A Framework for Enterprise IT Capacity Management

October 2013

Bhargav.Lavu@tcs.com
Agenda

- Background
- The Framework
- Examples of Success
- Conclusion
- Questions
A fresh Service led approach for managing the end to end capacity of key products with the following objectives was initiated –

- Improve infrastructure management

- Streamline and standardize processes in line with ITIL standards

- Better capacity utilisation

- Improve customer experience

- Drive operational efficiency through cost avoidance
The Framework - Introduction

• Simply manages demand and supply via,
  – Service capacity management
  – Component capacity management

• Services drive demand – so Service Capacity management belongs to the demand side

• Components supply capacity – so Component capacity management belongs to the supply side

Also,

• Service capacity management - Top down business centric view.

• Component capacity management - Bottom up application or infrastructure (IT) centric view

Both are needed for effective management of IT capacity
Service Capacity Management

- Demand Forecasting
  - Service Lines
  - Business Use information

- Capacity Incident Reduction
  - Standard regular checks
  - Impact of change on capacity

- Service Performance Analysis
Demand Forecasting

Forecasting demand is one of the most important and difficult steps in enterprise capacity planning.

Main approaches to forecasting demand are –

- **Business driven** - new product or feature launches, marketing campaigns, customer migrations for business or technology change reasons, changes to process flows etc. *Suited to systems experiencing rapid change*

- **Trending** - use of actual historical data to predict the future demand. *Suited to systems with stable or declining workloads.*

- **Seasonal analysis** – *use of historical information to isolate seasonal trends such as winter peaks in IPTV demand, reduced fault rates during dry summer months, etc.* *Suited to systems with highly periodic workloads.*
### Service Lines

- Draw inspiration from volumetric models.
- Simply measure the visit count for various systems in the enterprise.
- The visit count represents demand for the system in the business goal being modelled.

<table>
<thead>
<tr>
<th>Product Goal</th>
<th>Sub Goal</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Reset</td>
<td>Login</td>
<td>Get Boundary Keys</td>
</tr>
<tr>
<td>User Reset</td>
<td>Login</td>
<td>Refresh Session Key</td>
</tr>
<tr>
<td>User Reset</td>
<td>Login</td>
<td>Login</td>
</tr>
<tr>
<td>User Reset</td>
<td>Login</td>
<td>PresentTicket</td>
</tr>
<tr>
<td>User Reset</td>
<td>Login</td>
<td>Passive Update Guide Data</td>
</tr>
<tr>
<td>User Reset</td>
<td>Login</td>
<td>Client Logging</td>
</tr>
<tr>
<td>User Search</td>
<td>Search</td>
<td>All / By Title / By Name</td>
</tr>
<tr>
<td>User EPG</td>
<td>Scroll</td>
<td>Scroll browse bar</td>
</tr>
<tr>
<td>User DVR</td>
<td>SetSingle</td>
<td>Set single recording</td>
</tr>
<tr>
<td>User DVR</td>
<td>SetSeries</td>
<td>Set series recording using Record</td>
</tr>
<tr>
<td>User DVR</td>
<td>Clear</td>
<td>Clear recording</td>
</tr>
<tr>
<td>User DVR</td>
<td>Record</td>
<td>Record scheduled program</td>
</tr>
<tr>
<td>User DVR</td>
<td>Play</td>
<td>Play recorded show</td>
</tr>
<tr>
<td>User DVR</td>
<td>Frame</td>
<td>Frame recorded program</td>
</tr>
<tr>
<td>User VODInfo</td>
<td>GetInfo</td>
<td>View Menu (catalog)</td>
</tr>
<tr>
<td>User VODInfo</td>
<td>Select</td>
<td>Select Category: request</td>
</tr>
<tr>
<td>User VODInfo</td>
<td>Select</td>
<td>Select Sub Category: request</td>
</tr>
<tr>
<td>User VODInfo</td>
<td>SetAsset</td>
<td>Select Asset: request</td>
</tr>
<tr>
<td>User VODplay</td>
<td>Purchase and Play: Intent</td>
<td>100</td>
</tr>
<tr>
<td>User VODplay</td>
<td>Accept</td>
<td>Accept</td>
</tr>
<tr>
<td>User VODplay</td>
<td>Play</td>
<td>Play VOD</td>
</tr>
<tr>
<td>User VODplay</td>
<td>Pause &amp; play</td>
<td>100</td>
</tr>
</tbody>
</table>
## Capacity Incident Reduction

### Stage 1: Service Level Incident Analysis

1. Categorize the capacity incidents based on cause.
2. Short list the top IT systems associated with 80% of incidents (Pareto analysis.)

### Stage 2: Implement Recommendations

3. Analyse the causes and the IT systems to come up with recommendations to address issues.
4. Raise Service Records to implement the recommendations.

### Stage 3: Produce Service Level Report

5. Publish Service Level report on identified top IT systems (as in 2.) post implementation.

### Stage 4: LOB level Incident Report

6. Consolidate the incident report for all Services and publish a monthly report.

### For Disk related capacity incidents, urge application support team to…

- Implement automated housekeeping for log and data files in case not present
- Ensure thresholds set appropriately
- Ensure 100% coverage of standard monitors
- CM Devised tool to forecast exhaustion

### For other capacity incidents

- Documented procedures for application support team to handle CPU and memory resource utilisation, run-away processes, virtual memory configuration etc
- Problem records being for repeated issues (e.g. common looping tasks)
Service Performance KPI’s

1. Define KPIs - what **should** be measured

2. Define what **can** be measured

3. Gather the data
   - Who?
   - How?
   - When?
   - Integrity of data?

4. Process the data
   - Frequency?
   - Format?
   - System?
   - Accuracy?

5. Analyse the data
   - Relations?
   - Trends?
   - Targets met?
   - Corrective action?

6. Report and use the information, assessment summary, action plans, etc.

7. Implement corrective action

Identify
- Vision
- Strategy
- Goals
Service Performance KPIs – Analysis of Data

• Response time was good when volumes were high while response time was high when volumes were low (see left of chart)

• This defies capacity utilisation laws.

• Indicated presence of a background load that was affecting response times which was investigated and moved.
Component Capacity Management

For each IT System/Application

- Business volumes, application workload; Throughput and response time targets (SLA/OLA); Expected growth
- Capacity Utilization, transaction throughput, end to end response time
- Availability Requirements
- Existing Capacity Plan / Estimates; Actual measurements (results of performance tests)
- Capacity Incident Analysis

Gap Analysis

- Gaps in Performance monitoring and Capacity Planning
- Gaps in existing capacity planning approach followed by the Application team; Gaps in existing performance test results

Short Term Recommendations

- Quick fixes for improving application performance and stability
- Capacity Provisioning for the shorter term if needed
- Setting up the right toolset for performance measurement and capacity planning

Long Term Recommendations

- Design Recommendations for better performance and optimal resource utilization (For ex: Tuning of Billing platform)
- Defining the Iterative process for Release Capacity Analysis
- Refining the existing capacity planning model i.e. demand model based on the future workload
Capacity managers are often asked some key questions:

• What is the total installed capacity?

• How much are we using the existing capacity?

• When will we exhaust this available capacity?
Reporting – Component capacity exhaustion date

- Capacity forecast chart for peak daily transactions with available headroom.

- When forecast is running close to Safe working limit, actions are taken to increase headroom.

- Actions include potential optimisation or upgrades.
Governance

Looks at approaches and reports intended to help the business understand, and so guide, along the scope and depth of capacity management.

- **Scope**: Component – Service matrix
- **Depth**: Capacity Management Index
- **Capacity Risk management**
- **Quality**: Capacity Management KPIs
Scope: Component Capacity Matrix

<table>
<thead>
<tr>
<th>IT System Name</th>
<th>Service 1</th>
<th>Service 2</th>
<th>Service 3</th>
<th>Service 4</th>
<th>Service 5</th>
<th>Service 6</th>
<th>Service 7</th>
<th>Service 8</th>
<th>Service 9</th>
<th>Service 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>System 1</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 2</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>System 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>System 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>System 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>System 11</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>System 16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>System 17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>System 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>System 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>System 20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

- SOA encourages re-use of capabilities
- Service can be delivered over a large number of systems
- Aids discussion with technical and non-technical stakeholders
- Helps illustrate which are the most essential systems (capabilities) to the enterprise
- Aids impact assessment of new services built on SOA
Depth: Capacity Management Index

<table>
<thead>
<tr>
<th>CMI 1</th>
<th>No centralised capacity management</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMI 2</td>
<td>Support led capacity management</td>
</tr>
<tr>
<td>CMI 3</td>
<td>Cap Man led capacity management, trending, system (CPU) monitoring</td>
</tr>
<tr>
<td>CMI 4</td>
<td>CMI 3 + Demand Forecasting + Service Monitoring</td>
</tr>
</tbody>
</table>

- Few capacity managers, hundreds of IT systems
- Risk based approach
- Improved engagement as priorities are driven by stakeholders
Examples of Success

Some success stories are presented here to provide a flavour of the value created by the capacity management function -

• Capacity Incident reduction

• Proactive monitoring

• Sensitivity analysis

• CAPEX savings
Capacity Incident Reduction

- Analysed for systems with most problems
- Enabled by standardised checks targeted at recurring problems
- 99% reduction
Proactive Monitoring

- Post release checks OK as code not exercised
- Visual inspection
- Investigation and manual intervention managed prior to fix
- Process automated
Sensitivity Analysis

- Is investment needed this FY?
- What would happen if trading conditions changed?
- Demand forecasting enabled scenarios to be tested
- Enabled business decision to be made
Conclusion

• A framework for capacity management was devised and implemented for enterprise wide capacity management.

• A best of breed framework for capacity management which has proven to be effective.

• Has application in many large enterprises.
Questions