feedback (up). In addition, each Stack level produces a set of work products (efficiency metrics) that can be used for long-term tracking.

A refined diagram of the Capacity Planning Stack is shown in Figure 2.

- The Shared Services level of the Stack has been eliminated to simplify our discussion in this paper (assume that Shared Services is part of the Application or Infrastructure level).
- Efficiency (output) metrics are shown for each Stack level.

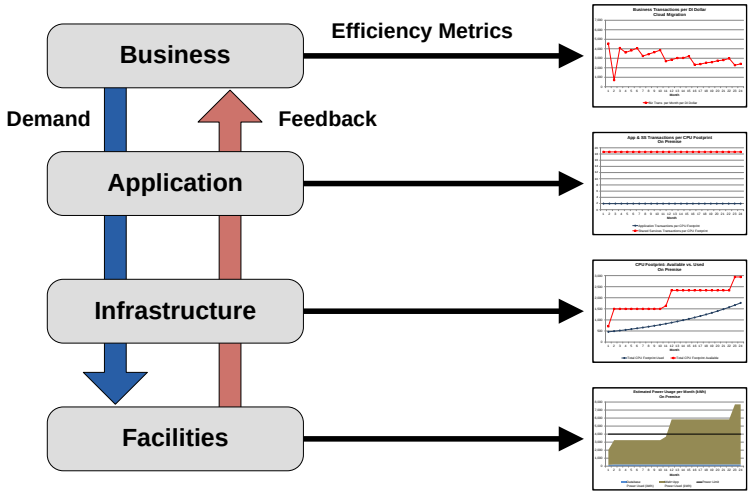

Figure 2. Refined Capacity Planning Stack

The three main workflows in the Stack are Demand, Feedback and Efficiency Metrics.

- **Demand flows down the Stack.** Business owners initiate and drive the capacity planning process. They pass their business requirements (e.g., support a new Shopping application with an expected customer base of 100,000) down to the Application level. The application planners translate the Business demand into application-level requirements (e.g., application transaction volumes, resource and performance requirements). The next step is for the Application level to pass their application requirements (demand) to the Infrastructure level. The infrastructure planners perform a similar task; translate the Application demand into infrastructure requirements that includes hardware, expected performance, headroom and SLAs. If additional hardware is required, they will make plans to initiate the procurement process. The final step is for the infrastructure planners to pass their demand to Facilities. The facilities team will evaluate the Infrastructure demand to determine if their data center has sufficient power, space and cooling to support any change in infrastructure. In addition, they may have to make adjustments within the data center to satisfy their own uptime and SLA requirements. At this point the Business demand has passed down through each level that supports the Digital Infrastructure.

- **Feedback flows up the Stack.** The feedback loop is used as the vehicle to describe what must be done to satisfy the demand at each level of the Stack. As an example, the feedback from Facilities will describe what needs to be done in the data center to support the Infrastructure demand. Facilities feedback may be simple (install a new rack of servers) or complex (upgrade

the data center's power feed to accommodate the additional power demand of the new servers). Feedback percolates up the Stack from level to level.

- **Efficiency metrics at each Stack level.** Efficiency metrics are generated at each level to document, describe and 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 SLAs. Infrastructure planners might use charts that show long-term trends of available headroom in their server and storage subsystems.

2.2 Stack Metric Taxonomy

There are a large (and overwhelming) number of Digital Infrastructure metrics that support capacity planning. In order to assist the capacity planner, a taxonomy of Stack metrics was developed and presented in [GIMA2014] that organizes and delineates the data requirements for each Stack level. The intent of the taxonomy was to provide the capacity planner with a way to view and discuss data requirements in a manner that is consistent with the Capacity Planning Stack.

Figure 4 illustrates the Stack taxonomy. The three main metric category are:

- **Business** - Metrics that describe the business as a whole (e.g., customer count and revenue) and the workload placed on the Digital Infrastructure (e.g., view account balance transactions). Traditional Natural Forecasting Units (described in the next section) are within the Business category.
- **IT** - Capacity planners are most familiar with this category which includes metrics such as CPU utilization and I/O rates.
- **Facilities** - The final category includes the metrics that describe power, space and cooling within the data center hosting the IT equipment.

A Venn diagram is used to illustrate the overlap and dependencies between metric categories.

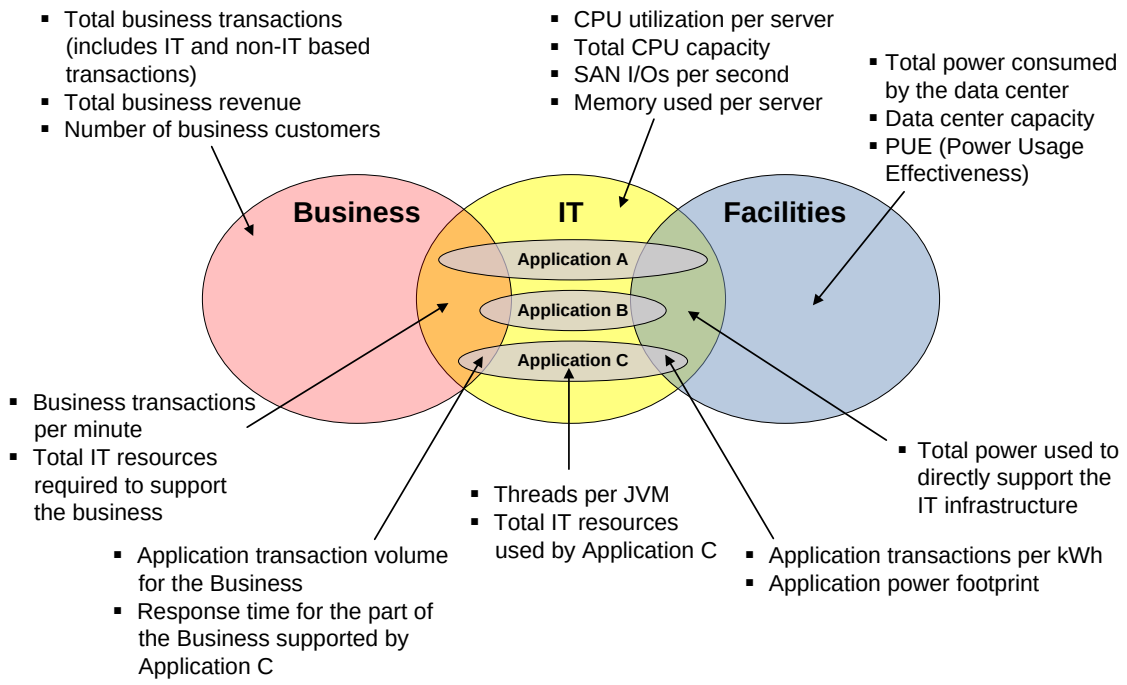

Figure 3. Fully populated Stack taxonomy

Error! Reference source not found. shows the perfect world. The taxonomy serves as a template for the capacity planner; it helps him identify, categorize and organize all the metrics available and those that are required to do his job. However, he is often faced with something less than perfect where the different groups within the Digital Infrastructure effectively function as silos. Today's capacity planner must find a way to create a more synergistic relationship that supports capacity planning across the entire Digital Infrastructure.

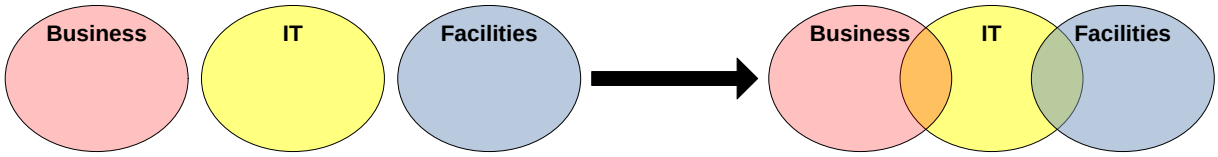

Figure 4. Evolving view of capacity planning metrics

Metrics and relationships must be identified that support the cascading effect of Business demand through Applications, IT and Facilities. This point cannot be overemphasized because it is precisely the intersections seen in the taxonomy that supports effective capacity planning.

2.3 Natural Forecasting Units

The Business demand used to initiate a capacity planning exercise is usually expressed in terms of a Natural Forecasting Unit (NFU). What makes the metric *natural* to the Business is that it is quantifies workload demand in terminology that they are familiar with. Consider the following examples of Business NFUs:

- Mortgage company: Number of existing loans & number of new loans
- Hotel chain: Number of hotels and beds & number of registered guests
- Insurance company: Number of current insurance policies & number of claims per month

NFUs were originally used by capacity planners to track infrastructure resource usage by workload (see [LO1986] and [REYL1987]). Take the mortgage company an example. The capacity planner's first task is to determine the relationship between the number of loans (NFU) and infrastructure resource usage (e.g., CPU, I/O and memory). The capacity planner would then use that relationship to predict the infrastructure required to support an increasing workload volume (e.g., add 5 quad-core servers). The primary challenge is translation; how do you translate Business NFUs into infrastructure resource demand.

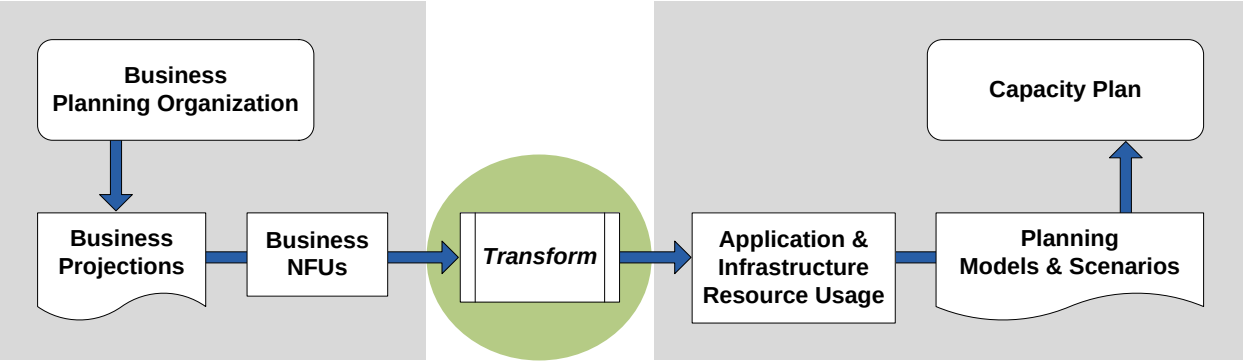

Figure 5. NFU flow in Capacity Planning

The Stack leverages the concept of an NFU in the sense that each level of the Stack has their own notion of a Natural Forecasting Unit and it is their NFU demand that is passed down the Stack from level to level.

The following table provides examples of the NFUs used by each level of the Stack.

	NFU Demand Factors (↓)
Business	NFU: Business volumetrics (e.g., number of loans)
Application	Task: Translate Business NFUs to application architecture, workload & metrics NFU: Application resource footprint, instance count and logical resources such as VMs, JVMs and threads
Infrastructure	Task: Translate Application NFUs into physical Infrastructure requirements NFU: Physical hardware requirements (e.g., servers, storage and network)
Facilities	Task: Translate Infrastructure NFUs into Facilities space, power & cooling NFU: Power draw, space & cooling requirements

Each level in the Stack uses their own language to describe NFUs. The challenge for the capacity planner is to be able to translate incoming demand NFUs into demand NFUs that are passed to lower levels. For example, how does the application capacity planner translate the number of loans into logical resource requirements such as VMs and JMVs that are passed to the infrastructure planners?

2.4 **Planning Horizon**

A capacity planning exercise is generally described in terms of a planning horizon. For example, the Business' new Shopping application may be stated as follows:

- *We need to be able to support our new Shopping application that will eventually support 100,000 customers.*
- *Our initial rollout will be 1,000 customers.*
- *We expect 5,000 after 3 months, 30,000 customers at the 6-month mark and the full 100,000 customers at the end of the year.*

Since the Business is looking forward for 12 months, does that mean that the other Stack levels should also adopt a 12-month planning horizon? The simple answer no, but the other Stack levels need to adapt their planning horizons to satisfy the Business demand. Each Stack level has their own planning horizon that is well-suited to meeting the needs (and demands) of the environment they manage. Those planning horizons must be coordinated to satisfy the Business demand.

The following table describes some of the differences and factors that guide and influence planning horizons.

Business	Business planning horizons are generally in the range of 6 months to 1 year. Factors that influence their planning horizon include: ▪ New application deployment ▪ Seasonal fluctuations (e.g., Black Friday and Cyber Monday) ▪ Acquisitions and mergers ▪ Organic workload growth
Application	The Application planning horizon is generally in the 3 to 6 month range. Factors that influence their horizon include: ▪ New application rollout ▪ DevOps ▪ Organic workload growth for existing applications
Infrastructure	Infrastructure planning horizon is also in the 3 to 6 month range. Although the Infrastructure and Application levels have similar planning horizons, the factors that drive them are different. The factors that drive the Infrastructure planning horizon include the following: ▪ Technology refresh ▪ Application support ▪ Capacity demand ▪ Procurement timelines

Facilities 	Facilities planning horizons are in the range of 5 to 10 years. Since they are concerned with the hosting data center they want to plan for as few changes as possible because change is expensive and time consuming. Changes to support additional power, space and cooling are not small or incremental; that is the primary reason for their long-term view. A consequence is that Facilities generally builds in more room for growth than the other Stack levels.

It should be clear that each Stack level has their own set of factors that drive their planning horizon. The important point is that the Stack must adapt to the Buisiness' timeline and planning horizon.

2.5 **Delivery vs. Planning**

The capacity planning group does not have delivery responsibility. Instead, delivery is the responsibility of the Application, Infrastructure and Facilities teams.

The capacity planning group's role is to plan and coordinate business service delivery across the Digital Infrastructure. It is critical for the capacity planners to have a working relationship with each of the Stack levels to plan for the successful delivery. Coordination results from the feedback that percolates up the Stack.

Demand flowing down the Stack will include timelines; how soon does each level require its demand to be satisfied. The feedback loop will contain information about when those demands can be satisfied. Coordination can be accomplished by scheduling the expected delivery timelines for each Stack level.

2.6 **Budget and Cost**

One of the important factors to consider during the capacity planning process is cost; how much will it cost to support the Business demand.

- One of the Business demand factors is **budget**. The Business' budget is the amount of money they are willing to invest to support their new business service (e.g., their new Shopping application). As demand flows down the Stack, each level will add detail to the budget to estimate how much of the overall budget they will require. As you might expect, if the budget ever reaches zero, then an immediate feedback loop will be initiated up the Stack. Alternative implementation choices may have to be made at higher levels or the budget may have to be increased. After the adjustment is made, the demand flow resumes.

- **Cost** is part of the feedback up the Stack. As each level adds their line items to the budget, the resulting cost will percolate up the Stack to the Business. Cost will describe the investment required by each level of the Stack and provide an estimate of the overall cost to support the Business demand. For example, the cost feedback from the Infrastructure level will describe to the application planners the expected investment for both Infrastructure and Facilities to support their Application demand.

3 Mapping IT Service Delivery Alternatives to the Stack

Originally the concept of the Stack was targeted at traditional owned and operated types of environments; however, it also applies to a variety of other delivery models: owned and operated, outsourced, hosted and cloud (IaaS, PaaS, SaaS and ITaaS). The Capacity Planning Stack applies to any mix of external IT services and internal IT management as we will demonstrate with the IT service delivery models discussed in this section. In all cases, the Business (with feedback from the other levels) makes the final decision regarding the Best Execution Venue [BROO2014] for IT services, based on cost, time to market, scalability and in-house skills. Management of external resources varies across the Application and Infrastructure Stack levels.

3.1 Hosted IT Service Delivery

Many organizations have been outsourcing IT for decades; this is not a new execution venue. However, it is still appropriate to describe how the Stack applies and is used for capacity planning in a hosted or outsourced environment. The Business and Application levels look very much the same and remain in-house. IT is typically still part of the in-house responsibility; they partner with the hosting provider who must plan for their own IT footprint with their Facilities team. Additionally, there will always be a portion of the Infrastructure that must remain in-house, namely networking (which includes internal LAN and access to the external hosted environment) and on-premise support infrastructure as a minimum.

Figure 6 shows the mapping of the hosted environment to the in-house Capacity Planning Stack.

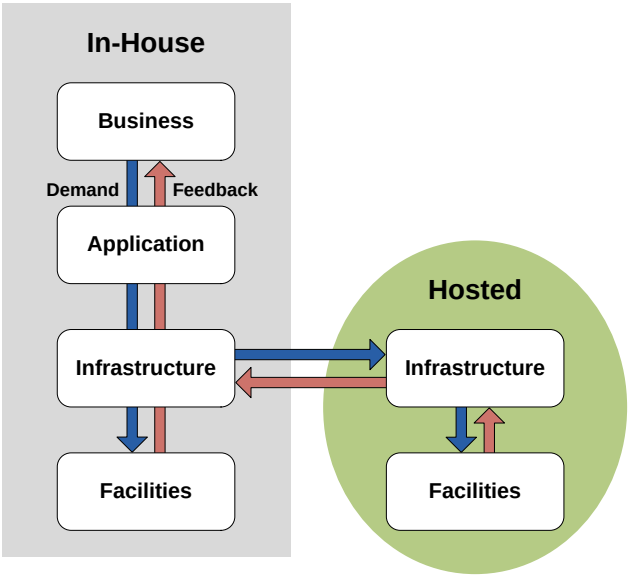

Figure 6. Stack Mapping for Hosted IT Service Delivery

In this execution venue, the Stack's workflow is augmented with a demand/feedback flow to/from the hosting provider. Business to Application workflow remains the same as in an owned and operated model. However, the Application to Infrastructure flow now has to occur via an internal Infrastructure team as well as an external team residing with the hosting provider.

Typically the internal infrastructure team is still responsible for capacity planning and interfacing with the application planners, but they now have the additional responsibility to communicate hardware requirements with the external hosting team. Communication between the internal and external infrastructure teams is simplified because they both speak the same language. The internal team's demand NFUs match the terminology used by the hosted Infrastructure team; they both talk about servers, storage and network components.

The internal team will monitor and manage both the internal and external IT systems, as well as define the plan and timeline for new server, storage and networking hardware needed in the hosting facility. The communication between in-house IT and the hosting provider requires the application and infrastructure teams to translate compute demands into server, network and storage (as in the traditional model):

- Type of server(s), processing speed and capacity, quantity
- Network traffic load and bandwidth requirements
- Storage capacity and performance required to support each application or business service

The external hosting provider responds with estimated, then actual costs for the space, power and cooling required (as well as any additional services, if applicable), while the internal IT team will continue to manage and monitor the performance, capacity and availability of the infrastructure.

The external infrastructure team must handle the capacity planning for their network and any hosted hardware. The responsibility of communicating with the hosting facilities team now lies completely within the confines of the hosting provider. The external infrastructure team must collect capacity requirements and facilitate communication with their data center operators by predicting and projecting when new hardware infrastructure will need to be deployed and helping to quantify the space, power and cooling requirements at the Facilities level (for all hosted or collocated infrastructures in the hosting data center).

Managed Service Providers (MSP) would fall into this category as well, though more of the infrastructure planning would fall on them versus the internal infrastructure team. In the MSP model, monitoring and management of all of the IT systems often falls to the provider.

3.2 **IaaS/PaaS**

In the public cloud, IaaS and PaaS look much like the Hosted IT Service Delivery model described in the previous section. The primary difference is the terminology used for communication between the internal infrastructure team and the cloud provider. Instead of capacity planning in terms of physical hardware components, the internal team works with the applications planners to specify the number of server "instances", storage size and network bandwidth required for their applications. Additionally, the internal infrastructure team handles capacity planning for network traffic to/from the cloud.

The internal team has no insight into the actual hardware running in the cloud. The application planners continue to specify their Infrastructure demand for infrastructure resources, but the internal infrastructure planners translate those requirements into the language of the IaaS/PaaS provider.

The communication between in-house and the IaaS/PaaS provider requires the application and infrastructure teams to translate compute demands into cloud instances:

- Type of Instance (e.g., AWS On-Demand, Reserved or Spot) based on scalability and capacity requirements and lifetime of the application or service
- Processing speed and capacity for the instances
- Network traffic load and bandwidth requirements

- Storage capacity and performance required for each instance for the application

The cloud provider responds with estimated, then actual costs and utilization, while the internal IT team must manage the instances, performance, capacity, and availability of the instances over the lifetime of the application. There are a number of tools in this growing market for tracking and managing public and private clouds (and that topic will be a target for future discussions).

Figure 7 shows the relationship between the Cloud Providers and the In-House IT Stack.

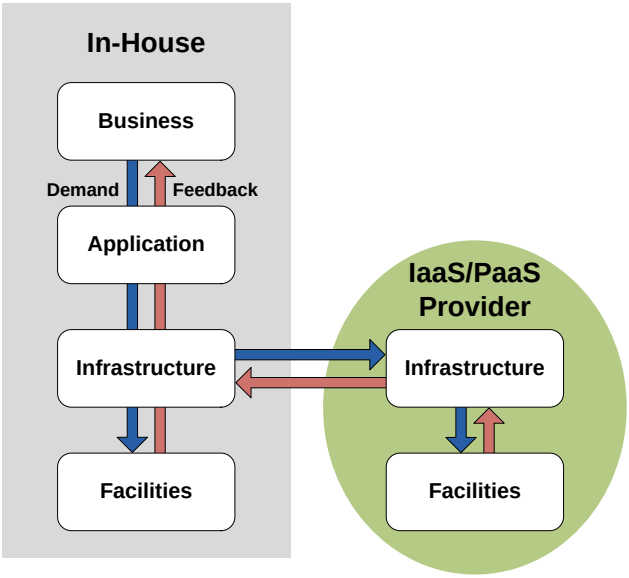

Figure 7. Stack Mapping for Cloud Provider Service Delivery (IaaS & PaaS)

Again, as with the hosting model, the internal, in-house Stack remains in place to support network connectivity to the cloud as well as internal network and support requirements for on-site infrastructure and users. Note that there are cases where the application team works directly with the IaaS/PaaS API's to stand up test and DevOps environments, but ultimately, management of the external cloud environments and instances must be the responsibility of the in-house infrastructure team. The (unmanaged) ease of standing up cloud instances can result in:

- Shadow IT (application teams acquiring cloud services without organizational approval),
- Underutilized instances (similar to server sprawl in the data center), can remain unused and completely forgotten but still costing money
- Increased complexity
- Poor management and capacity planning for the IT infrastructure
- Inefficient use of infrastructure incurring significant costs over time (cumulative monthly cloud instances charges)

These factors provide additional motivation for a central organization to manage all of IT at the Infrastructure level.

3.3 **SaaS**

In the SaaS model, the interface to the internal Digital Infrastructure lies with the Business (or Lines of Business). In this case, the internal application and infrastructure teams have little or no responsibility to monitor, manage or do capacity planning for the SaaS provider (except for the network connectivity). The Business would define requirements for the SaaS application in terms of:

- The number of users for the SaaS application
- The storage capacity required to support the user workload
- Any customization of the SaaS application required to meet Business requirements

The Business does not have any visibility into the SaaS provider's infrastructure. The in-house infrastructure team will handle the network monitoring and capacity planning for the pipe connecting the internal data center to the SaaS data center(s). Inside the SaaS provider, they may have their own data center, outsource, or utilize public cloud instances for their execution venue. In any case, the SaaS provider manages everything down the Stack in their execution environment.

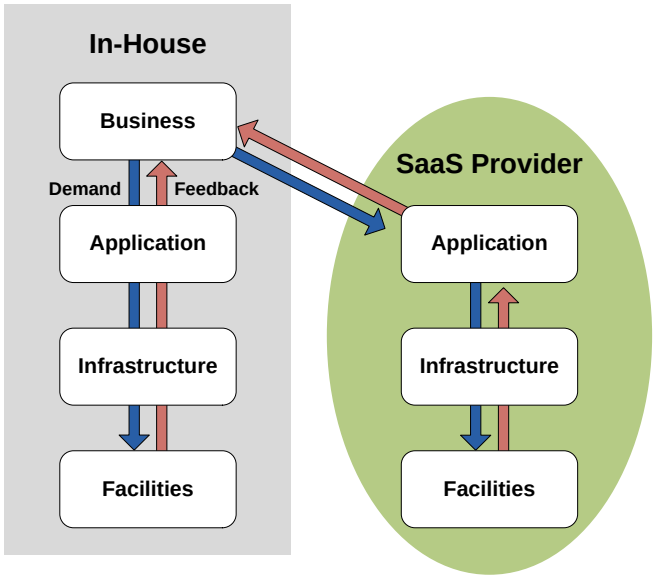

Figure 8. Stack Mapping for SaaS

Typically, the in-house application team is also involved in making the decision to utilize a SaaS offering, translating Business demands into SaaS requirements and evaluating the effort/cost of a home-grown solution. But ultimately, the Business makes the decision based on cost, availability, speed to deliver and focus of internal resources; the individual Business lines control the budget and the relationship with the SaaS provider. The feedback to the Business is the cost of the SaaS application, estimated and then actuals as the system goes live. This model limits the analysis of the performance and capacity of the application to the SaaS provider. In the ITaaS model (next section), the management of the SaaS solution moves inside the IT function to the Application level so that the capacity and performance analysis can be included within their IT responsibilities (in conjunction with the SaaS provider).

3.4 **ITaaS**

In the evolution of IT-as-a-Service, where many enterprises are evolving today, the focus of the internal teams turns to determining the Best Execution Venue for the required business services and applications.

- ITaaS is a competitive business model where an IT department views the Lines of Business as having many options for IT services, and the internal IT organization has to compete against external providers to serve these LOBs.
- The catalyst for ITaaS within enterprises is typically mobile service delivery and the creation of the "anytime, anywhere and any device" workplace as it introduces consumer IT trends into the enterprise.
- The IT department needs to start thinking in terms of the creation of a portfolio of services – some developed internally, some provided externally via dedicated contracts, and others sourced from public cloud providers.
- Tooling, automation, process and policy become paramount for long-term optimization and sustainability of the service portfolio. This is why service automation, integration and management are such growing requirements in the "new style" IT services market, where ITaaS is the ultimate goal.

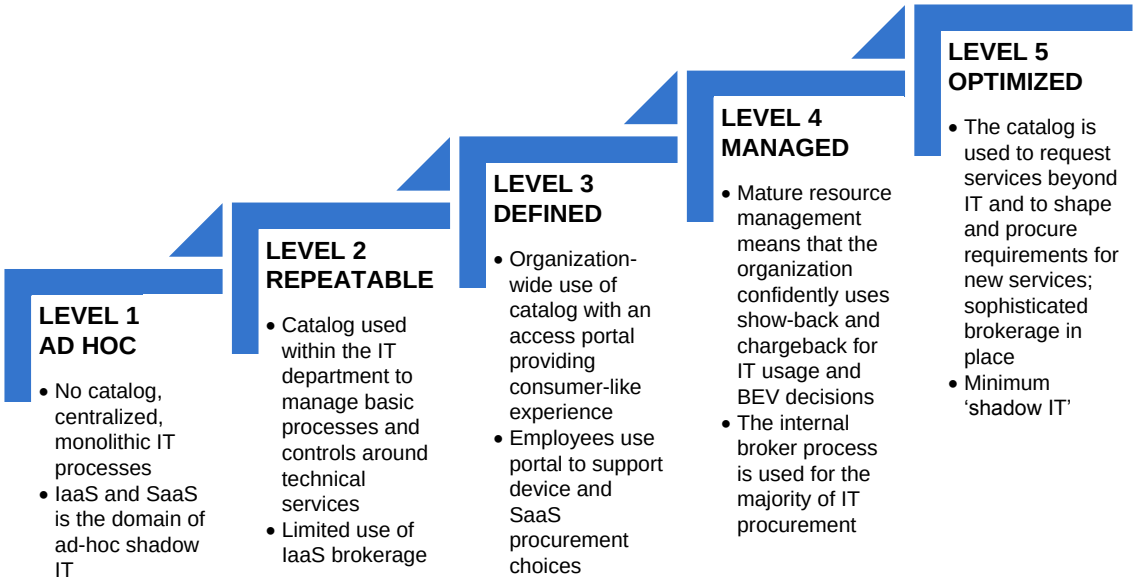

Figure 9. ITaaS Maturity Model (451 Research)

In the ITaaS model, the internal IT organization becomes responsible for working with all external IT service providers. In the same manner as described in the previous sections, the internal team is responsible for capacity planning across internal and hosted infrastructure (servers, network and storage) as well as cloud infrastructure (instances, storage and network capacity) and all network connections to outside providers. Figure 10 shows the relationships that support ITaaS and Best Execution Venues.

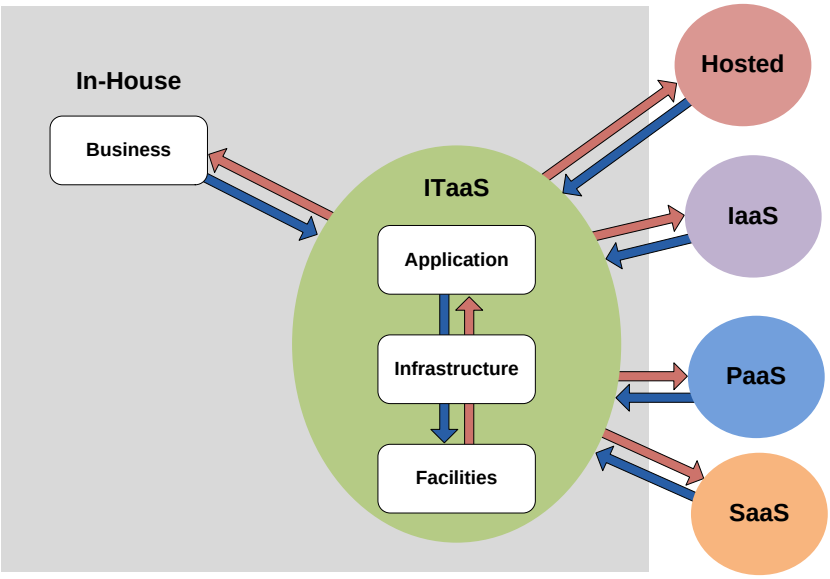

Figure 10. Stack Mapping for ITaaS

In this model, the SaaS connection to the in-house Digital Infrastructure is managed by the Application level instead of the Business as ITaaS drives the organization to manage all IT services, internal and external. Application, Infrastructure and Facilities work together to define the Best Execution Venue for specific business services, evaluating resource requirements, costs, management, operations, migration and staffing factors for venue choices. Ultimately, the Business, with decision support from IT, takes the responsibility for choosing the final execution venue, based on cost, risk, time to market, business goals and trends for future growth.

4 Summary

The Capacity Planning Stack was introduced in [SPEL2013] as a way to simplify, structure and focus the practice of capacity planning for today's Digital Infrastructure. In that paper we demonstrated how the Stack could be applied to capacity planning for an in-house and hybrid cloud environment. In this paper we extended to scope to include a variety of execution venues (e.g., SaaS and ITaaS). The significant consequence of the Stack is that it encourages (and demands) communication and coordination between the different areas within an organization; namely the Business, Application, Infrastructure and Facilities.

The basic structure of the Stack was not modified to support capacity planning for different service delivery alternatives. What did change were the points of communication between in-house and external service providers; but each of those players utilized their own subset of the Stack.

The basic Stack workflow of demand and feedback identifies and supports the required communication channels between the Business, Application, Infrastructure and Facilities whether they are all contained in-house or are distributed amongst a set of cooperating organizations. Furthermore, successful level-to-level communication requires that language and terminology differences be recognized, and it is the capacity planner's responsibility to be the facilitator for successful translation and communication.

Capacity planning for today's Digital Infrastructure requires a rigorous methodology, metrics and language for efficient IT Service Delivery. As ITaaS gains momentum, it is imperative for the capacity planner to view

this new world in terms of the Stack so that he can be the primary facilitator to recognize, adapt and translate the language used in Information Service Delivery. Navigating the evolving landscape will require capacity planners to emerge as leaders in the transformation from traditional IT to ITaaS.

5 References

[BROO2014] Carl Brooks, "*The Best Execution Venue; Positioning for the Next Wave of Change in Enterprise IT Infrastructure*", 451 Research, July 2014.

[GIMA2014] Richard Gimarc, Amy Spellmann & Adrian Johnson, "*Stack Metrics: A Taxonomy of Metrics Supporting the Capacity Planning Stack*", CMG International conference, November 2014.

[LO1986] T. L. Lo and J. P. Ellis, "*Workload Forecasting Using NFU: A Capacity Planner's Perspective*", CMG International Conference, December 1986.

[REYL1987] John M. Reyland, "*The Use of Natural Forecasting Units*", CMG International Conference, December 1987.

[SPEL2013] Amy Spellmann & Richard Gimarc, "*Capacity Planning: A Revolutionary Approach for Tomorrow's Digital Infrastructure*", CMG International Conference, November 2013.