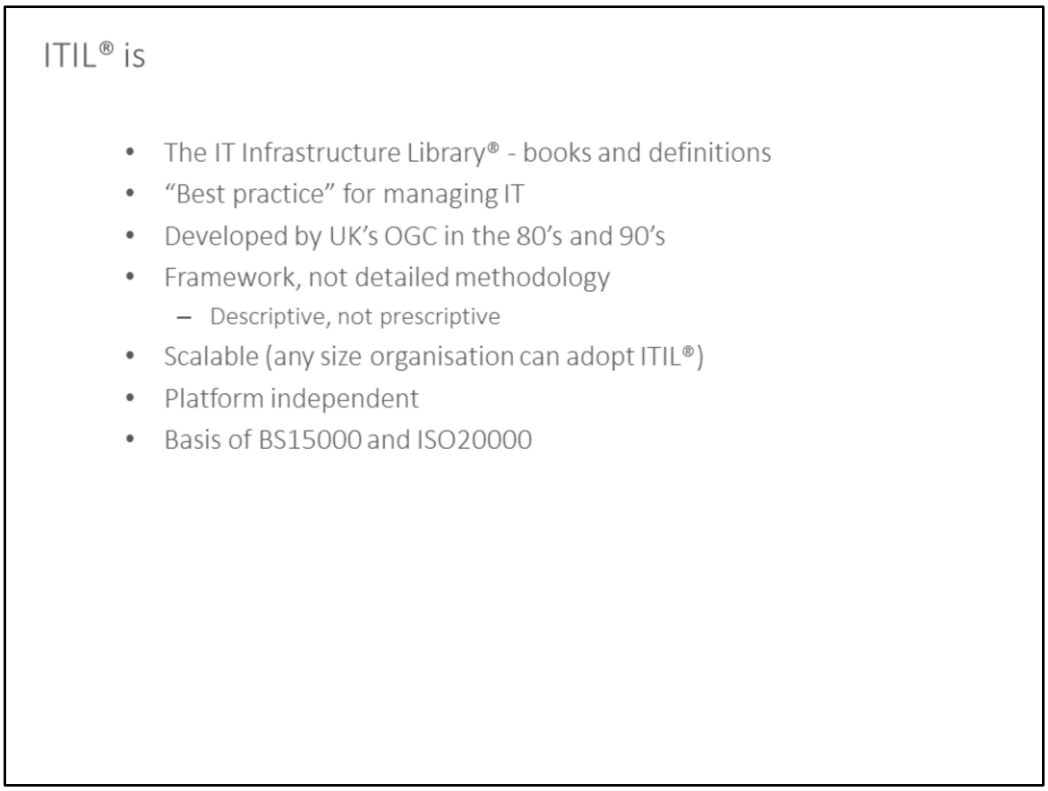
ITIL Capacity Management for the Newbie

Jamie Baker – Product Manager

Jamie.Baker@metron-athene.com

Agenda

- Define ITIL® Capacity Management
- Difference between Capacity Management and Performance Management?
- How to be successful in Capacity Management
- Talking with Business stakeholders
- Creating a baseline and it's effect on forecasting?
- Different types of forecasts?
- Examples of Capacity Management scenarios for Mainframe



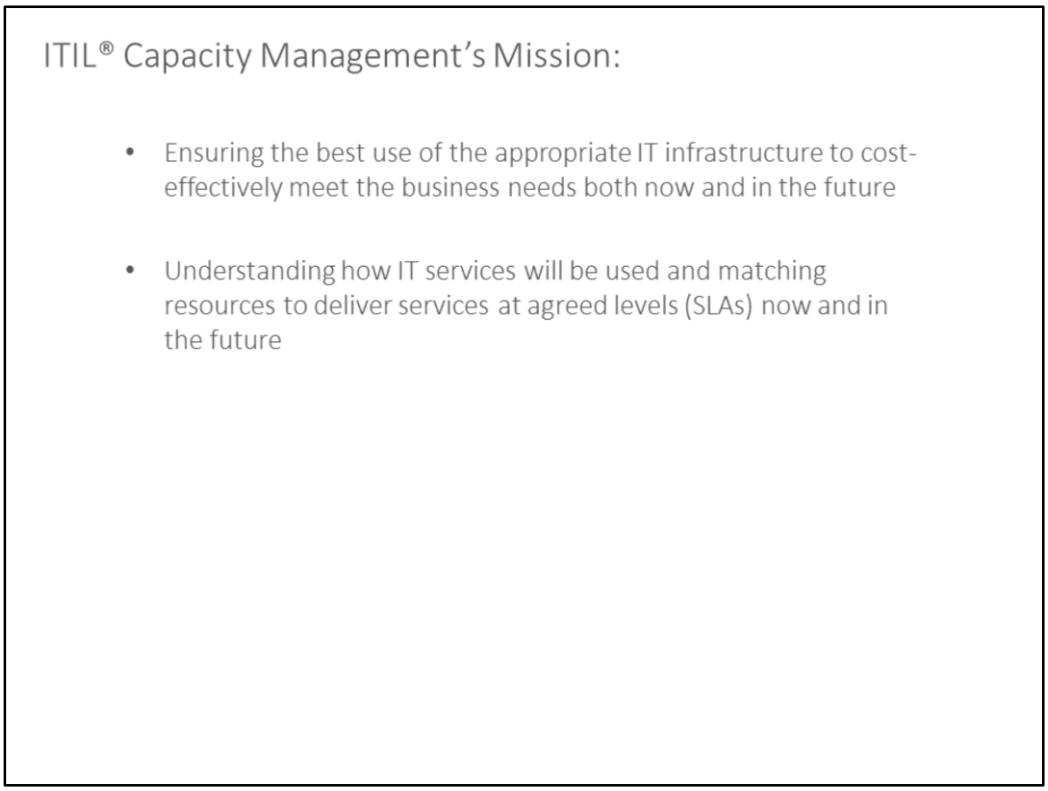
ITIL IS….

ITIL Worldwide

ITIL users are worldwide, and meet under the auspices of the IT Service Management Forum (ITSMF). There are chapters in the UK, Netherlands, Belgium, Australia, South Africa, Canada, France, Switzerland, Austria, Germany and the USA, with new chapters being added all the time.

ITIL in the organization

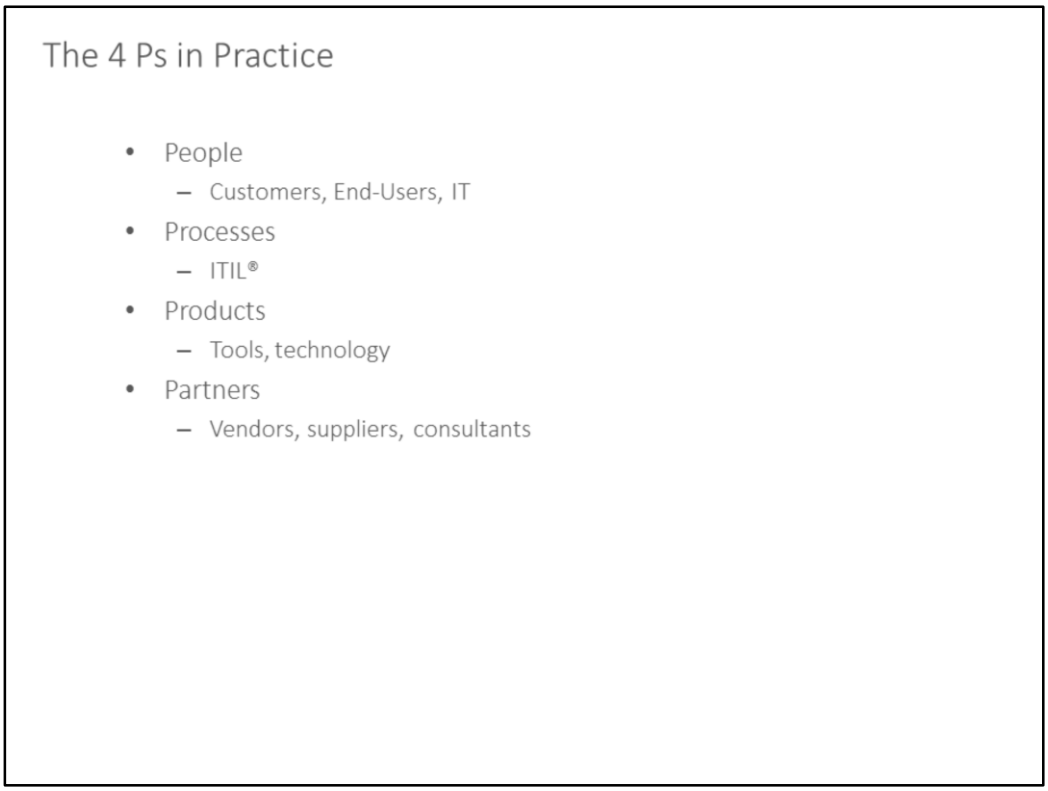
The ITIL process model defines the activities within a typical organization. The major importance in such definitions lies in the identification of the boundaries and resultant necessary interfaces and data flows. The major benefit lies in the standardization of terminology and definitions of the terms used. ITIL can be viewed as providing a common filing system for those involved in ITSM. It does not define what goes into each file or how to get it there.



ITIL Capacity Management Mission

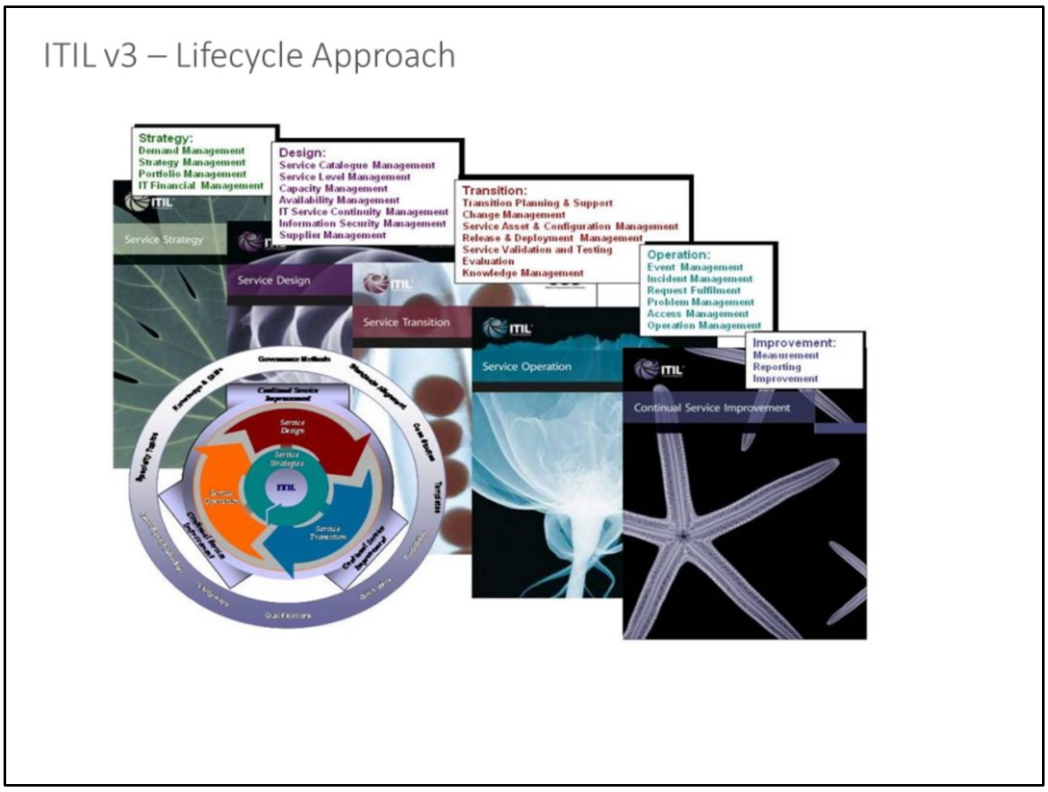
The goal of Capacity Management as stated by ITIL

"To understand the future business requirements (the required service delivery), the organization's operation (the current service delivery), the IT infrastructure (the means of service delivery), and ensure that all current and future capacity and performance aspects of the business requirements are provided cost effectively".



The 4Ps in practice

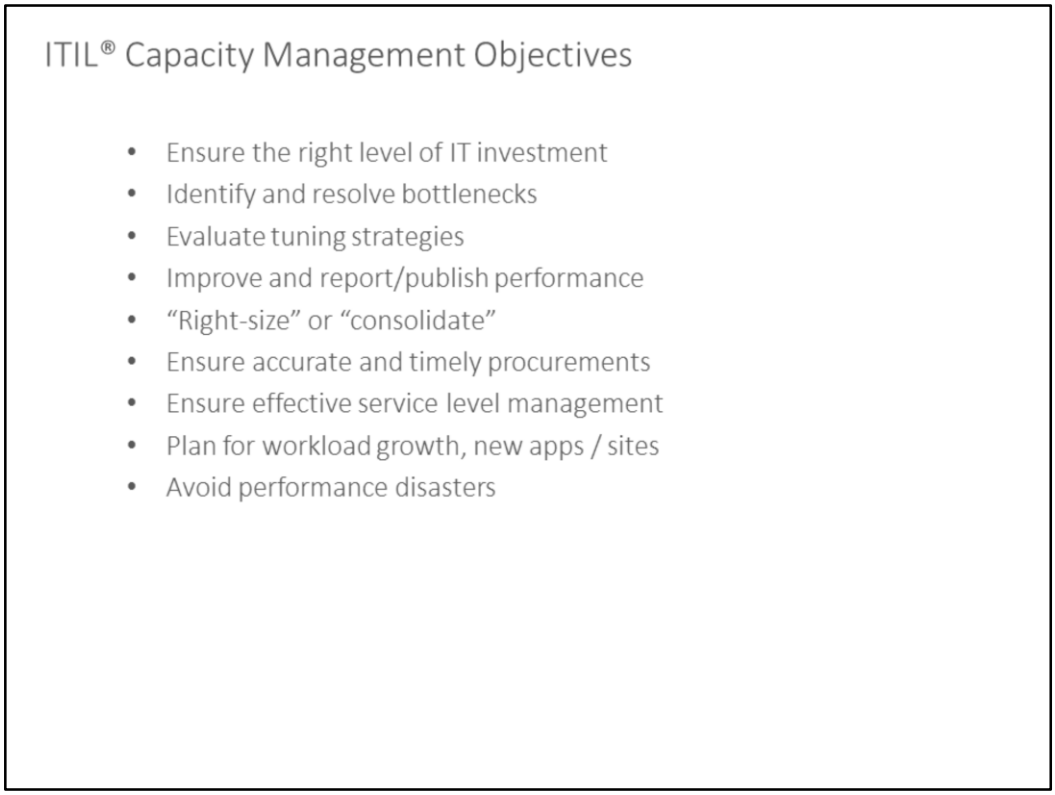
ITIL projects and implementations frequently rely on the interaction of the 4 Ps: People, Processes, Products, and Partners.



ITIL v3 Lifecycle Approach

Dominant pattern in lifecycle is sequential (SS through SD-ST-SO and back to SS via CSI)

Those responsible for design, development, and improvement of processes can adopt (and better benefit from) a process-based control perspective



ITIL Capacity Management Objectives

Essential Objective

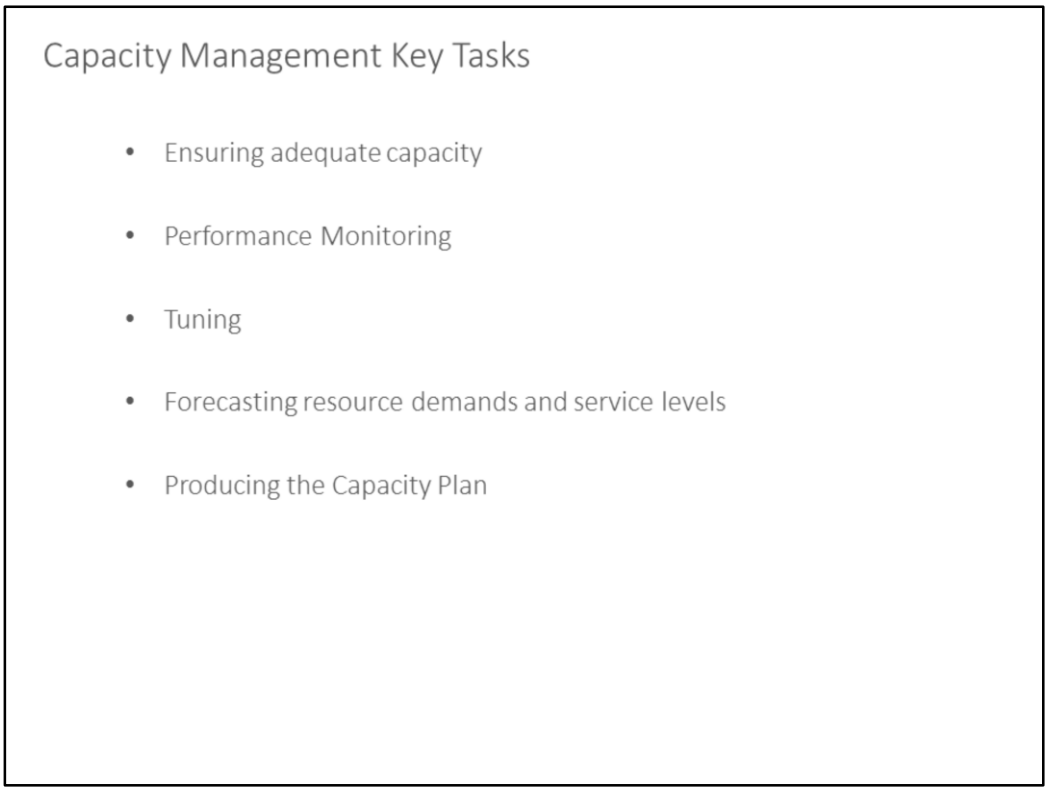
The essential objective is to achieve **the most cost-effective balance** between **business demands** and **the size and form of the IT Infrastructure needed to support it.**

Business Objectives

- The Capacity Management process, including the liaison with Performance Engineering, Service Level management, development, QA etc. has the following business objectives:
- **Ensure the right level of ITI investment**:
  Match the equipment to the need; Optimize on computer expenditure; Do not waste money on redundant hardware; Ensure users are able to meet business demands
- **Optimize the resources available**, "right-sizing" or "consolidating servers" as necessary
- **Ensure accurate and timely capacity procurements** to minimize disruption and expenditure:
- Have reliable hardware plans; Properly size the impact of upgrades
- **Ensure effective service level management** in terms of response times and throughputs
- **Help prepare for new application implementations** or new sites or new acquisitions

Potential More Detailed Objectives

Configurations should be matched to workloads
Impact of upgrades are properly sized and procured in a timely fashion
Service Levels are maintained
The impact of workload growth can be predicted
Potential bottlenecks are identified and bypassed
Tuning strategies are evaluated.



Capacity Management Key Tasks

Ensuring Adequate Capacity

It is the responsibility of Capacity Management to ensure that adequate capacity is available to meet the needs of the business, as the business needs change and evolve, in a cost-effective and timely manner. Effective Capacity Management is based on **business requirements**, not just on the current hardware and software performance (or lack of it). For example, if the business need is being met on a server which "everyone knows" would benefit from being tuned, such effort would be better directed elsewhere.

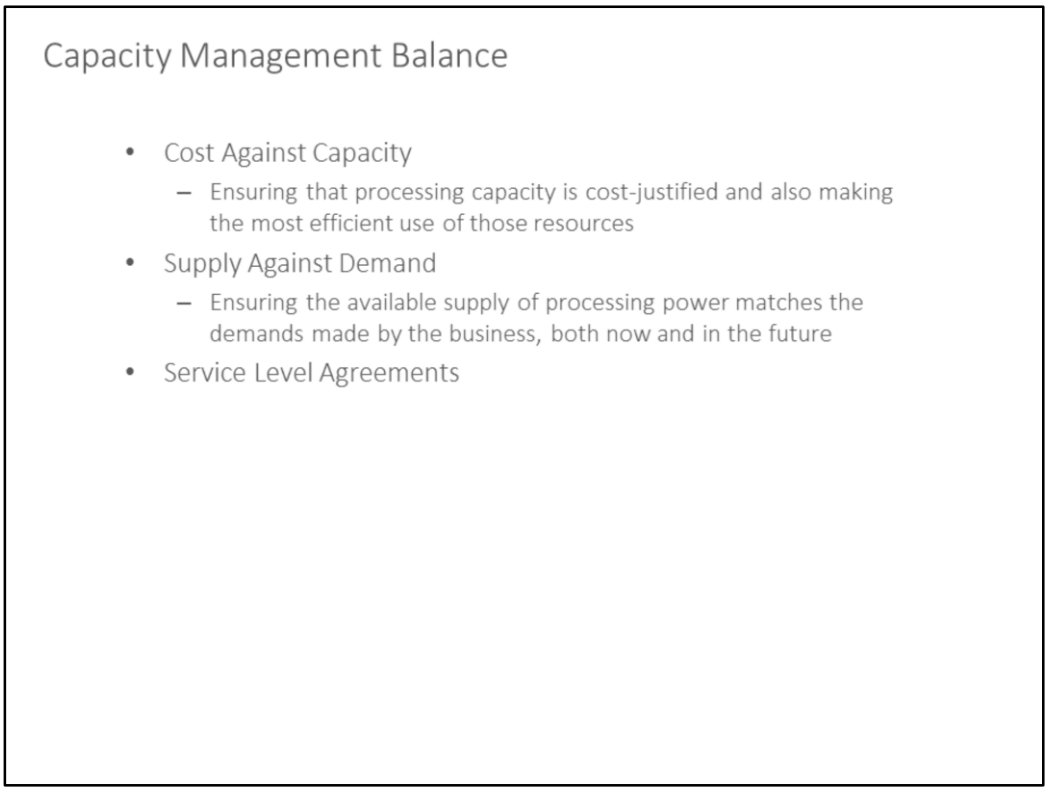
Incidents

In ITIL® terminology, an Incident is an event that interferes with the normal operation of the services supporting the business. It is entirely possible for an incident to be raised (through the Help desk) which, when analyzed, appears to be caused by a performance problem related to lack of capacity in some particular area. In this case, it is the responsibility of Capacity Management to assist in the resolution of the problem and the subsequent closing of the incident. Since this may involve unbudgeted expenditure, this would be regarded as a failure, or at least a shortcoming, in the Capacity Management process. Capacity Management should attempt to ensure that capacity-related incidents are extremely rare. This, after all, is the purpose of forward planning.

Forecasting Demand and Predicting Service Levels

This is the central task around which all the others revolve. To carry out this activity successfully, the Capacity Management team needs access to:

- Performance data
- Technology forecasts
- Business requirements, forecasts and statistics
- Modeling tools and techniques
- Budgetary information.



Capacity Management Balance

The art of Capacity Management lies in finding the optimal balance for two related scales:

Cost versus capacity

Supply versus demand.

The former ensures that the processing capacity is cost justified and optimized. The latter ensures that the power matches the demands made to meet the business need. Capacity Management aims to understand what an organization needs to achieve in the future, map this onto IT resources required to achieve those goals and provide a plan showing what IT resources are needed and when in order to achieve those goals. Without this forward-looking activity, you could be in for any number of unpleasant surprises, such as:

Performance crises

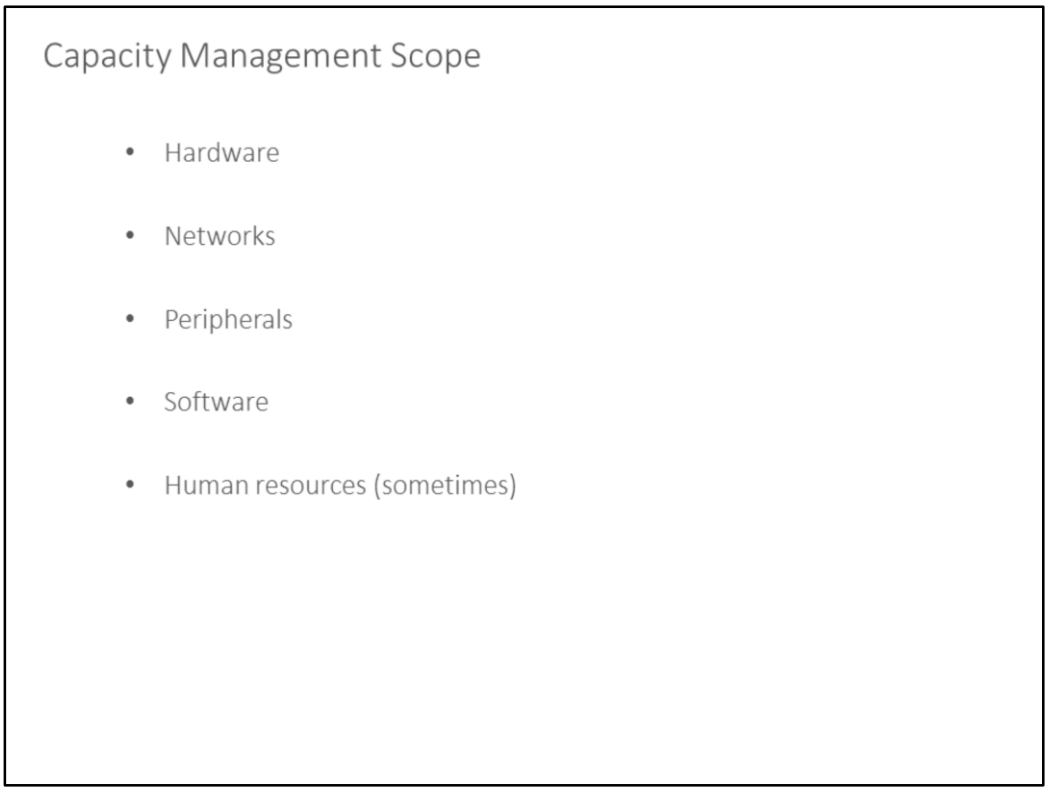
Unnecessary hardware expenditure

User dissatisfaction.

Capacity Management is responsible for ensuring adequate capacity is available at all times to meet the requirements of the business. It is directly related to the business requirements and is not simply about the performance of the system's components, individually or collectively.

Service Level Agreements

Ideally, these business requirements are quantified in formal Service Level Agreements. These should not only define the availability and continuity requirements for a given application service, but should also define the application transaction traffic levels and corresponding throughput and response time performance levels required.



Capacity Management Scope

Hardware - This is the traditional focus of Capacity Management, not least because hardware planning can be supported (and simplified) by the use of modeling tools and other related technologies.

Mainframes and Servers - This is typically the area in which most of the Capacity Management effort is expended, and justifiably so: it is planning decisions about purchase and upgrades of mainframes and large servers that have the most significant financial implications.

Workstations - A large organization may have tens of thousands of employees, at hundreds of sites, each with their own workstation. It is not feasible to carry out any kind of planning process for each workstation individually. Clearly some kind of sampling is necessary, based on job function, location or other attribute.

***Storage (NAS and SAN)* -** A special feature of storage planning is that two quite distinct capacity issues are involved:

- Capacity in terms of storage space available
- Capacity in terms of the ability to support high rates of I/O transfer requests.

Networks - Again, this is an area of planning that can benefit from the use of specially designed tools, to measure network load and predict capacity requirements. Again, though, there are two distinct capacity issues:

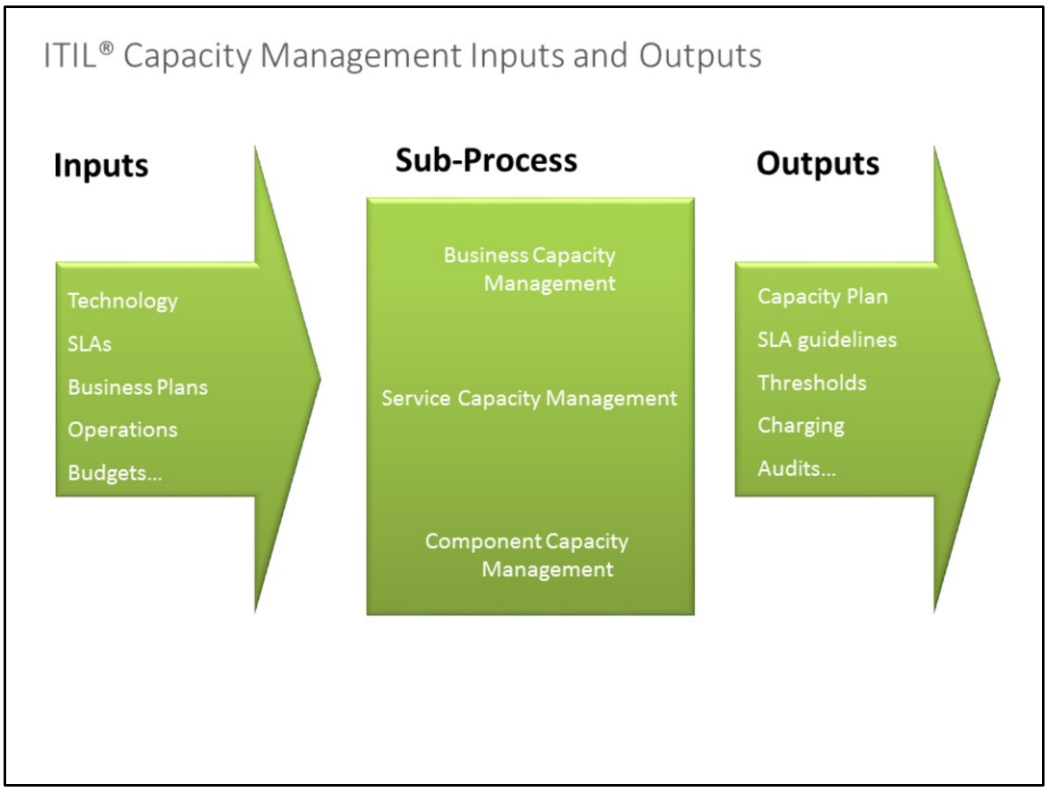
- Network performance (ability to move the required amount of data around at a satisfactory speed)
- Network connectivity (supply of sufficient numbers of physical connections to satisfy the needs of all users).

Software - Capacity Management is responsible for identifying future demands for service as required by the business. Services are provided by applications developed in-house, and/or by packages purchased from third parties; and they run under the control of operating systems such as UNIX and Windows, and sub-systems such as Oracle or SQL Server.

What is a Service

A service is an offering that IT can provide to users and customers.

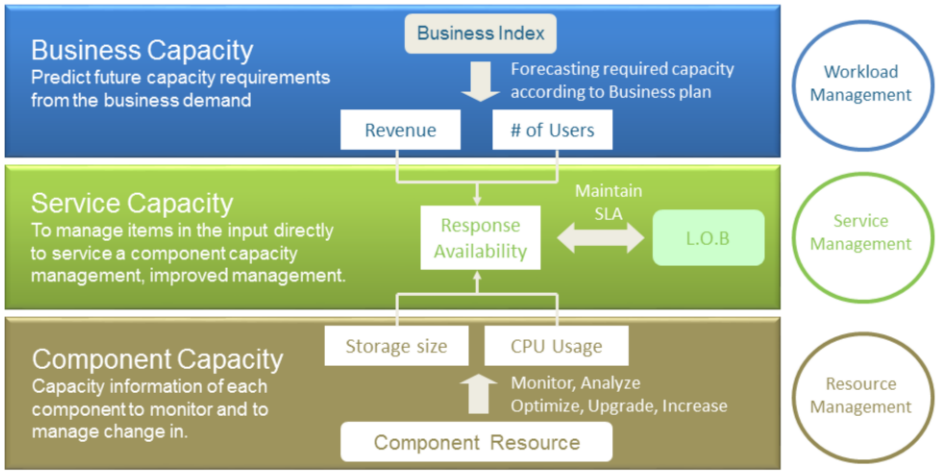
The key part of the definition provided above is that the capability should meet an already-communicated business objective. In other words, business requirements should direct IT to design or purchase services to meet needs, not the other way around (the technology defining the service).



ITIL Capacity Management Input and Outputs

The ITIL® description of Capacity Management indicates the need for a repository for the relevant information in a Capacity Management Database, with the primary deliverable of a capacity plan (and related performance reports) and major activities such as demand management, modeling and application sizing.
The inputs and outputs are indicated in this slide.
These activities have to be actioned at all three levels of Capacity Management: Business, Service and Resource.
Essentially this is achieved by **aggregation** and **simplification**. Thus the huge amount of detailed resource data can be aggregated to reflect service level agreements for individual transactions or applications. These in turn can be aggregated and weighted by appropriate business impact factors to provide the corporate performance management dashboard, or Business level Capacity Management.

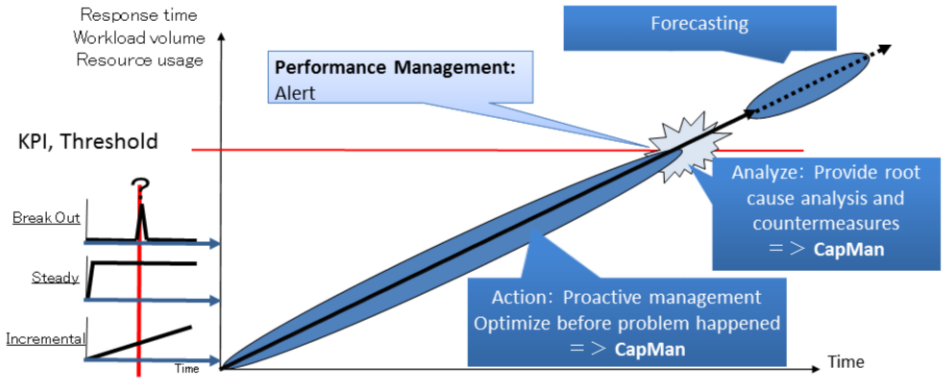
Capacity Management

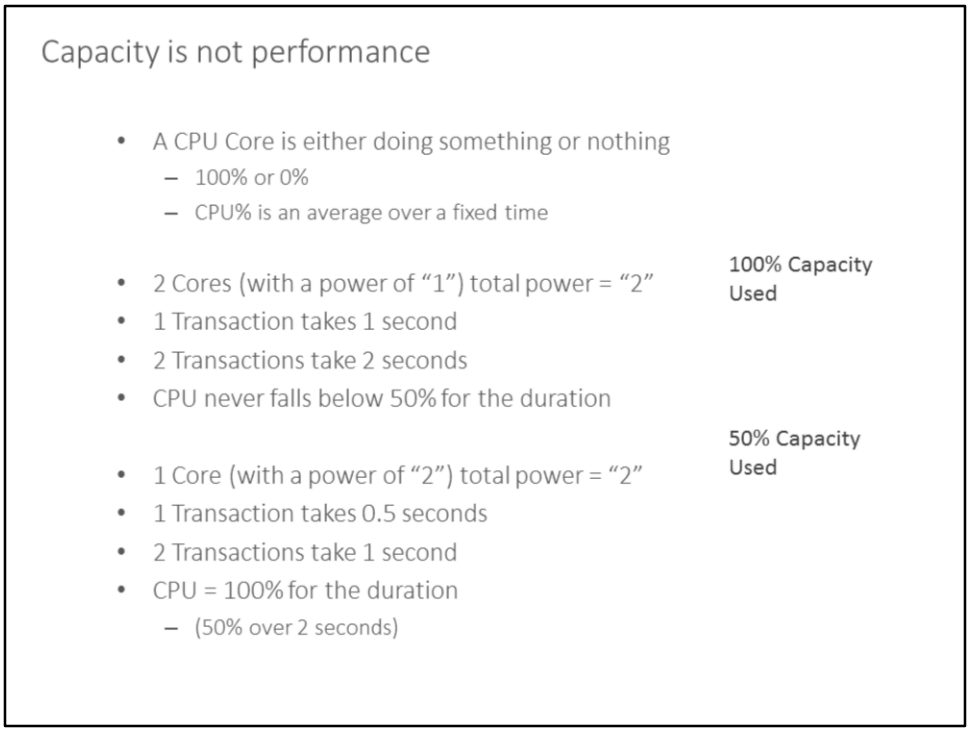


Difference between Capacity Management and Performance Management?

Performance Management vs. Capacity Management

- Performance Management: Alert problems
- Capacity Management: Prevent problems





Capacity doesn't just rely upon performance though.

A CPU at the smallest possible time frame, is either 100% busy, or 0% busy. It either used that clock tick to do something useful, or it didn't. The CPU Busy % that we all know and love is an average over a time period.

We're going to look at two scenarios, each processing 1 transaction and then 2 transactions presented at the same time.

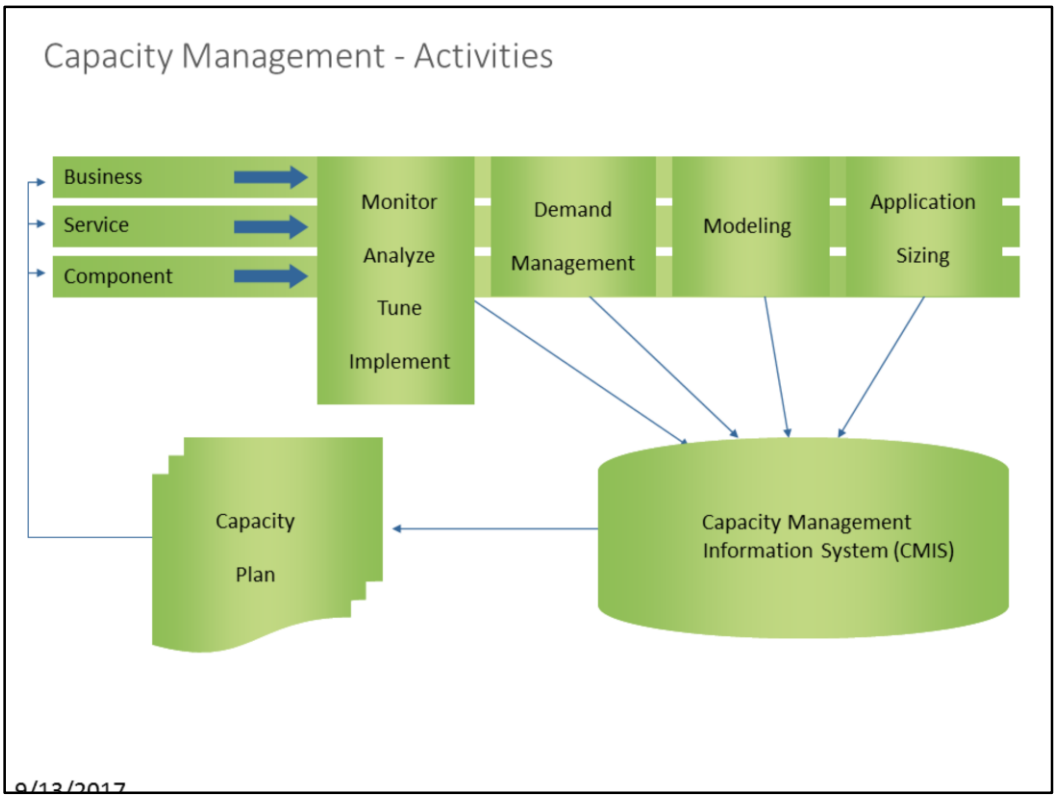
1st Scenario
We have 2 cores each with a "power" of 1. So the total power (resource) is 2.
One transaction takes 1 second, 2 transactions (presented at the same time), take 2 seconds.
And the CPU utilisation during those 2 seconds was 50%.

2nd Scenario
We have the same power. 2. But in a single core.
This time the first transaction completes in half a second, a second transaction presented at the same time completes after 1 second.
During the processing, CPU was at 100%.

If we were just looking at CPU% during the processing we'd think the 2nd scenario was out of capacity and service times were being affected. But if we look at it over the time period of the first scenario, then we can see we are at 100% capacity in the 1st scenario and 50% capacity in the 2nd.

This is all due to the nature of the workload being single threaded.

How to be successful in Capacity Management



Capacity Management Activities

Monitoring, Analysis, Tuning, Implementation
The main iterative activities are essentially part of Performance Management . These are key to all sites, especially as the number of servers increases.
Included in this area by implication (but not explicitly) are:
Event Management, Alarms and alerts
Intranet status and exception reporting
In ITIL® terms, the tuning and implementation activities incorporate a necessary level of reporting. Depending on the application and organization, it may be regular reports or exception reports, to management, users or technicians, via the intranet, email or paper.

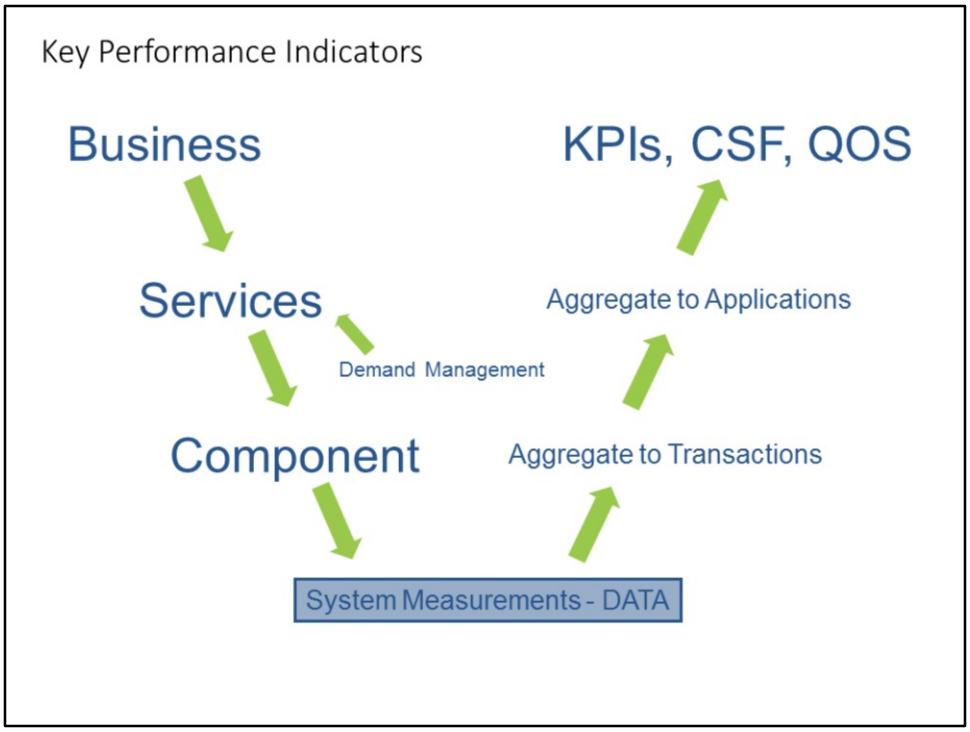
Demand Management
Demand Management is the control of resources to meet specific levels of demand that the business is willing to support. For example, user demand might have to be limited for a period if additional capacity cannot be purchased immediately. Demand Management is underpinned by modeling, which can show what level of demand can be supported for a given level of resources. This is essential in disaster planning, showing what demand can be supported if a given component in the infrastructure fails.

Modeling
Modeling and prediction is at the heart of proactive Capacity Management. It enables you to know what service can be provided for a planned workload or what workload can be supported on a given set of resources.

Application Sizing
A major risk with implementing a new application or making significant changes to existing services is that they no longer perform to required service level targets. Many development teams use load testing and benchmarking techniques to allay some of the fears.

Key Performance Indicator

- Measurements, agreed to beforehand, that reflect the critical success factors of an organization
- Not necessarily one metric, can be a relationship
- Key Performance Indicators Must Be Quantifiable
- Reflected in a scorecard



Key Performance Indicators (KPI) are critical in the measurement of performance goals within an environment. To ensure a KPI is meaningful, it must have a discrete way of measurement. You want to ensure the number presented to be ambiguous and eliminate any gray area for interpretation.

KPI's can consist of single or multiple components. The KPI components can consist of technical, service and business metrics. The critical factor is to present what is meaningful to your audience.

Quality of Service (QOS) is tied to how the performance of the environment relates to the end-user or workload.
The component that allows for the display of this information a centralized Capacity Management Information Store (CMIS). A centralized repository allows for the collection of business, service and technical metrics in one database and report on a single or multiple components as necessary.

Talking with Business stakeholders

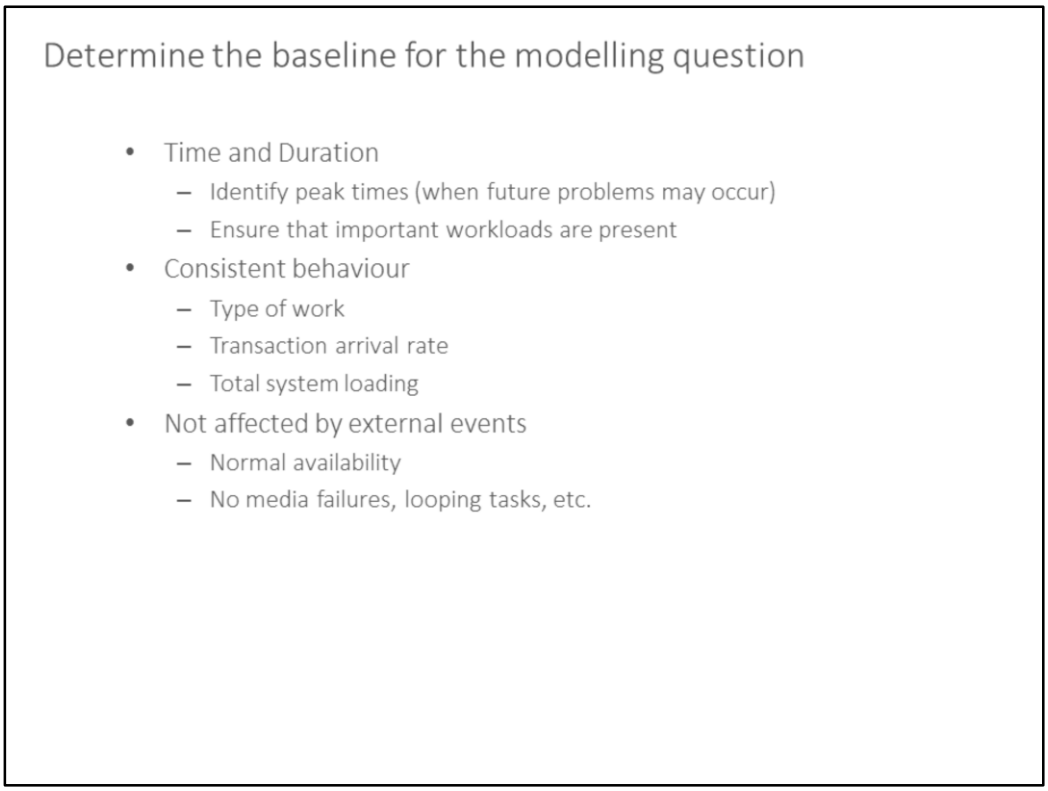
What is the Story?

- What is happening in the current environment
- Concise information
- Display forecasts
  - Trends
  - Models
- Gather further information

Different Capacity Plans for different audiences

- "C" level executives
  - Information concise
  - Elevator talk
  - Information to the point / summary and findings first
  - Do I need to spend more money?
  - Leave detail reports in pocket
- Business owners
  - What are the trends for my area?
  - How does it affect my area ?
  - What do I need to budget for?
- Technical
  - Show me the details
  - Show me the trends

Creating a baseline and it’s effect on forecasting?



Choosing a modeling period

When deciding on a modeling period, consider all the points shown above. In particular, always be aware what you are building a model **for** (in other words, what specific aspects of present and/or future system activity you are investigating).

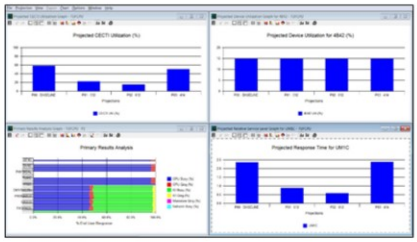
Furthermore, ensure that your chosen time period is a period of "normal" (ideally, busy) activity for which consistent and complete data is available. There is no point choosing a period during which the system behaved abnormally, because that is not a sensible basis for planning.

Different types of forecasts?

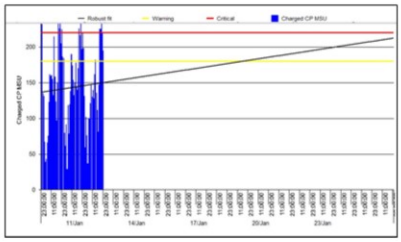
Forecasting

Quality of Service to Stakeholder

Modeling



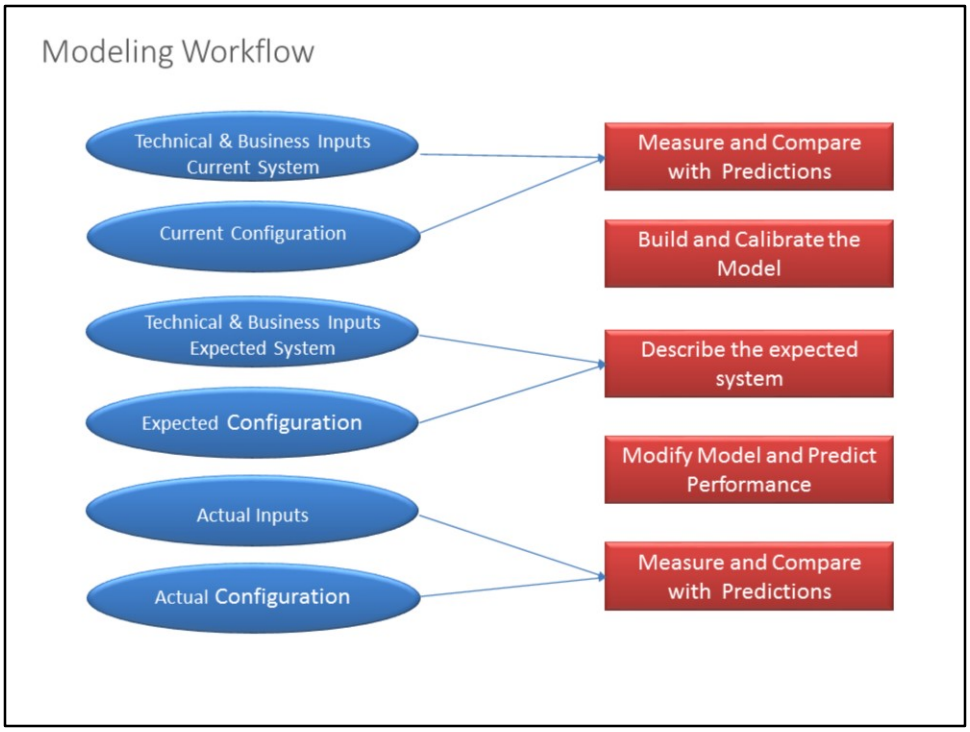
Trend



Modeling vs. Trending

- Trending
  - Easy to do
  - Scales well for multiple systems e.g. Windows , UNIX
  - Easy to alert on
  - No relationship to the performance (service) of a system or application
  - Looks at just one metric

- Modeling
  - Relates service to utilization
  - View whole system interactions
  - ITIL recommends modeling
  - Moderate effort

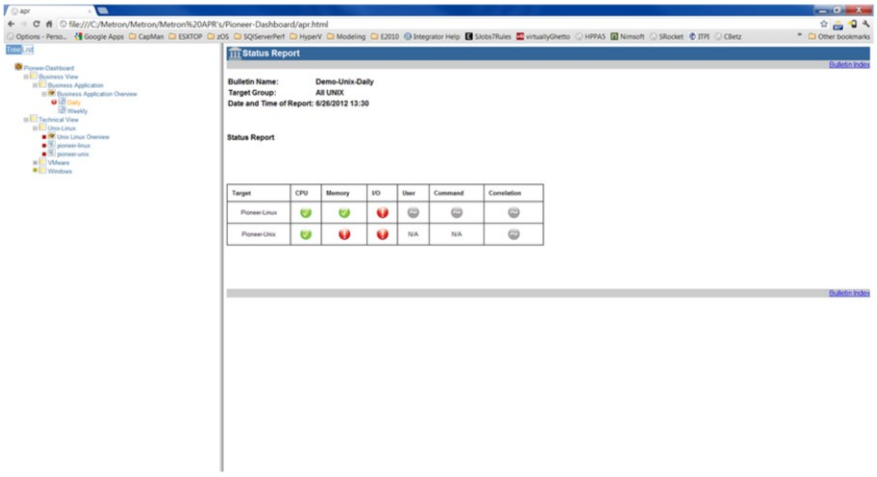


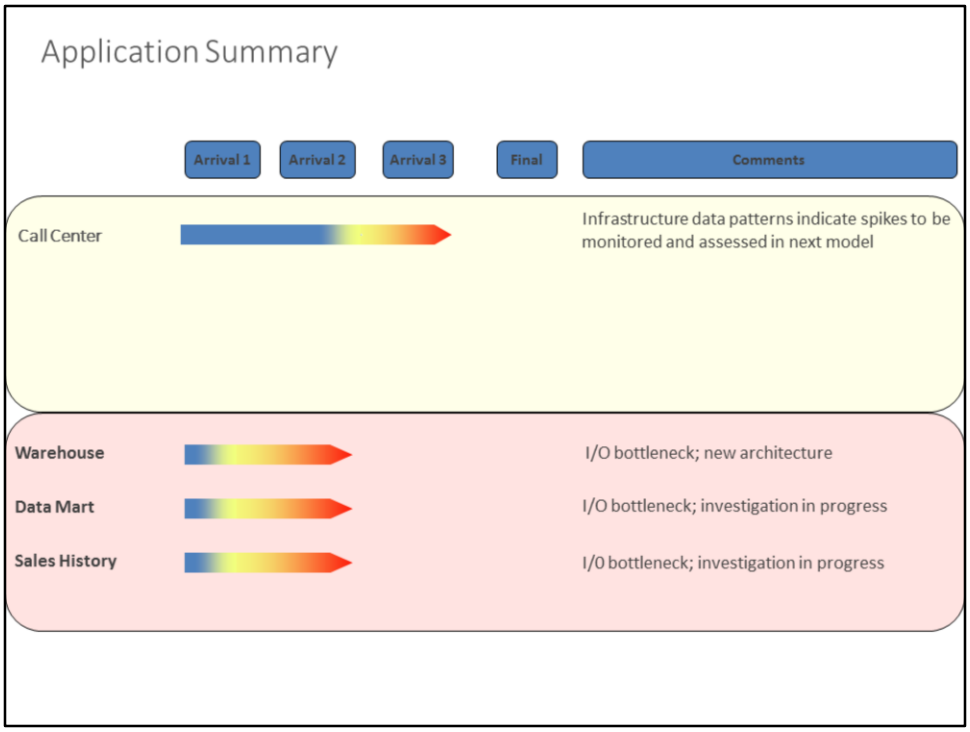
There are six major steps in the Analytical Modeling process as described above. Each one is critical in it's own right along with in total of the process. Following this flow will assist in ensuring not only you have the right information but understand the outcome of the model. With every model there are assumptions that have to be built in. Documenting those assumptions allow you to understand any discrepancy when the actual does not match the projection. This feeds into an appropriate baseline for your model as we will discuss in more detail in subsequent slides.

Agenda

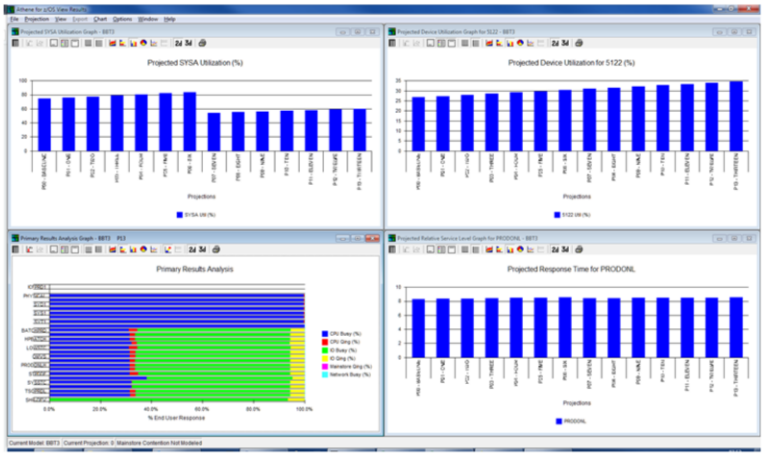
Examples of Capacity Management scenarios for Mainframe

Dashboard – Overview Scorecard Detail




Application Summary
Arrival 1
Arrival 2
Arrival 3
Final
Comments
Call Center
Infrastructure data patterns indicate spikes to be monitored and assessed in next model
Warehouse
I/O bottleneck; new architecture
Data Mart
I/O bottleneck; investigation in progress
Sales History
I/0 bottleneck; investigation in progress

Results CPU Change CPU/CEC Modeling – Hardware & I/O



Summary

- Capacity Management is a Business activity
- Capacity Managers are the liason between the business and technical areas
- Gather all the business and technical data possible
- Ask questions
- Continually document assumptions

ITIL Capacity Management for the Newbie

Jamie Baker – Product Manager

Jamie.Baker@metron-athene.com