Starting as a team of one

Starting a capacity management function has a beginning, and that can be starting from scratch with one person. Determining where to start with building a capacity management capability requires evaluating what is possible in the near and long term, defining what needs to be done and how to do it requires the evaluation of many factors. Then defining goals and objectives to accomplish the mission, short and long term goals are established as more is known.

Defining the 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”.

So the Capacity Management Mission must:

• Ensure the best use of the appropriate IT infrastructure to cost-effectively meet the business needs both now and in the future
• Understand how IT services will be used and match resources to deliver services at agreed levels (SLAs) both now and in the future

Initial roadmap for implementing Capacity Management

The development of a capacity management capability should be validated with IT management and combined with initiatives to improve the IT Service Management environment and capability.

In the context of the ITSM initiative, the implementation points outlined below were deemed critical to the success of a creating a Capacity Planning capability.

• Establish the IT Capacity Management team
• Ensure monitoring is validated and deployed for identified IT resources
• Analyze/trend existing resource data and establish capacity reporting thresholds
• Plan, design, build and implement the CDB
• Implement capability to integrate CMDB with the CDB
• Understand/measure/report on the quality of service delivery
• Create Capacity Plan document containing history, predictions, and recommendations
• Make recommendations to ensure service is delivered as required
• Refine the process to translate business plans and forecasts into IT requirements

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 performance tuned then such effort would be better directed elsewhere.

Initial Capacity Management tasks were implemented one service at a time

• Review process documents and best practices
• Capture performance data
• Establish service level and business volume metrics
• Build relationship with operations and application area
• Start producing capacity plan
• Configure alerting
• Ensure adequate capacity across monitored components
• Forecast resource demands and service levels

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.

ITIL Capacity Management Objectives
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.

Long Term Capacity Management Objectives

- Ensure the right level of IT infrastructure investment
- Identify and resolve bottlenecks
- Evaluate tuning strategies
- Improve and report/publish performance
- “Right-size” or “consolidate”
- Ensure accurate and timely procurements
- Ensure effective service level management
- Plan for workload growth, new apps / sites
- Avoid performance disasters
- Performance Monitoring

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 IT 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.

**Spreadsheet analysis**

Spreadsheets are a great way to get up and running when other tools are not available, they can be used for many tasks related to analyzing and reporting on capacity.

The example below shows a worksheet that is part of a spreadsheet used to forecast CPU and memory capacity for a VMware cluster.

![Spreadsheet Example](image)

The problem with using spreadsheets as a complete solution is the maintenance becomes unwieldy as more components and different data types are managed. Software specifically designed to address capacity management address this scalability problem with automation and scheduling.
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.

Network data was the first area of focus for our client because data was available, where spreadsheets were used to create value quickly.

An example worksheet showing how network capacity was evaluated, produced in the form of heat map type report, is shown below and was created with conditional spreadsheet cells.
Below is an example of a heat map type report but created automatically with our athene® capacity management tool.
Spreadsheets are a powerful tool for analyzing data, but they become a maintenance burden as your capacity management operations scale up. In the example below you can see that limitations occur when too much of too many types of data must be analysed and reported on repeatedly.

Investing in capacity management software has an upfront cost that is quickly recovered through savings across many areas. The example below shows positive returns only six months after implementation.
Factors to include when doing your own ROI might be:

**Costs**
- License Cost
- Maintenance
- Training
- Implementation Services
- Personnel
- Servers

**Benefits**
- Hardware Purchase Deferral
- Software Savings from Consolidation
- Less Downtime
- Better Performance (productive and happy users)
- CM staff productivity
- Sys Admin Savings

**Capacity management tool selection**

**Evaluation Criteria and Results**
Products were evaluated by our client for their suitability as a CMIS in the following categories:

- Data collection
- Data Management
- Reporting
- Dashboards
- Delivery
- Analytics
- Modeling

In order to make an objective determination for products, each test was assigned a relative importance score of 1 to 10 and each product was assigned a separate functionality score of 1 to 10. The two scores were multiplied to obtain a weighted score of 1 to 100 for each test. A score of 70 or greater is considered exceptional, a score of 31 to 70 is adequate, and less than 30 is unsuitable.

Each category was evaluated based on aligned functionality in each tool.

The decision to purchase a capacity management tool is a required step to advance your capabilities beyond manual processing but choosing a capacity management tool is not as easy as it sounds.
Selecting a tool involves evaluating factors related to the goals and objectives set out in the beginning. There are various ways to evaluate tools and no one way is best. The person or persons making the selection will bring their own methods and priorities to the process. Criteria may be evaluated in a questionnaire as shown in the example below and don’t forget that experience may also play a big role in deciding what is the right tool.

<table>
<thead>
<tr>
<th></th>
<th>Average Importance Score</th>
<th>Average Functional Score</th>
<th>Average Weighted Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Analytics</strong></td>
<td>7.2</td>
<td>9.6</td>
<td>69.1</td>
</tr>
<tr>
<td>Analyst Summaries</td>
<td>6.8</td>
<td>10.0</td>
<td>67.5</td>
</tr>
<tr>
<td>Interval Reports</td>
<td>6.7</td>
<td>9.3</td>
<td>61.8</td>
</tr>
<tr>
<td>Profile Reports</td>
<td>8.3</td>
<td>9.5</td>
<td>77.9</td>
</tr>
<tr>
<td><strong>Data Collection</strong></td>
<td>8.4</td>
<td>8.3</td>
<td>72.9</td>
</tr>
<tr>
<td>Custom Data</td>
<td>9.5</td>
<td>7.5</td>
<td>71.3</td>
</tr>
<tr>
<td>iSeries</td>
<td>5.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Oracle</td>
<td>7.5</td>
<td>9.1</td>
<td>68.3</td>
</tr>
<tr>
<td>UNIX</td>
<td>10.0</td>
<td>10.0</td>
<td>100.0</td>
</tr>
<tr>
<td>VMware</td>
<td>9.5</td>
<td>9.0</td>
<td>85.5</td>
</tr>
<tr>
<td>Windows</td>
<td>8.0</td>
<td>9.4</td>
<td>76.0</td>
</tr>
<tr>
<td><strong>Modeling</strong></td>
<td>9.0</td>
<td>8.3</td>
<td>74.2</td>
</tr>
<tr>
<td>Baseline</td>
<td>8.5</td>
<td>9.5</td>
<td>80.8</td>
</tr>
<tr>
<td>What-If Modeling</td>
<td>9.1</td>
<td>8.0</td>
<td>72.6</td>
</tr>
<tr>
<td><strong>Reporting</strong></td>
<td>8.3</td>
<td>7.5</td>
<td>63.3</td>
</tr>
<tr>
<td>Automatic Reports</td>
<td>8.7</td>
<td>6.9</td>
<td>62.2</td>
</tr>
<tr>
<td>Bulletins</td>
<td>7.8</td>
<td>9.0</td>
<td>70.4</td>
</tr>
<tr>
<td>Dashboards</td>
<td>7.5</td>
<td>4.1</td>
<td>30.8</td>
</tr>
<tr>
<td>Delivery</td>
<td>8.3</td>
<td>9.8</td>
<td>82.1</td>
</tr>
<tr>
<td><strong>Trend</strong></td>
<td>7.0</td>
<td>9.0</td>
<td>63.0</td>
</tr>
<tr>
<td>Trend Wizard</td>
<td>7.0</td>
<td>9.0</td>
<td>63.0</td>
</tr>
<tr>
<td><strong>Data Management</strong></td>
<td>8.2</td>
<td>8.8</td>
<td>72.8</td>
</tr>
<tr>
<td>Custom Data</td>
<td>8.2</td>
<td>8.8</td>
<td>72.8</td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td>8.2</td>
<td>8.3</td>
<td>69.1</td>
</tr>
</tbody>
</table>

In order to make an objective determination, each test was assigned a relative importance score of 1 to 10 and each product was assigned a separate functionality score of 1 to 10. The two scores were multiplied to obtain a weighted score of 1 to 100 for each test. A score of 70 or greater is considered exceptional, a score of 31 to 70 is adequate, and less than 30 is unsuitable.
Capacity Management Outputs
For analysis the user needs complete control over factors such as time periods for viewing data, aggregation, trends and workload definition. Business data needs to be automatically correlated with Application Views or individual component information to enable analysis across Business, Service and Resource perspectives. A list of some of the outputs below:

- Automatic Reporting
- Capacity Plans
- Baselines and Profiles
- Thresholds and Proactive Alarms
- Capacity Reports (regular, ad-hoc, exception)
- SLA and SLR recommendations
- Costing and charging recommendations
- Proactive changes and service improvements
- Revised operational schedule
- Effectiveness reviews
- Audits

Modeling
Modeling enables tuning actions to be checked and tested prior to implementation, both at individual component and service wide levels. This avoids time wasted implementing ineffective actions.

Where available, response time data can be used to refine models of tuning activities or planned changes to the environment to show the effect on performance from a user perspective, from both a server and service or application wide perspective.

The Capacity Plan
In practice at many sites, this is not a fixed regular management report but rather an ad hoc report based on a current situation and an urgent need to address specific problems. So the structure for such reports is often not as formal as indicated for the ideal.

Business Scenarios
This part of the plan explains the capacity implications of business strategies, for example the ability to support acceptable service levels for a new marketing initiative without upgrading.

Service Summary
Either in the periodic Capacity Plan, or available 24x7 on the Intranet, a service wide report is required on observed performance using application level views. Trends showing when acceptable levels of utilization will be breached are needed to be automatically available, and detailed changes to the infrastructure to be modeled and reported on in detail from a service perspective.
Resource Summary
All data imported into the Capacity Management Database is available for automatic reporting either in the Capacity Plan or via the Intranet. Data is automatically grouped into whatever time frame, short, medium and long-term, that is appropriate. Trends showing when acceptable levels of utilization will be breached are automatically maintained.

Cost Model
The ability to ‘try before you buy’ is required for the various options for coping with changes in capacity and new business initiatives. This helps you see what you are getting for the money you spend, ensuring the best purchasing decisions are made for the business.

Why is it important?
Having the ability to bring metrics from across the board to your CMIS is very important. The CMIS is the core of the capacity management process, by storing and relating the data that enables decision making. This data is also invaluable to other process and all areas of the business, as illustrated in the diagram below:

Data Capture for Managed Systems
Capturing UNIX Data
**Standard UNIX Utilities** - Most UNIX variants, and indeed nearly all those variants commonly encountered, have standard calls for obtaining various types of performance data. These utilities are sar (for CPU, memory and I/O data), ps (process statistics), vmstat (virtual memory and paging), iostat (in some cases, provides I/O statistics in the absence of sar) and acctcom (process termination and accounting data).
By using the standard UNIX utilities you can gather a large amount of data with a relatively low performance overhead and without the need for kernel intrusive code. But the format of the data can vary depending on the UNIX variant and the data maybe slightly inaccurate (but is still more than adequate for business use) when compared with Kernel based monitoring for example.

**Obtaining Data from Kernel Tables** - Some UNIX monitors are implemented in the form of operating system patches, which read performance data direct from tables in the UNIX kernel. This will provide a wider range of metrics than the standard utilities, but patches are highly dependent on the precise release level of the operating system kernel, and its own patch level. Patches will have to be re-worked when the release level or patch status of the kernel changes.

**Asynchronous Collection** - Whichever method of data capture is used, it is a very good idea to have the local capture of data running as an independent process that does not require a continuous link with any other process or server. This means that even if the data is not collected for some period of time, perhaps because of a problem with the server holding the Capacity Management Database (CMIS), no data will be lost. Captured data can be stored locally until such a time as it is successfully collected.

**Capturing Windows Data** - By comparison with UNIX, the situation with Windows is very easy to deal with. Windows stores performance counters about a variety of system “objects” (processors, memory, network cards, disks, etc) in the Registry. These registry entries are well documented and stable, and are easily read by means of a suitable program.
Management of the CMIS

Administration
The administrator role should be defined at the process implementation stage. The depth of the role will again depend on the size of the environment and the toolset.

Security Implications and Concerns
The security requirements should be determined prior to implementation as certain data elements (such as business forecasts) may be more sensitive than others. The capacity management process should work with Availability to plan and implement the appropriate security processes and protocols.

As guidance the following should be implemented:

- Password controls to limit access data
- Read-only access to non-administrators

Backups and Recovery
The CMIS underpins the Capacity management process and because of this criticality a good deal of consideration should be given to how and when the CMIS is backed up. The scheduling of the backups are of particular importance as we need to ensure that all data is captured and that backing up the data does not impact the operational performance.

The backup process should be documented and where appropriate regularly tested to ensure data integrity.

Tools to Access the Data
The choice of the appropriate tools to access the data will again be dependent on the underlying toolset and the data type in question. Most performance management tools will provide an appropriate interface to access and analyze the data. It is recommended that where feasible the data remains in a read-only format to allow for analysis/aggregation without compromising the data integrity.

Regular Audit and Review of CMIS
The CMIS should be audited regularly to check the following:

- Data integrity
- Data validity i.e. is the data still relevant
Analysis

The Purpose of setting Thresholds - By itself, a measured value of some aspect of the system hardware, as provided by a performance monitoring tool or application, is not worth much. Suppose the CPU utilization varies between 30% and 90% over the working day. Without further analysis and knowledge, all we can say is – so what? The idea of setting threshold values is that over time, the business users and the performance analysts will realize that while certain key measurements remain below (or in some cases, above) specified limiting values, then the performance of the system will be satisfactory.

Exception reporting - If your organization has thousands (or even hundreds) of servers deployed to support various aspects of the business, it is impractical to study performance reports in detail for each one of them. Hence, exception reporting is the norm. This allows performance analysts to concentrate on the more important servers, without wasting time on servers where performance criteria may be less rigid.

Traffic lights
Practically everybody in the world understands the concept of traffic lights:

- green = good
- orange = warning/poor
- red = alarm/bad

Aggregation and time summarization
Creating charts that accomplish desired goals and objectives requires flexibility in how data can be manipulated and displayed. Having an ability to easily accomplish this during report creation is a key feature to have.

Linear Extrapolation
Alarming on trends provides the ability to get proactive notification of thresholds being breached. This is in contrast to monitoring alerts that give notice of problems that often generate incidents and require immediate attention.

Correlation
Having the ability to correlate data, especially across multiple types and variables, is an important consideration when selecting capacity management software.

- Validity against CMS i.e. do the details for common CI’s tracked between the CMIS and CMS still match.
Seasonal Variation
As shown in the picture below, a business can experience significant peaks and troughs in activity over long timescales, even if there is an underlying trend.

Analysis and reporting on peak and average

Naturally it makes sense to base your thresholds on times when the activity is high, and to ensure that your forward predictions of required capacity take into account any seasonal peaks. Capacity management software needs to handle this important task.

Post implementation
Implementing new software almost always presents challenges that must be overcome. Discovering what you didn’t know and experiencing unexpected problems is typically even more challenging. Working through issues takes perseverance and patience, but is all part of the process.

Working with the vendor
Overcoming problems with operating software can often involve vendor support. Support from the vendor is an important part of the relationship formed when purchasing software. Below is a list of vendor experiences that were key to ensuring getting maximum value from the software.

- POC and evaluation
- Sizing exercise during the evaluation
- Product training
- Capacity management workshop
- Onsite consulting
- Product support calls (time difference can be a factor)
Some level of problems can be expected when implementing any large software package and below is a list of issues to plan for, ordered by priority for this case study:

- Getting an accurate list of servers to capture data for
- Understanding naming conventions
- Grouping servers for reporting
- Data management processing
- Software bugs and environmental issues
- Time required to administer and use
- Software limitations

As functionality, data, and processing increase, the likelihood of a problem occurring can increase.

**Growth of resources managed and personnel**

As operations expand after implementation and more data is brought into the CMIS, the amount of work required can exceed the time limits of existing personnel.

Below is a list of tasks that required additional people, ordered chronologically:

- Report creation and maintenance
- Administration
- Capacity Plans
- Automation
- Modeling

**Lessons learned**

Lessons learned along the way are important knowledge that should be documented and used for continual improvement.

- Consider requirements for data management early
- Reporting decisions change
- Capacity plans are always a work in progress
- Modeling can be more difficult than it sounds
- Measure success and value
Looking Forward
Goals and objectives for improvement are important to maintain when looking to improve and expand the value of capacity management across the organization.

- Interface with change management or review process
- Performance testing evaluation
- Expand scope of coverage in organization
- Incorporate other data types like storage
- End to end performance management
- Demand management

Critical Success Factors (CSFs)
For Capacity Management to be successful, accuracy in the information received is crucial:

- Business forecasts
- IT Strategies and Plans

Ensure that you are communicating with teams to ensure value.

It is also important that the organization be able to leverage new technology to maximize performance and optimize costs, where appropriate.

Being able to demonstrate cost effectiveness is vital because Capacity Management is considered a proactive process, the inability to clearly demonstrate value will doom the process. Likewise, not having management support to plan and recommend solutions that are actually implemented could render the process ineffective.

All of this will enable you to plan and implement the appropriate IT capacity to meet business need.
Capacity Management Capability Maturity

Improving capacity management maturity is a continual process, just as the ITIL framework specifies for service management.

Getting to an optimized state requires identifying gaps and filling them in.