Going beyond ITIL®:

IT Capacity Management with SAS

Exploiting SAS business intelligence and advanced analytics capabilities to effectively align IT capacity to business demand using a best-practice and ITIL-compliant IT capacity management methodology

Table of Contents

The content provider for Going beyond ITIL: IT Capacity Management with SAS was Ingo Schulz, product manager SAS IT Intelligence at SAS International headquarters, Heidelberg, Germany.

Introduction: Managing IT Capacity

Data centers represent a large pool of capital investments, yet the utilization of these capital resources is generally low compared to that expected for other business investments. Whether it is because of seasonal peaks or other reasons, these systems are far from 100 percent utilized on average. In order to become accepted as a business partner and value provider instead of a cost center, IT organizations have to adapt the provisioning of IT capacity to its driving business demand – by providing not less, but also not more capacity than required. This paper outlines how to overcome the challenges of modern large scale environments by implementing an ITIL (IT Infrastructure Library) compliant best-practice IT capacity management methodology based on advanced analytics and business intelligence tools.

IT capacity management is a balancing act: balancing service cost against capacity investment and supply against demand. In the mainframe environment with its invaluable resources, IT capacity management has been a well accepted area for many years – not so in the distributed world. But, the typical data center which might have managed just a few systems one decade ago is managing thousands today. Though these large infrastructures represent an immense pool of capital investment, the utilization of these capital resources is generally low compared to that expected for other business investments. Additionally, the complexity of a distributed infrastructure makes it harder to achieve high efficiency, and due to its complexity and decreased agility, error rates are increased. In this new world IT capacity management now faces many more challenges than before.

This paper will take you beyond ITIL's best practices and outline how to:

- Develop an environment for an automated IT capacity management process
- Master ITIL's resource and business IT capacity management processes
- Use advanced analytics to align your demand forecasts with business requirements
- Leverage Business Intelligence tools to effectively deploy IT capacity management reporting.

The main benefits of an automated IT capacity management environment are cost savings by acquisition deferring or even avoiding expenditures. Additionally, companies will reduce the risk of capacity problems and improve efficiency in their service management processes. In order to achieve the maximum benefits it is important to:

- Setup a surrounding framework like ITIL to interface and integrate IT capacity management with other processes, e.g. Configuration Management, etc.
- Implement the IT capacity management processes through a best-practice methodology, like the IT capacity management methodology described in this paper, and

IT Capacity management helps to effectively align IT capacity with business demand while optimizing IT resource costs

- Automate the whole process through sophisticated tools, e.g. for consolidating all data into an IT Capacity Database or for sophisticated forecasting.

Positioning of IT Capacity Management in ITIL

ITIL is a framework of best-practices and the de-facto standard for IT Service Management. IT capacity management, positioned in ITIL's strategic "Service Delivery" area, consists of:

- Monitoring of resources
- Tuning the resources
- Analyzing current demand
- Forecasting future demand
- Producing a capacity plan
- Influencing the demand
- Monitoring of performance and throughput of services.

ITIL organizes the IT capacity management process in three sub-processes:

- Resource capacity management, for ensuring that all components in scope are monitored, measured, analyzed, and reported on.
- Service capacity management, for ensuring that the performance of services is monitored and processed accordingly
- Business capacity management, for ensuring that future business requirements are considered for IT capacity management.

In this paper we will focus on the sub-processes of resource and business capacity management.

ITIL outlines the major steps of how to setup the sub-processes, but it does not provide enough information for a real implementation, e.g. how to integrate future business demand. Therefore, we will describe a straight-forward methodology of how to go beyond ITIL and actually implement the resource and business capacity management processes for IT capacity planning.

The IT Capacity Management Methodology

The following methodology is based on input and discussions with Alan Knight, the founding chairman of the UK CMG and a widely known expert in the IT capacity management area for many years.

The methodology can be applied to IT capacity management for whole IT infrastructures as well as isolated scenarios, e.g. will a new disk system support next year's business demand?

Figure 1: IT Capacity management methodology

The methodology is pragmatic and straight-forward:

1. **Define Scope**: Define the services, resources, components, and their corresponding priority to be taken into IT capacity management.

2. **Define Standards**: Define and consolidate data sources, define metrics, thresholds, reporting formats, and reports for the defined scope.

3. **Establish current resource usage reporting**: Based on the scope and the standards analyze the current resource usage and identify seasonality, peak periods and characteristic values in order to report on current resource usage.

4. **Establish resource demand forecasting**: On the basis of historical demand represented by characteristic values, analyze future demand.

5. **Assess other impacts**: Integrate planning data like future business, service, and technical data into the statistical forecast in order to align the capacity to future demand and evaluate results.

6. **Communicate through a business aligned capacity plan**: Integrate financial data and reports into a business aligned capacity plan and communicate recommendations in business language.

This IT capacity management methodology can be applied for dedicated and shared resources. For shared resources, every workload or application using the resource has to be monitored, analyzed and planned for on its own.

The methodology can also be applied to all types of resources and their components, e.g. CPU, memory, disk space, or networks. In the examples and figures included in this paper the main focus will be on CPU resources as these are probably the most complex resources to handle.

- The pragmatic, best-practice, universal IT capacity management process for all environments

Applying the Methodology

In the following sections we will go through each step of the methodology in more detail.

Step 1: Define Scope

In general the scope will be defined by a business reason, e.g. does the infrastructure have enough capacity to support the effects of the next marketing campaign for a certain business service? Given this business reason, one has to identify the critical components supporting this business service, i.e. which capacity shortages on which resources could harm the provisioning of the service as defined in your Service Level Agreements (SLA's).

The more detailed and sophisticated the IT capacity management the higher the costs. Therefore, as costs for different resource types vary the most expensive and important resources should receive priority for the more detailed planning of the business service. For those resources of moderate cost and importance one could use simple trending or other simplified planning techniques, and for the less expensive resource types just plan for excess capacity, e.g. instead of upgrading to 2GB of memory, plan for 4GB.

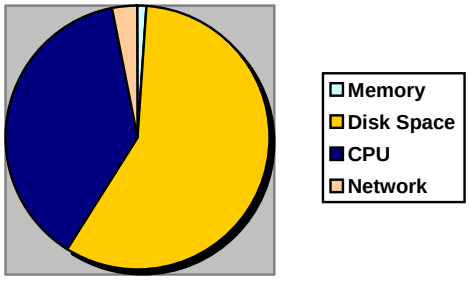


Figure 2: Typical relative resource costs

One could also assign a priority based on the criticality or the discontinuity agreements of the resource in order to allow a better granularity in reporting.

Once this step is completed one should have defined a list of services, resources, e.g. certain servers and corresponding components, such as CPU and disk space, that are in scope for further detailed IT capacity management.

Step 2: Define Standards

In order to be able to automate the whole process and make it efficient one has to define some standards.

2a. Define and consolidate data sources

Based on the scope, one needs to decide on corresponding data sources or tools to measure resource usage. This decision will be influenced by the platforms in scope, by the availability of tools, and if one is dealing with shared resources like mainframes or dedicated resources like simple web servers. Even if one has no commercial tools for measuring resource usage, ask the platform experts for platform integrated tools such as. df, du, or sar on UNIX or freely available tools such as Nagios.

Having identified all necessary data sources one should consolidate all necessary IT measurement data and other data, e.g. configuration data, containing thresholds and absolute capacities or business data, into one central IT capacity database (CDB).

ITIL also describes some features such as the CDB that should offer features such as auto-aging, checking for duplicates, normalization or archiving. It is important that this database is the designated core data source for all IT capacity management tasks, incorporating all your data in one central data store.

2b. Define device independent metrics

In order to facilitate automation and analysis one should standardize on one common metric per resource type, e.g. for CPU, memory, disk space. In general, IT capacity management deals with how resource usage changes over time. However, as the absolute capacity might change, e.g. a new disk is added, it is important for reporting that the units in which the metrics are expressed are device independent. If this is not the case, then it would not be possible to produce a graphical representation that involves a change of resource capacity over time.

In addition, it is important to use device independent units to enable consolidation. For example, measuring the CPU load as a percentage would not allow for a statement of how much workload in absolute values is left over for consolidation. It would be very difficult to move workloads across hardware platforms.

Typical examples for device independent units are:

- Storage capacity being reported in mega, giga, tera, or peta bytes.
- Data transfer capacity being reported in megabits/bytes per second.
- Processor capacity being reported in MSUs or SPECints.

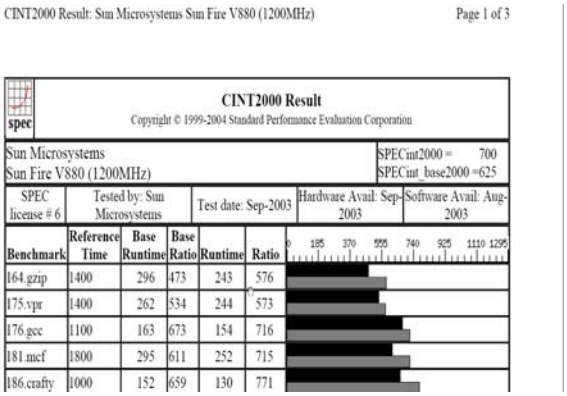
Using MSU as the measurement of capacity allows device independence for mainframe IT capacity management.

Device-independent metrics allow for long-term planning and comparing workloads on different platforms and thus enable consolidation planning

Mainframe Service Unit (MSU) Capacity is specific to the IBM z/OS and OS/390 operating systems (OS). Measured in MSUs or Millions of Service Units per hour, it is a capacity measure applying to both the entire machine and partitions (LPARs) of the machine. The MSU capacity of the entire physical system (or the "Central Electronic Complex") is known as the "CEC Capacity." Each LPAR running an OS on the system has its own MSU capacity, which is known as its "image capacity." This image capacity has become an industry standard and is the value used in sub-capacity software licensing. In the simplest case, a machine may have just one LPAR in which case the image capacity equals the CEC capacity.

The image capacity of an LPAR is determined either by the capacity of the resources assigned to the LPAR or can be set explicitly on the HMC (Hardware Management Console) in the Defined Capacity setting. A customer would use Defined Capacity in situations where LPARs share resources with each other and the capacity of a given LPAR is intended to be restricted to only a fraction of the resources available to it.

SPECint is a device independent metric for CPU processing power for non-mainframe servers, defined by the Standard Performance Evaluation Group. The Standard Performance Evaluation Group hosts several performance metrics but the SPECint in its release 2000[1] is probably the most important one.



CINT2000 Result: Sun Microsystems Sun Fire V880 (1200MHz) Page 1 of 3

CINT2000 Result

Copyright © 1999-2004 Standard Performance Evaluation Corporation

Sun Microsystems
Sun Fire V880 (1200MHz)

SPECint2000 = 700
SPECint_base2000 =625

SPEC license # 6	Tested by: Sun Microsystems	Test date: Sep-2003	Hardware Avail: Sep-2003	Software Avail: Aug-2003

Benchmark	Reference Time	Base Runtime	Base Ratio	Runtime	Ratio
164.gzip	1400	296	473	243	576
175.vpr	1400	262	534	244	573
176.gcc	1100	163	673	154	716
181.mcf	1800	295	611	252	715
186.crafty	1000	152	659	130	771

0 185 370 555 740 925 1110 1295

Figure 3: Sample SPECint report

SPECints are measured by executing the freely available tools of the SPECint suite. Alternatively one could also access the official SPECint reference database at http://WWW.SPEC.org which hosts the SPECint results for various hardware platforms. Figure 3 shows an example of such a reference report where the computed SPECint value for a Sun Fire V880 is 625. As a best practice one should always use the SPECint_base value as this value was computed using normal compiler settings not leveraging the platform dependent compiler optimization options as in the SPECint2000 value.

2c. Define thresholds

For planning purposes and for mechanisms such as exception reporting it is best to deal with three kinds of thresholds:

[1] Since August 2006 SPECint 2006 is available.

- Absolute capacity - the maximum theoretically achievable capacity of a resource
- Alert threshold - exceeding this threshold will affect the defined Service Level Objectives
- Warning threshold – exceeding this threshold should raise a warning.

The warning threshold is particularly useful for automated exception reporting as proactive attention can be brought to resource utilisation issues before they become problems.

The alert threshold is used for planning purposes to define the effective capacity of a resource so that the life span of the configuration can be estimated.

In addition, a resource will have an absolute capacity which cannot possibly be exceeded. Attempting to use a resource to this level can affect performance to varying degrees, depending on the resource type.

Defining the thresholds should be accomplished in collaboration with corresponding hardware experts, but in general the actual thresholds are mainly dependant on:

- Underlying Service Level Objectives (SLO's)
- Priority of the resource – the more critical the resource the lower the threshold
- Type of resource – e.g. CPUs can handle short term peaks achieving the total capacity, but disk space capacity cannot handle this type of peaks
- Type of resource usage – demand peaks exceeding thresholds could be tolerated in batch mode, whereas in online mode they should be avoided
- The data centers personal risk potential.

The most accurate way to derive thresholds is through actual measurement of performance impacts in a testing environment. If such a testing environment does not exist or measurement is too expensive, one could derive some good thresholds with simulation modeling. However, for deriving CPU thresholds one will need to take into account the number of processors: the more processors the higher the alert and warning thresholds as well as the higher the overhead for process distribution. The system overhead can be identified by the corresponding hardware expert. In general, good overhead values to start with are 7% for an eight processor machine or 5% for a four processor machine.

The following figure shows the response time multiplier for different loads for various multi-processor configurations. It shows that a uni-processor machine at 80% load is already five times slower than without any load, while a 2-processor machine gets five times slower at around 88%. The plots are derived from queuing theory calculations and provide a good starting point for setting thresholds.

■ Define different thresholds for pro-active management like exception reporting and long-term capacity planning

- Queuing theory can be used to derive first threshold candidates

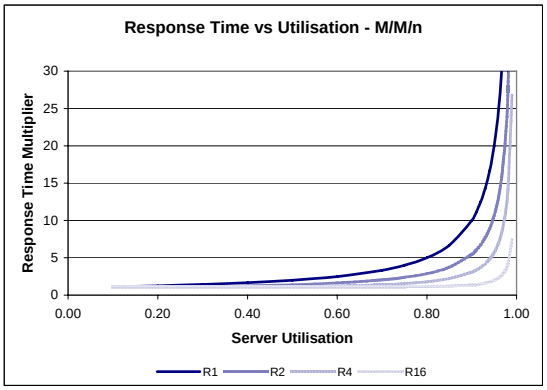


Figure 4: Response time multiplier vs. load for several processor configurations (Source: Alan Knight)

Knowing the SPECint value for a certain server and the number of processors, it is easy to derive initial simple thresholds. Given an eight processor machine with a base value of 625 SPECint, one could compute the following thresholds:

- Absolute capacity: 625 * 8 * 93% = 4650
- Alert threshold, e.g. at 90%: 4650 * 90% = 4185
- Warning threshold, e.g. at 80%: 4650 * 80% = 3720

2d. Define reports

The list of reports will vary depending on reporting requirements and the tools available for reporting, but the general guideline is to first identify the stakeholders, identify their requirements, and make it as easy as possible to accomplish their tasks. In large scale environments one would also have to deal with other issues, which will be covered in the next chapter.

Some sample stakeholders in an IT capacity management environment are the IT Director, as a sponsor, the capacity planner for dealing with future resource demands (ITIL's business capacity management), and the computer performance Analyst for handling current capacity issues (ITIL's resource capacity management process).

As a sponsor, the IT director is interested in the financials and recommendations resulting from the IT capacity management processes, so they will be the primary consumer of a formal capacity plan and IT dashboards or IT scorecards monitoring IT capacity management key performance indicators (KPIs).

The capacity planner and author of the capacity plan will need all reports, which underpin the capacity plan and the resulting recommendations. As the plan deals with future demand, they will need the corresponding forecast reports, the underlying financials, and also reports for evaluating the IT capacity management process.

The Performance Analyst who deals with current resource usage will need overview reports, detailed resource usage reports, and exception reports.

This list cannot be generalized to meet the requirements for every environment. In each case the requirements of the consumers and the features of the available reporting environment have to be analyzed.

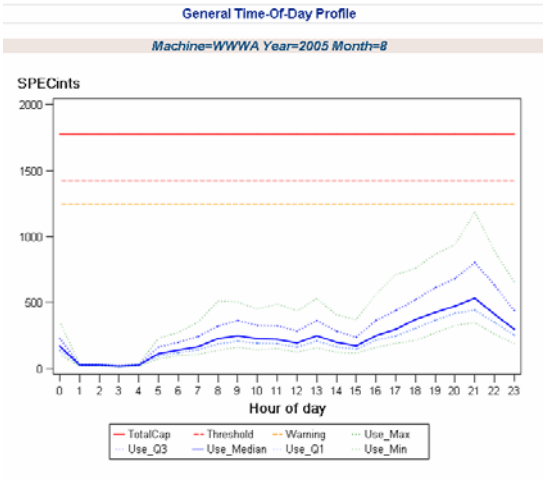


Figure 5: Sample report: Time of day profile

2e. Define reporting standards

Last not but least, in order to simplify analysis and to make it appear even more professional there should be a standardization on certain reporting standards and guidelines, for example:.

- Use of unique report titles and standardized header and footer lines containing all applied filters, the date and time of the report generation
- Standard colors and line styles, e.g. solid lines for total capacity, actual values, dashed lines for thresholds, …
- Standardized labeling and sizing of axis
- Unique report identification coding is especially powerful in larger organizations where many reports are produced and distributed. Requests for report change or queries can be simplified with a report coding system in place.
- Reporting naming or numbering standards

Reporting standards are important to setup an efficient and easy to understand communication channel

After all the sub-steps of “define standards” have been accomplished one should have achieved the following:

- All data sources consolidated into one core Capacity Database for further analysis and reporting
- Standardized a set of device independent metrics for all resource types in scope
- Defined all thresholds for the resources in scope
- Identified all stakeholders and their reporting needs

- Established reporting standards.

Step 3: Establish Current Usage Reporting

Based on the scope and the defined standards the establishment of reporting on the current resource usage can be accomplished along with ITIL's resource capacity management process.

The key question is: How to report on monthly resource usage? Does one use the average or the maximum resource usage? The first step to enable reporting and future demand planning is to identify characteristic values of a workload that are representative for the resource usage. These characteristic values can be used to report on historical resource usage as well as for planning the future resource demand.

It is best practice to derive characteristic values from the peak period of resource usage. The peak period is a period of time for which resource usage will be planned or reported on so that the plan or report will accurately reflect the demand for a resource e.g. for a typical day, or for a typical month.

The peak period will always be a part of the duration considered as important for the business application or IT service, i.e., the times during which either SLO's apply, or, where no formal SLO's are defined, the times are accepted as important. The "important" time of day might be the entire 24 hour day or just the time from 9am to 5pm. Peak periods may be defined for individual workloads or for whole servers.

Characteristic values needed for monitoring and reporting derive from SLO's

As already stated, a characteristic value will always be taken from the peak period. It may be defined in a number of ways in terms of the duration and point in time of the period, the statistics used to characterize the resource usage and the point in the business cycle at which the period is measured. It mainly depends on existing SLOs: if one has to plan for the peak hour of the peak month of the year, the maximum resource usage must be used as a characteristic value for reporting and planning. However, if just an average demand on an average day is needed, then taking the average resource usage might be enough while the peak might be too high. A best-practice in this case is to use the 75th percentile to define the representative value.

Seasonality is a repeated pattern of resource usage that can be found on the weekly, monthly or yearly level. Because there is a need to deal with seasonality at the daily, weekly, or monthly levels one must identify a value that will accurately represent resource demand.

The characteristic value can be defined at both the workload and the resource levels. Typical examples for characteristic values are:

- The 75th percentile of 15 minute resource utilisation averages during the 08:00 to 18:00 period on weekdays
- The maximum hourly average resource demand on the peak day
- The 75th percentile of the hourly averages from a working month.

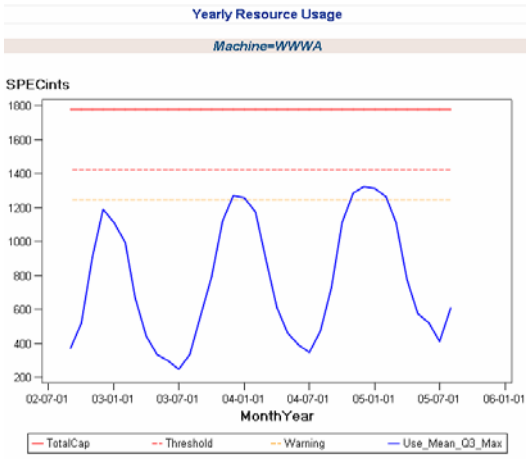


Figure 6: Yearly resource usage based on characteristic values

As a best practice the definition of characteristic values can be performed as follows:

- Hour: hourly averages of ten minutes intervals
- Day: 75th percentile of the daily hourly averages or all daily values
- Month: Maximum of the 75th percentiles of the hourly averages.

As illustrated in the time of day profile in Figure 5, the 75th percentile always is a good compromise on the day level as the average might be too low, while the maximum might be too high for planning resulting in over sizing the resource.

Having identified the characteristic values one is now able to plot a usage trend on a larger scale as shown in Figure 6. This will also serve as a basis for forecasting characteristic values in order to get a forecast for a representative demand.

It is a common best practice to deploy reports through an IT capacity management portal as shown in Figure 7.

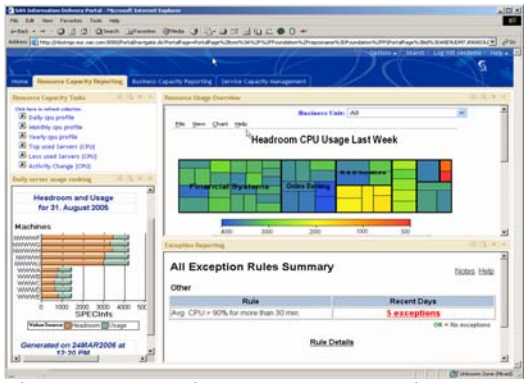


Figure 7: IT Capacity management Portal

Publishing through portals is an effective and efficient way to provide the most convenient reporting environment to information consumers

A portal offers the following advantages:

- Provides an easy access to web-enabled reports
- Offers a single-point-of-management for IT capacity management tasks,

- Fully customizable to each user's job requirements, as well as their personal preferences
- Offers the capability to produce drill down reports
- Enables an effective communication because the reports are available everywhere, and…
- ….offers an entry point for on-demand reports through your IT capacity management "toolbox" consisting of predefined IT capacity management procedures.

Especially for large scale environments, it is most effective to deploy reports through a portal. Other best practices for reporting in large scale environments are:

- Automation of all processes where practical
- Management by exception: Definition of exception rules that automatically create exception reports, e.g. when the alert threshold has been exceeded
- Exploitation of computational intelligence to interpret results through usage of filters, sorting, grouping, e.g. report just on the top 10 resources, workloads, etc.
- Interactive analysis by enabling drill down report structures
- Usage of on demand reporting where possible in contrast to batch creation of reports, e.g. through definition of a "toolbox" of predefined routines for IT capacity management containing parameterized routines for time of day profiles, etc. as shown in the upper left area of Figure 7
- Intelligent representation of large amounts of data, e.g. through a 'tile' or 'heat' chart as shown in the upper right area of Figure 7

The tile chart is a powerful visualization technique for sifting through masses of data at a glance. It uses size and color, so that problem areas are highlighted. Size and color work in tandem to emphasise the areas of concern: Size can represent importance, color can represent status, e.g. size of a rectangle could represent the power of a server in terms of SPECints and the color could represent the load on this server ranging from deep blue to red (see Figure 8). A tile chart can be implemented to support drill-down capabilities to link specific tiles to other tile charts or websites. A tile chart also offers a menu for dynamically re-configuring the hierarchy, size and color, the color scheme, and titles and footnotes. Additionally, it provides mouse-over capabilities to display details of the underlying tile.

The tile chart is a very powerful visualization technique for sifting through masses of data at a glance

By using a tile chart for visualization analysts can focus and drill-down into red rectangles to find out if and why servers are overloaded or also focus on blue areas for identifying consolidation opportunities.

Once this step is completed one should have accomplished the following:

- Identified peak periods, seasonality and characteristic values for all workloads
- Created the major reports for the major consumers
- Established resource capacity reporting through a flexible portal

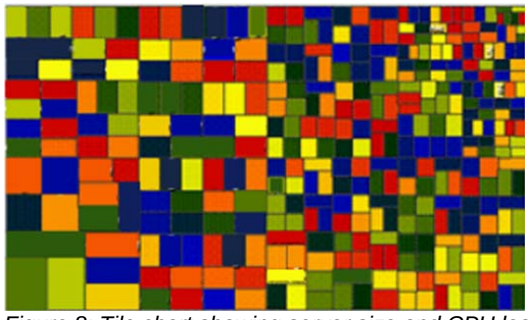


Figure 8: Tile chart showing server size and CPU load

Step 4: Establish Future Demand Forecasting

ITIL's business capacity management sub-process is accessed by using the historical resource usage data – represented by the computed characteristic values – to forecast the future demand.

There exist many techniques to forecast time series data, ranging from simple regression techniques to very sophisticated models. Which techniques are applied mainly depends on the available statistical knowledge and the availability of advanced analytical tools. The advantage of using simple models is that they are supported through a lot of tools. However, they may be very inaccurate as e.g. they often do not support the notion of seasonality, while more sophisticated models require much statistical knowledge or advanced analytical tools.

- Advanced analytics use historical values as the base for predicting future resource utilization

The next figures illustrate two simple models using an exponential smoothing model. The first model (Figure 9) is very simple and just continues the trend of the time series. The second model (Figure 10) also takes the seasonality into account. Obviously, the effects of choosing one over the other are dramatically different.

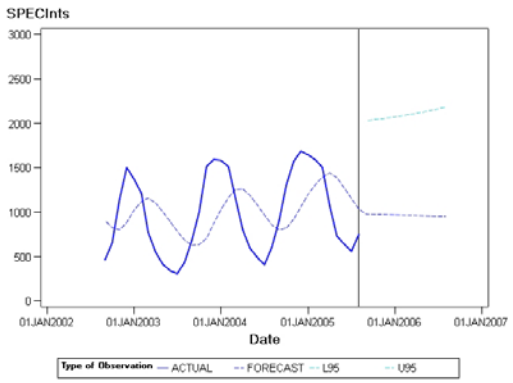


Figure 9: Exponential Smoothing Model (ESM) with trend

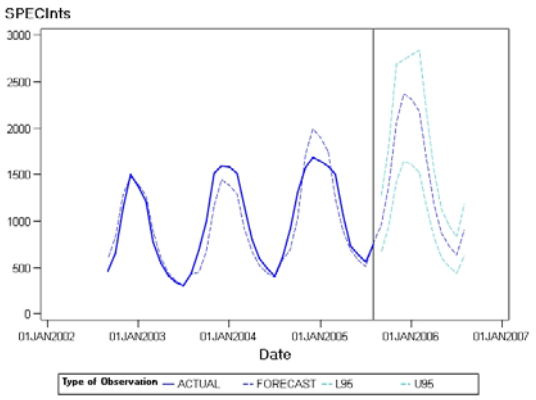


Figure 10: Winters model: Exponential Smoothing with seasonality

Creating a forecast from time series data will always involve:

1. Identifying the class of models to be considered
2. Picking the appropriate models
3. Estimating the parameters of the model
4. Creating the forecast (including confidence intervals).

Whilst the challenges of statistical forecasting are:

- The availability of skilled statistical analysis resources
- The possible generation of many forecasts
- Frequent forecast updates maybe required
- Forecasting model may not be known for each time series
- Calendar events and business drivers (inputs) may influence the forecasting accuracy.

■ Forecasting in large-scale environments needs high performance and highly automated analytic tools

As one may conclude, especially in large scale environments there may be a need advanced analytics, e.g. in the form of high performance forecasting environment, to master the challenges of statistical forecasting as it may not be possible or cost efficient to manually develop and evaluate the forecasting models for several hundred servers. A high performance forecasting environment potentially offers the following features:

- Scalable production of forecasts at high speed
- Automated Identification of model, parameter estimates, and selection of best model
- Support of input of independent variables by automatically applying correlation analysis
- Input of future calendar events e.g. to support the effects of future marketing campaigns or technical changes

- Support for different forecasting model families
- A GUI, but should also support batch automation
- Ability to guide the novice to produce acceptable forecasts, as well as guide the expert in the analysis of different models
- Must be capable of integration into production workflows.

The following figures give a few insights into how the production in such a high performance forecasting environment could appear:

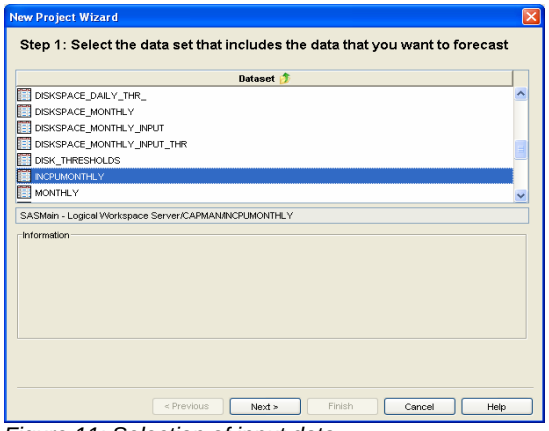


Figure 11: Selection of input data

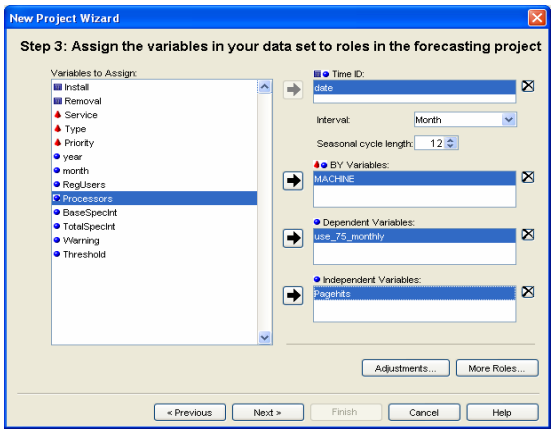


Figure 12: Assignment of table columns to forecasting roles

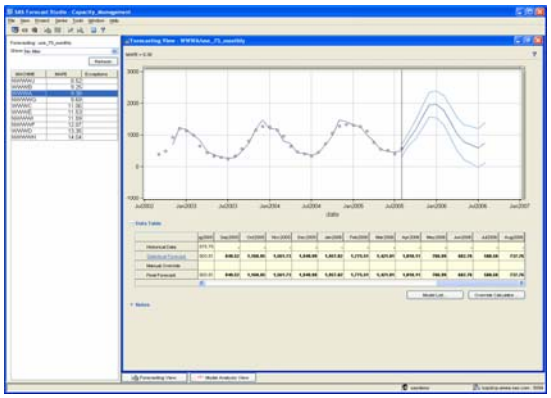


Figure 13: Resulting forecast for all resources as plots and tables, and with corresponding statistical error

If there is no possibility to access such a high performance forecasting environment there could still be the possibility to automate the production of simple regression models based on the measurements in the IT Capacity Database and produce good trend models. But the analyst must be aware that these models may be inaccurate due to seasonal components in the time series data.

After completing this first step towards ITIL's business capacity management, all statistical forecasts or at least trend models for all relevant workloads and resources in scope should have been established.

Step 5: Assess Other Impacts

- Historical values are just the base that has to be enriched with technical and business planning information to accurately forecast and align IT capacity to business demand

In this step we will enrich our pure statistical forecasts with planning data in order to accurately represent future business demand. It's like driving a car: the driver wouldn't drive forward by only looking into the rearview mirror in case a bend appeared in front of them. Therefore in this step the analyst also has to integrate planning data to accurately plan for future resource demand, e.g. expected business volumes or infrastructure changes such as an application upgrade.

In order to integrate business data the analyst must integrate input data from the business side - historical business volumes as well as expected business growth. This input is called Business Forecast Unit (BFU).

A BFU is a metric in business terms that expresses business growth and volumes, which must be a direct driver of resource consumption, i.e. one can compute the resource consumption by one BFU. Typical examples for BFUs are the number of business transactions, served users and web hits. It is very important that the BFU is a direct driver for resource consumption.

In general, correlation analysis for identifying if a BFU is a direct driver for resource usage and how a BFU drives resource usage can be deployed. Unfortunately, BFUs are also defined in time series format, thus simple correlation is not appropriate for time series, i.e. simple correlation analysis between time series data. If two series both have a trend tendency they would show a high correlation even though they are not really correlated. In order to analyze the correlation the transformation of both the input and output series must be performed and then analysis of cross correlations between these series. The transformation is called 'pre-whitening' and can be accomplished through an ARIMA (**A**uto-**R**egressive **I**ntegrated **M**oving **A**verage – time series forecasting model) procedure (see Figure 14). The pre-whitening technique will provide the analyst with a regression factor representing the correlation between the BFU and resource usage, e.g. in figure 14 the regression factor is 0.000024 which means that 1/0.000024=40000 page hits will drive resource usage by one SPECint. One can also identify if there is a lag between the BFU and the resource usage. A high correlation at lag 0 means that the input has an immediate impact on the output. A high correlation at lag 1 means that the input has a high impact on the output one time unit later.

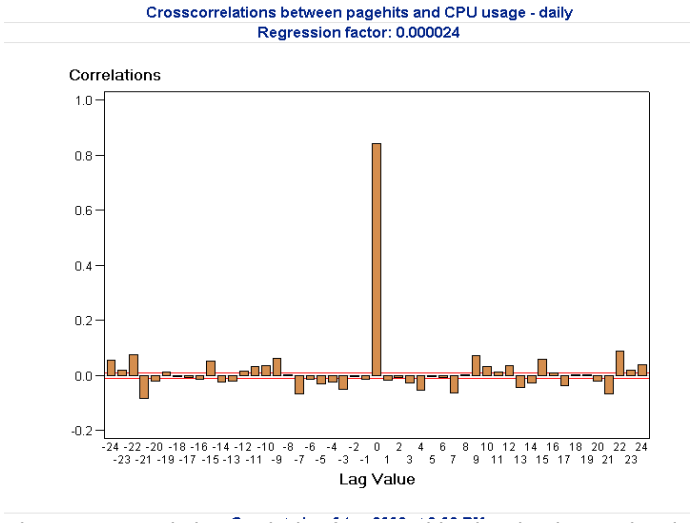


Figure 14: Correlation analysis after pre-whitening the time series data

This has always been, and still is, quite a complex statistical process that can only be applied by statisticians.

Fortunately, advanced analytical tools such as high performance forecasting environments that support the input of additional input variables (see input of independent variables at the bottom of Figure 12,), automatically apply this technique and automatically integrate the input into the forecast. As a result, the analyst can produce forecast graphs, as shown in Figure 15 that integrate business volume forecasts in terms of web page hits.

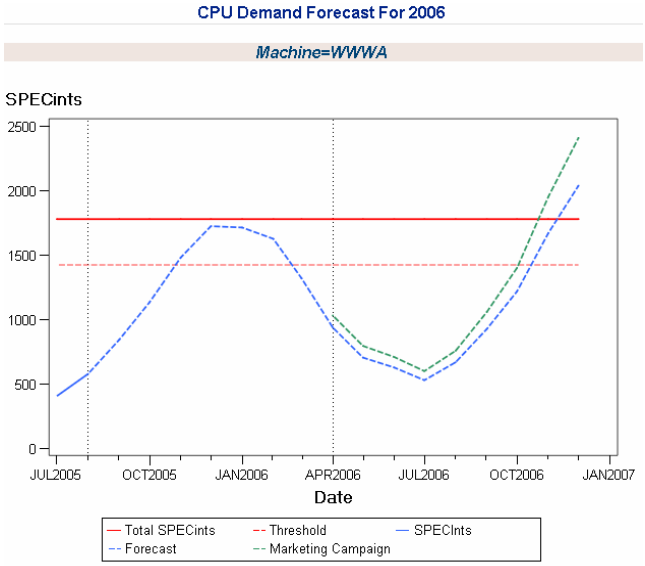


Figure 15: Integrating business growth through planned marketing campaign on 1st April

If the analyst already knows the direct effect on resource usage, e.g. by analyzing stress test results of the new release of an application, they can compute the resulting demand by first forecasting the demand without any input and then modifying the resulting forecast, e.g. increase the resource demand by 10% from 1st July due to release of new online application (see Figure 16).

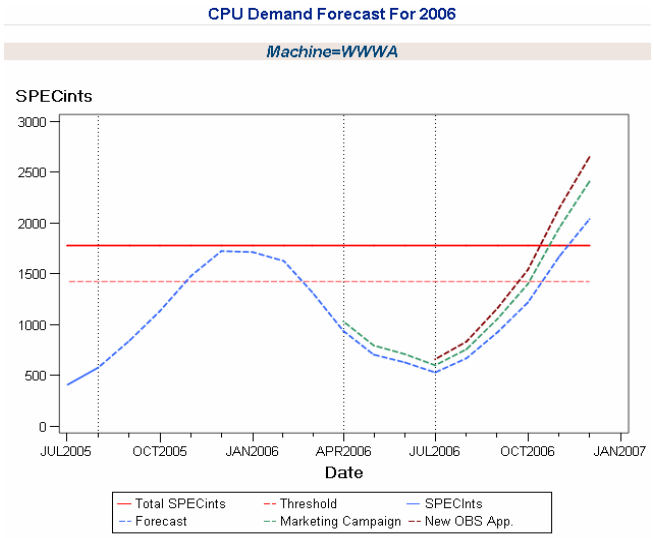


Figure 16: Integrating additional 10% demand increase due to new application release

After completing this step one should have integrated all other impacts affecting resource usage into your statistical forecasts.

But what are the results? In Figure 16 one can identify that the server will not support the resource demand adequately in 2006. So, there is a need to analyze the situation and arrive at a recommendation for the IT director.

As the server will run out of capacity the analyst has to evaluate the alternatives: upgrade the server, buy a new server, or consolidate the workload onto another existing server. Assuming that the overloaded server does not support any more processors subsequently there needs to be a check for consolidation opportunities.

- IT Capacity management involves analyzing resource optimization alternatives

The main reasons for consolidating workloads are to utilize IT investments more effectively, avoid the costs for buying new hardware, and reduce the complexity of the IT infrastructure by having less components to manage. How does one consolidate?

1. Find a consolidation opportunity: Which other servers on the same platform offer the most headroom?

2. Combine the two workloads by summarizing the historical workloads and analyze the combined workload.

3. Forecast the single workloads and analyze the combined workload for the future.

One could also check on the daily level if the two workloads to consolidate peak during the same peak period. If they do then a plan should be instigated with aim of investigating additional headroom. The following figures represent the described steps to consolidate.

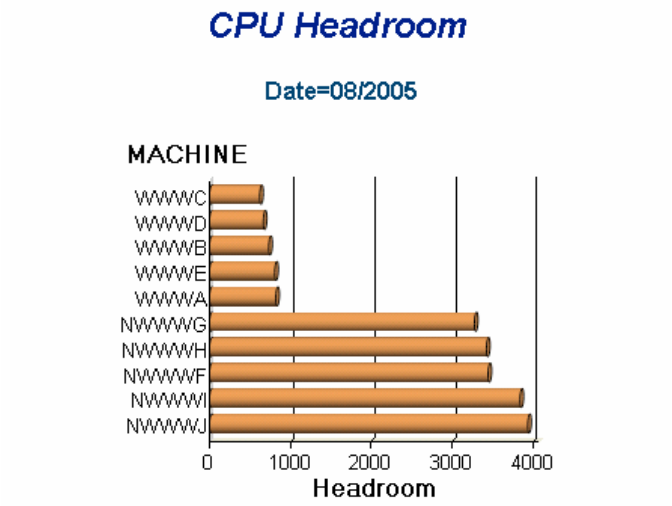


Figure 17: Sorting servers by CPU headroom

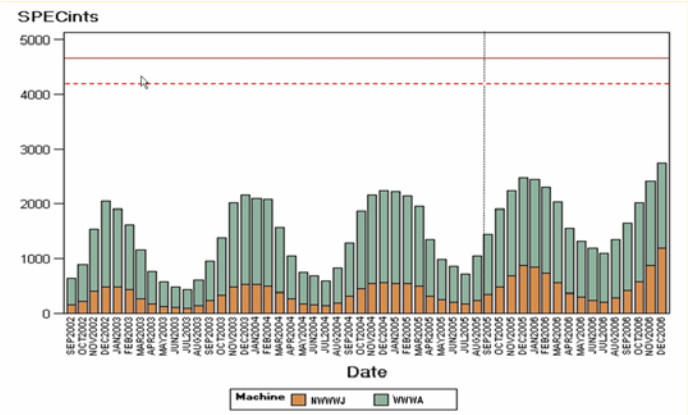


Figure 18: Combining the workloads, and forecasting the single workloads

Having identified a consolidation opportunity there finally has to be a comparison of the financial aspects of both scenarios. In the replacement scenario we assume that a new SUN Fire V880 server would cost $50 thousand. Distributing this cost over five years would result in a yearly acquisition cost of $10 thousand. In addition we assume a yearly maintenance cost of $10 thousand, a first setup cost of $4 thousand, and a cost of one thousand dollars for the removal of the old server. This would add up to the following cost scenario:

- Optimization alternatives have to be analyzed from a financial point of view

Replacement Scenario (first year costs)	
Task/Item	**Cost (in thousand dollars)**
Acquisition V880 p.a. on 5 years	$10
Maintenance	$10
Setup	$4
Removal old	$1

Total cost after 1st year	**$25**

Figure 19: Cost analysis first year for replacing server WWWA

Replacement Scenario (following year costs p.a.)	
Task/Item	**Cost (in thousand dollars)**
Acquisition V880 p.a. on 5 years	$10
Maintenance	$10
Total costs for following years	**$20**

Figure 20: Cost analysis for following year costs for replacing server WWWA

Adding up the costs for the first year and the following four years would result in a **total cost of ($25 + 4 * $20) = $105 thousand for a five year period for the replacement scenario**.

Now one has to compare the replacement scenario to a consolidation scenario. In the consolidation scenario there are no acquisition costs. We could assume a more expensive maintenance of additional $5 thousand dollars (in addition to the already existing maintenance cost of $10 thousand that one has to spend anyway) and one could assume that the costs for the setup are higher because of a more complex setup ($10 thousand instead of $4 thousand). This would add up to the following costs.

Consolidation Scenario (first year costs)	
Task/Item	**Cost (in thousand dollars)**
Maintenance (additional)	$5
Setup	$10
Removal old	$1
Total cost after 1st year	**$16**

Figure 21: Analysis first year costs for consolidating workload of server WWWA

In the following years we just have to take into account the increased maintenance costs of $5 thousand in addition to the already existing maintenance costs of the consolidation server.

Consolidation Scenario (following year costs p.a.)	
Task/Item	**Cost (in thousand dollars)**
Maintenance (additional)	$5
Total costs for following years	**$5**

Figure 22: Cost analysis for following year costs for replacing server WWWA

Adding up the costs for the first year and the following four years would result in a **total cost of ($16 + 4 * $5) = $36 thousand for a five year period for the replacement scenario**.

Comparing both scenarios leads us to a cost saving of $69 thousand in the consolidation scenario. Therefore the IT capacity manager will recommend the consolidation to the IT director.

Step 6: Communicate through a Capacity Plan

Once all reports are created, the results will need to be communicated to management through a standardized IT capacity plan. In contrast to the technical forecasts and usage reports, the IT capacity plan is intended for an IT management and financial audience. In addition to the recommendations, the underpinning financials also have to be embedded. The following is a sample capacity plan structure as proposed in ITIL:

- Results and recommendations have to be communicated in a transparent and non-technical language through a standardized IT capacity plan

1. Introduction
   - Scope
   - Methods
   - Assumptions
   - Management Summary
   - Business Scenarios
2. Service Summary
   - Current Services
   - Service Forecasts
3. Resource Summary
   - Current Resource Usage
   - Resource Forecasts
4. Recommendations
   - Options for Improvements
   - Cost Model
   - Recommendation
5. Appendix
   - Glossary
   - Reporting Guidelines

In order to simplify and automate the process of capacity planning document generation, it makes sense to combine the information contained in the CDB and accompanying report generation logic with desktop tools and applications. So, for example, the capacity planning MS-Word document could be automatically refreshed through the embedding of automatic data retrieval/manipulation and reporting features. Figure 21 shows an example of such capability.

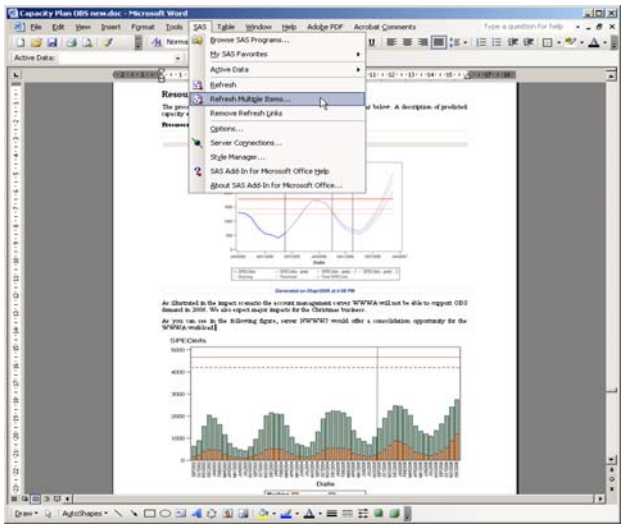


Figure 23: Automatic refresh of reports in Microsoft Office

Summary

A well performed IT capacity management process effectively aligns IT capacity to business demand. IT capacity management ensures optimum business continuity, reduces costs by avoiding or deferring unnecessary expenditure, and improves the utilization of the existing IT capital investment.

IT capacity management is well positioned in ITIL, but doesn't have the provision of detail of how to actually implement an IT capacity planning process. In order to maximize the benefits of IT capacity management, one has to go beyond ITIL and apply a pragmatic IT capacity management methodology and automate the whole process through the use of advanced analytics and powerful business intelligence tools.

Based on the defined scope and standards of the IT capacity management environment the major areas for automation are to:

- √ consolidate all input data into one core IT Capacity Database
- √ publish reports created in batch and provide access to on-demand reports through an IT capacity management portal
- √ forecast future IT capacity demand and assess other impacts through advanced analytic capabilities
- √ integrate financial data into the IT capacity management process
- √ communicate through a dynamically refreshable capacity plan