The Languages of Capacity Planning: Business, Infrastructure & Facilities

Amy Spellmann; St. Louis CMG 4/19/16
The Languages of Capacity Planning:
Business, Infrastructure & Facilities

Amy Spellmann - amy.spellmann@451research.com
Richard Gimarc - richard.gimarc@ca.com

November 5, 2015
CMG 2015
Session 511
The Languages of Capacity Planning

The Challenge

**Capacity Planning Stack**
- Multi-level hierarchy
  - Demand (↓)
  - Feedback (↑)
  - Efficiency Metrics (→)
- Supports all elements of today's Digital Infrastructure
- Implementation is straightforward & transparent

**Challenge:** Formulate & communicate demand
- How does Application talk to Business?
- How do Infrastructure planners talk to Facilities planners?
The Languages of Capacity Planning

Our Approach

- Understand demand
- Revive the notion of Natural Forecasting Units (NFU)
- Use NFUs to express demand
- Apply this approach to a variety of execution environments
A Bit of History

2013  We presented the Capacity Planning Stack as a new way to view, analyze & communicate Digital Infrastructure capacity

2014  We presented a taxonomy that organizes the metrics supporting the Stack

Today  The Language of Capacity planning; utilizing the Stack to describe communication between the levels of the Stack for any service delivery model
The Capacity Planning Stack

- **Business**
  - Users of applications, infrastructure, facilities
  - Focus on delivering business functions to end users

- **Application**
  - Application owners
  - Focus on the efficient delivery of applications to the business

- **Shared Services**
  - Provide shared services (e.g., message brokers, databases)
  - Focus on efficient delivery of shared services to applications

- **Infrastructure**
  - Hardware (e.g., servers, storage, network)
  - Focus on optimized environment

- **Facilities**
  - Data center space, power, cooling
  - Focus on efficient design, maintenance, operations
Demand & Feedback

**Demand (↓)**
- Volume & priorities
- Logical resource requirements
- Performance requirements & SLAs
- Budget

**Demand (↓)**
- Physical resource footprint & instances
- Performance requirements & SLAs
- Budget

**Feedback (↑)**
- Cost
- Total time to satisfy
- Expected performance

**Diagram:***
- Business
- Application
- Shared Services
- Infrastructure
- Facilities
A Revolutionary Approach to Capacity Planning

The Capacity Planning Stack - a revolutionary approach to capacity planning that simplifies, structures and focuses the practice.
The Capacity Planning Stack Taxonomy

- Total business transactions (includes IT and non-IT based transactions)
- Total business revenue
- Number of business customers

- Business transactions per minute
- Total IT resources required to support the business

- Application transaction volume for the Business
- Response time for the part of the Business supported by Application C

- CPU utilization per server
- Total CPU capacity
- SAN I/Os per second
- Memory used per server

- Total power consumed by the data center
- Data center capacity
- PUE (Power Usage Effectiveness)

- Total power used to directly support the IT infrastructure

- Threads per JVM
- Total IT resources used by Application C

- Application transactions per kWh
- Application power footprint
Benefits of the Stack Taxonomy

- Provides a structured approach to evaluating today’s Digital Infrastructure
- Simplifies & guides the selection of metrics required to support decision making
- Equips the capacity planner with a structured way to think about, organize, communicate and collect the metrics that apply to the decision support process
- Paves the way for federated capacity planning across today’s Digital Infrastructure
### Planning Horizons

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Business</strong></td>
<td>Business planning horizons are generally in the range of 6 months to 1 year. Factors that influence their planning horizon include:</td>
</tr>
<tr>
<td></td>
<td>- New application deployment</td>
</tr>
<tr>
<td></td>
<td>- Seasonal fluctuations (e.g., Black Friday and Cyber Monday)</td>
</tr>
<tr>
<td></td>
<td>- Acquisitions and mergers</td>
</tr>
<tr>
<td></td>
<td>- Organic workload growth</td>
</tr>
<tr>
<td><strong>Application</strong></td>
<td>The Application planning horizon is generally in the range of 3 to 6 months. Factors that influence their horizon include:</td>
</tr>
<tr>
<td></td>
<td>- New application rollout</td>
</tr>
<tr>
<td></td>
<td>- DevOps</td>
</tr>
<tr>
<td></td>
<td>- Organic workload growth for existing applications</td>
</tr>
<tr>
<td><strong>Infrastructure</strong></td>
<td>Infrastructure planning horizon is also in the range of 3 to 6 months. 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:</td>
</tr>
<tr>
<td></td>
<td>- Technology refresh</td>
</tr>
<tr>
<td></td>
<td>- Application support</td>
</tr>
<tr>
<td></td>
<td>- Capacity demand</td>
</tr>
<tr>
<td></td>
<td>- Procurement timelines</td>
</tr>
<tr>
<td><strong>Facilities</strong></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>A consequence is that Facilities generally builds in more room for growth than the other Stack levels.</td>
</tr>
</tbody>
</table>
Natural Forecasting Unit (NFU)

- Business metric that can be related to the use of computer system resources (e.g., CPU, I/O, memory, network traffic)

- Examples: Number of accounts, customers, loans, hotel beds, cars manufactured, product orders, insurance contracts, ...

**Business Planning Organization**

**Business Projections**

**NFU**

**Transform**

**Computer Resource Usage**

**Planning Models & Scenarios**

**Capacity Plan**

*Link business plans to IT plans*
NFUs & the Stack

- Each Stack level has its own notion of a Natural Forecasting Unit
- NFU demand is passed down the Stack from level to level.

<table>
<thead>
<tr>
<th>Stack Level</th>
<th>NFU Demand Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Business</strong></td>
<td>NFU: Business volumetrics (e.g., number of loans)</td>
</tr>
</tbody>
</table>
| **Application** | Task: Translate Business NFU to application architecture & metrics  
                 NFU: Application resource footprint and instance count (logical resources such as VMs, JVMs and threads) |
| **Infrastructure** | Task: Translate Application NFU into physical Infrastructure requirements  
                      NFU: Physical hardware requirements (e.g., servers, storage and network) |
| **Facilities** | Task: Translate Infrastructure NFU into Facilities space, power & cooling  
               NFU: Power draw, space & cooling requirements |
NFU - Examples

**Mortgage company**
- Number of existing loans
- Number of new loans

**Bank**
- Number of accounts
- Number of customers using electronic bill payment

**Insurance company**
- Number of current insurance contracts
- Number of claims per month

**Hotel chain**
- Number of hotels & beds
- Number of registered guests

---

**Challenge**
- How do you transform an NFU into computer system resource usage?

**How much CPU, I/O, etc. are required to:**
- Service an existing home loan
- Create a new loan
- Generate monthly statement
- Process a customer claim
- Register a new guest
- Generate bill for checkout
### Best Execution Venue

*Where is the best place to run a workload?*

<table>
<thead>
<tr>
<th>External Public Cloud</th>
<th>Software-as-a-Service</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>External Public Cloud:</strong></td>
<td>Servers and storage on demand (AWS, Rackspace)</td>
</tr>
<tr>
<td><strong>Private Cloud:</strong></td>
<td>We own (or pretend we own) the servers and storage (Cloudstack, OpenStack, hosted private cloud)</td>
</tr>
<tr>
<td><strong>Hybrid:</strong></td>
<td>Private and public resources managed together as needed (RightScale, Dell)</td>
</tr>
<tr>
<td><strong>SaaS:</strong></td>
<td>Applications on demand (Salesforce, Google Docs)</td>
</tr>
</tbody>
</table>
Best Workload Execution Venue: Cloud

For each of the following categories of workload/business functions, what is your primary deployment method likely to be in the next two years (internal private cloud, external public cloud, hybrid cloud, or SaaS)?

- **Batch Computing Applications**: 68% Internal Private Cloud, 14% Hybrid Cloud, 14% External Public Cloud, 4% SaaS
- **Back-office Enterprise Applications**: 53% Internal Private Cloud, 18% Hybrid Cloud, 9% External Public Cloud, 20% SaaS
- **Customer-facing Enterprise Applications**: 53% Internal Private Cloud, 26% Hybrid Cloud, 13% External Public Cloud, 8% SaaS
- **Test and Development of Applications**: 42% Internal Private Cloud, 27% Hybrid Cloud, 27% External Public Cloud, 4% SaaS
- **Collaborative Applications**: 35% Internal Private Cloud, 17% Hybrid Cloud, 27% External Public Cloud, 19% SaaS
- **E-business Hosting**: 18% Internal Private Cloud, 43% Hybrid Cloud, 25% External Public Cloud, 14% SaaS
- **Cloud-native Applications**: 11% Internal Private Cloud, 23% Hybrid Cloud, 37% External Public Cloud, 29% SaaS
Business Demand Aligned to Best Execution Venue

- **Strategy**
  - Use Case
  - Business Demand
  - Fixed vs Variable Costs

- **Workload**
  - Best Execution Venue
  - Availability
  - Performance

- **Platform**
  - Optimization
  - Automation
  - Orchestration

- **Topology**
  - Owned vs Third Party
  - Availability of System

**Business**

**Application**

**IT Infrastructure**

**Facilities Infrastructure**

**DEMAND**

**DECISSIONS**

- $ / Risk / ROI
- $ / user & transaction
  - DL Transaction / $
  - Elasticity
  - Portability

- Best Platform Venue
  - $ / logical unit
  - Time to Implement
  - System Redundancy

- Topology Decisions
  - $ / logical unit
  - Efficiency
  - System Redundancy
Traditional

- NFU Demand (↓)
  - Business to Application
    - Business volumetrics
  - Application to Infrastructure
    - Resource footprint & instance count
    - Logical resources
  - Infrastructure to Facilities
    - Physical hardware requirements

- Feedback
  - Describe what will be implemented to satisfy the demand at each level

- Budget (↓) & Cost (↑)
Hosted

- Internal is identical to the traditional On-Premise model
- External communication of demand to hosting provider
  - Hardware requirements
  - Feedback is cost based on space, power and cooling at the provider’s site
IaaS can be Public or Private Cloud

NFUs & Communication/Language

- Internal is identical to the traditional model
- External communication of demand to IaaS provider
  - In-house Infrastructure to Cloud Provider
  - Physical hardware requirements
  - Feedback is cost based on instance usage & type, storage and network actuals
PaaS

- Internal is identical to the traditional model
- External communication of demand to PaaS provider
  - Application to PaaS Provider
  - Resource footprint & instance count
  - Logical resources
  - Primary use is development
  - Feedback is cost based on instance usage & type, storage and network actuals

PaaS can be Public or Private Cloud

NFUs & Communication/Language
SaaS

- Internal is identical to the traditional model
- In-house LOB to SaaS Provider
  - Number of users, size of data stores, network traffic requirements, any specific SaaS customization or add-on features
  - Feedback is cost based on actual users, special features, storage and network

SaaS will be external Cloud

NFUs & Communication/Language
IT-AS-A-SERVICE

Managing Capacity Across Best Execution Venues
What is ITaaS?

- **ITaaS is a business and operating model**
  Recognizes that lines of business (LOBs) have options for IT resources and services and the IT organization must compete for their business

- **ITaaS is enabled by a software-defined architecture**
  For private, public and hybrid clouds where infrastructure (e.g., servers, storage and networking) is virtualized, automated by software and delivered as a service

- **Uses self-service catalogues**
  Expose services (applications, tools, resources) to LOBs and users

- **IT organization acts as an advisor and broker**
  To recommend and/or curate additional resources and services, on-demand, to maintain resiliency and accommodate changing business needs

- **Performance metrics**
  Used to measure customer satisfaction and competitive positioning
ITaaS - Best Execution Venues

- **Every application has a best execution venue**
  - Some are mature, others are evolving
  - All are headed toward “the cloud”

- **Cloud computing is mainstream**
  - This means choice, access and diversity for IT
  - Benefits are material - business drives choice of venue

- **There are ways to systematically make the best choice of venue for each workload**
  - Enterprises are making these choices today
  - Providers are/should be targeting these workloads and creating the venues
  - There is a snowball effect across the IT landscape as cloud begets automation and automation begets growth
ITaaS - Maturity Model

LEVEL 1: AD HOC
- No catalog, centralized, monolithic IT processes
- SaaS andaaS is the domain of ad-hoc shadow IT

LEVEL 2: REPEATABLE
- Catalog used within the IT department to manage basic processes and controls around technical services
- Limited use of SaaS brokerage

LEVEL 3: DEFINED
- Organization-wide use of catalog with an access portal providing consumer-like experience
- Employees use portal to support device and SaaS procurement choices

LEVEL 4: MANAGED
- Mature Resource management means that the organization confidently uses showback and chargeback for IT usage and DCV decisions
- The internal broker process is used for the majority of IT procurement

LEVEL 5: OPTIMIZED
- The catalog is used to request services beyond IT and to shape and procure requirements for new services; sophisticated brokerage in place
- Minimum 'shadow IT'
ITaaS - Basic

- Business communicates with ITaaS
- ITaaS makes its own internal decisions based on
  - Business demand
  - ITaaS internal operation
  - ITaaS will choose the best execution venue
- Feedback is cost based on what is required to provide the satisfy the Business demand

ITaaS communication

NFUs & Communication/Language
ITaaS - Evolved

- ITaaS will choose the best execution venue

**Graphic Description**

- ITaaS communication
- NFUs & Communication/Language

**Diagram Elements**

- ITaaS
  - Application
  - Infrastructure
  - Facilities

- Hosted
- IaaS
- PaaS
- SaaS

- In-House
  - Business
Case Study: Global Fortune 500 Consumer Product Co.

Pre transformation state:
- Owned/operated multiple data centers
- Utilized third party managed services outsource provider
- Typical infrastructure design/develop/deploy times of 6+ months
- Traditional siloes of technology, development, operations and management
- All APP development efforts were bespoke standalone projects with no sharing of infrastructure

Cloud-First post transformation state:
- Started with consumer facing APPS deployed in private cloud.
- Experienced sticker shock of 10% rise in APP environment cost increase
- Version 2 deployed to AWS cloud technologies. On demand, elastic and turned OFF when not in use
- 20% reduction in TCO vs. pre transformed (non cloud) state
- Design/build times of less than 2 weeks start to finish
Previous Client Organizational Structure: Complex/Silos

- Client IT
- Design/Orchestration Mgmt
- Client Cloud Operations
- Client IT Operations
- Strategy Cloud Partners
- On Premise Cloud Partner
- VEND1 Partner
- VEND2 Partner
- Next Partner
- Legacy Ops Partner
451 Research Recommended IT Organization

Client ITaaS Strategy

Client ITaaS Operations

Hybrid IT Partner (Traditional/Private Cloud)
- Outsourcing?
- Legacy Owned
- Private Cloud
- DRaaS

Public Cloud Partner
- SaaS
- IaaS
- PaaS
- DRaaS

Best Execution Workload Partner
- Consumer Facing
- Enterprise Services
- Orchestration Execution
- Data portability

Hypervisor Partner
- Core Tools
- Automation
- Compute
The Languages of Capacity Planning

**Summary**

- Started with today's Digital Infrastructure
- Created the Capacity Planning Stack to describe the capacity planning process across the breadth and depth of the Digital Infrastructure
- Developed a taxonomy to organize our universe of metrics
- Revived the notion of an NFU - used as touch points between Stack levels
- Illustrated how the Stack can be applied to a variety of execution venues
- Final step was to show how using the Stack is critical for ITaaS
The Languages of Capacity Planning
What does this mean to you?

✔ Use the Stack to think about, discuss & approach capacity planning – all execution venues
✔ Leverage the Stack taxonomy to organize and describe your metric requirements
✔ Think NFU
✔ Language “challenges” can often be traced back to NFU usage & assumed understanding
✔ Planning for the cloud fits into the Stack’s framework

✔ Navigating the evolving landscape will require capacity planners to emerge as leaders in the transformation from traditional IT to ITaaS
Make an IMPACT in La Jolla