The Association of System Performance Professionals

The **Computer Measurement Group**, commonly called **CMG**, is a not for profit, worldwide organization of data processing professionals committed to the measurement and management of computer systems. CMG members are primarily concerned with performance evaluation of existing systems to maximize performance (eg. response time, throughput, etc.) and with capacity management where planned enhancements to existing systems or the design of new systems are evaluated to find the necessary resources required to provide adequate performance at a reasonable cost.

This paper was originally published in the Proceedings of the Computer Measurement Group's 2000 International Conference.

For more information on CMG please visit www.cmg.org

Six Levels of Sophistication for Capacity Management

George I. Thompson
IBM Global Services

Six levels of Capacity Management sophistication are defined and reviewed. The levels evolve from using the most basic approaches to including Capacity Management as part of a total IT Enterprise Resource Management Architecture. Included are illustrations based on actual Capacity Management programs with an emphasis on simpler approaches and distributed computing. Topics include methodologies for workload characterization, forecasting, and predicting resource capacity with some comments regarding the current state and future of Capacity Management.

Six Levels of Capacity Management Sophistication

The following conceptual framework has been a useful reference for marketing & implementing a Capacity Management (CM) program, and the personal development of automated CM solutions. It also could be used for developing a CM program audit standard. There have been many previous good papers and books written on Capacity Management processes[1], which generally describe fairly similar process steps.

This paper's objective is not to focus on the horizontal process steps, but rather to suggest a formal outline of how we may use less sophisticated approaches than what is considered ideal, due to a lack of resources, a lack of available data or measurements, a lack of software automation, or even differences in management philosophy about the cost/benefit of using Capacity Management solutions. Each describes how to perform solutions in an increasingly sophisticated fashion.

Simple methods and solutions may need to be used initially to demonstrate an initial Capacity Management benefit or to get the program started. Then, over time, more costly, sophisticated solutions, if technically feasible, can be justified as enhancements to derive further benefits. In my experience most IT environments operate at fairly low levels of sophistication even for the more mature mainframe environments.[2]

The lower levels of sophistication will seemingly provide similar deliverables as the more sophisticated levels by using rules of thumb for some requirements or simply leaving out more detailed process steps with the corresponding supporting data and analysis. Data requirements may be provided in a supplemental fashion, using an unspecified process not formally included in the CM program. This can have interesting side effects regarding responsibility and accountability, since the end result of the less sophisticated approach will generally be expected to have less accuracy. Also, often out of necessity, the lower level approaches usually require the involvement of more staff with a diversity of knowledge and expertise with the resulting additional cost of staff time analysis and meetings.

Proceeding from describing the low levels to the higher levels, it is implicitly assumed that the lower level processes are retained, although some aspects may be replaced or eliminated with the higher level, more efficient or rigorous solutions.

The CM levels of sophistication are summarized as:

Level 0 No formal ongoing CM program ; CM occurs as an occasional project.

Level 1 Formally measure, trend, & forecast peak period utilization and plan resource capacity with an ongoing periodic review program.

Level 2 Accurately partition resource utilization by meaningful workloads. Integrate workload data with other IT ERM[3] components. ["Light" – single attributes; "Heavy" – multiple attributes including company /department, application, & process type workloads.]

Level 3 Include an automated workload forecast system. ["Light" corresponds to Level 2; "Heavy" – Provides for distribution of summary forecast to lower levels to allow for forecasts re-summarized by different categorizations.]

Level 4 Predict service levels & capacity requirements using (level 3) workload forecasts. ["Light – Use rules of thumb; "Heavy" – Use configuration modeling tool]. Inte-

grate processes with other IT ERM Components.

Level 5 Use business application criteria with an application model to predict service levels and forecast resource usage requirements of the most resource intensive client applications. ["Light" – Use simple correlations; "Heavy" – Use Modeling Tool / Ongoing Review]

The potential for process integration is best understood with a review of some broader concepts, which will also provide a contextual definition of Capacity Management.

Information Technology Enterprise Resource Management (IT ERM)

"Business" Enterprise Resource Management (ERM) Systems by vendors such as Oracle, PeopleSoft, and SAP AG have provided automated business application solutions that provide a significant amount of data and process level integration to those companies who have successfully implemented them. In a similar manner Information Technology (IT) ERM automated management of operations, performance, capacity, finance, asset inventory, client management, etc. could similarly be designed and implemented with well-defined integrated data and processes.

"Business" ERM vendors cited above, have come much closer to the ideal of data and process automated integration than "Infrastructure" IT ERM level vendor solutions. Mainframe platforms are more mature than client/server distributed platforms in terms of automated IT ERM solutions, but may still lack significant data and process integration -- although it has been a stated objective of most vendors for years. In this paper, Information Technology Enterprise Resource Management (IT ERM) refers to such an ideal solution.[4]

In a highly sophisticated IT ERM implementation, a wealth of additional cost, transaction volume, and quality of service information would improve IT automated management services and information reporting about workloads and services for business customers. At a lower level of sophistication, there is more reliance on manual interfaces, limited data, and non-integrated automated processes so much less meaningful information could be acquired.

The organization of an IT ERM implementation would include the following underlying major process categories:

O Operations Management

Real Time Monitoring of System Issues
Reactive Mode of Operation
Exception Alerts
Problem Identification / Resolution

O Performance Management

Near Term Monitoring System Statistics
Both Reactive and Proactive
System Tuning / Configuration
Workload Management
Service Level Focused

O Capacity Management

Long Term Oriented
Pro-active
Resource Requirement Focus
Forecast Workloads
Predict Capacity

O Investment / Financial Management

Budgeting / Tracking Costs
Planning Expenditures
Recovering Costs
Reporting Costs

O Asset Management

Maintain adequate Asset Inventories
Automate Software Distribution
Maintain Hardware Maintenance
Provide Graphical Location References

O Security Management

Provide Secure Application Access
Authenticate Users
Provide System Recovery Management
Provide Data Recovery Management
Provide Facilities Security

O Customer Partnering

Provide Adequate Service Levels
Understand IT Business Plans
Promote Business IT Strategies / Solutions
Report Resource Usage and Cost Trends
Provide System User Inventory

This paper will only focus on the Capacity Management component of IT ERM as described above, understanding that the opportunities for data and process integration would expand in a similar fashion for any similar analysis of the other IT ERM categories. At the lower of levels of Capacity Management sophistication, data and process integration with the other IT ERM components is limited or non-existent. As all components reach higher levels of sophistication, then the potential for data and process integration would significantly improve.

Detailed Descriptions of CM Levels

Level 0 No Formal Ongoing CM Program – CM occurs as an Occasional Project.

At this level, CM is not considered an ongoing process. However, utilization levels may be informally tracked and when they reach certain thresholds, when service level issues occur, or when management simply realizes the necessity due to new implementation plans, a capacity planning project may be initiated. This is the level used most often in the IT industry. A "snapshot" of the current environment is often studied, forecasts of future requirements are included, and resource predictions are conducted. Some of the more sophisticated techniques discussed in the more sophisticated levels of this conceptual framework may be used on a project basis. On the other hand, the method may be as unsophisticated as an educated guess.

The main distinction of Level 0 and the other levels is that there is no formal ongoing CM program. Data and statistics used for other components, particularly performance, may be used although they may not be designed specifically for Capacity Management purposes.

Level 1 Formally Measure, Trend, & Forecast Peak Period Utilization and Plan Resource Capacity with an Ongoing Period Review Program.

At this level, CM is an ongoing process based on trending resource utilization. Part of the Level 1 process is a regular review of the plans by a multi-disciplined Capacity Management team for each major business, both technical experts and IT business planners. The technical experts should be able to provide supplemental data, such as workload information, the timing of events, and help perform analyses to explain unexpected variances and develop forecast estimates. The IT business planners can provide early warning about future planned changes to the environment that may affect capacity requirements. The periodic meeting, usually limited to a significant related group of applications for a specific company or a company division, is held with staff from various IT departments, and if necessary, capacity plans are updated after acquiring the necessary management approvals and budget adjustments.

The frequency of the review of the charts should be based on unexpected variances in utilization. A monthly review is a good starting point, but often it may evolve to be less frequent. A supplemental approach is to establish a frequency and call a special meeting if unexpected significant changes are reported. Another alternative is to plan short sr. management reviews of updated charts as agenda items during ongoing monthly status meetings, with in-depth staff reviews planned less frequently.[5]

Typically, a capacity analyst prior to the formal CM review meeting issues a memo to specify action items to be reviewed and identify unusual variances that have occurred since the last meeting and issues as identified by sr. management. During the CM team meeting the capacity analyst, facilitates the discussions and manages the agenda to focus only on the most relevant servers and develop a list of action items. Although it is a working meeting and issues can be discussed briefly, care should be taken by the capacity analyst to assign significant activities as action items and keep the pace deliberate. After the meeting, a post-meeting memo summarizes the results and lists the new action items. Management is notified of any budget and funding issues as part of the memo summary and by action item assignment. Materials may also be produced in a standard fashion for sr. management meetings and updated on web sites.

At this level of sophistication, the capacity analyst can only take responsibility for in-depth forecasts if it is assigned as an action item and if the activity is funded (such as a Level 5 application modeling project, possibly outsourced to a consulting firm). Often, the application support staff or vendor will assume responsibility for application forecasts, since they have the most knowledge and information about the application. The capacity-planning staff, usually with limited staff resources, may assume a facilitating or consulting role.

Figure 1 represents an example of a Level 1 CPU capacity plan. It includes a chart with line plots of actual resource utilization, forecast utilization, and a capacity plan to provide the necessary resource. Significant past and future events are documented and tracked as part of the review process.

A “Level 1” automated forecast system could provide some of the following features [Many would be common to a “Level 3” forecast system as well.] –

1) Notice in Figure 1 that a benchmark value is used for the vertical axis rather than CPU utilization. By normalizing the CPU utilization values to benchmark equivalents, the trend on the chart remains relatively accurate after an upgrade. The current benchmark type and full capacity value should be identified on the chart.

2) The system should allow for a customized peak period definition per server. A standard “Prime” period was chosen for the server in Figure 1. Based on an hourly utilization analysis, like the sample chart in the lower part of the figure, a custom period may need to be defined to best represent the actual peak processing. The forecast statistic chosen for the peak period was the maximum weekly average for the month.

3) The automated forecast system would include the capability to develop a regression line based on selected actual data points, which default to the last 12 months. The regression line in Figure 1 would probably align better with the forecast if the data points prior to 2/00 were removed from the regression.

4) The forecast can reflect seasonal adjustments. Several different algorithms could be available and an ability to scale or override the seasonal adjustments should be available. A seasonal adjustment option is illustrated as a 2nd forecast line in Figure 1.

5) A user interface is available to adjust the forecast by month, specifying changes in the forecast reflecting significant events, specify relative changes by month either by slope, relative growth, or by percent change. The example in Figure 1 shows a significant event in February, 2000, will increase workload demand and capacity requirements significantly.

6) A narrative is stored in the forecast database describing significant business issues or rationalizing the forecast. In the example the narrative is displayed under the chart as “Server Remarks”.

7) The system should also provide for an upgrade criteria specification. The specification can adjust for the fact that the peaks vary widely from the average and also be based on rules of thumb for workload types. For instance, a lower value would normally be used for a server with on-line transactions, rather than a server with predominately batch transactions. Of course a very sophisticated modeling tool (Level 4) could be used to help specify it, or it could be based on a rule of thumb, experience over time, and/or other factors known to the CM team.

8) Both past and future significant events and causes of significant forecast variances should be maintained and identified as annotations on the planning chart. This way a simple record of past events can be correlated with the changes they had on resource usage, and future events identify the significant changes in the forecast estimates on the planning chart. This capability increases the opportunity to apply lessons learned from past events to similar future events, particularly when the forecasts are rather unsophisticated.

9) In distributed client/server environments there can be hundreds of servers. The system should provide reporting to indicate which servers had the most significant variances from the planned forecasts. This feature is a bit more rigorous than features typical in most vendor products that simply provide a threshold value without a relationship to a planned forecast value.

10) Optionally, the forecast should be able to adjust automatically based on new monthly actual data. This coupled with the exception reporting mentioned above, may eliminate much of the manual effort and cost associated with forecasting & reviews.

11) It should be possible to save certain forecasts for reference. Before forecasts are changed automatically by monthly actual data point updates, they would be saved for a specified number of generations. It should also be possible to save a forecast as a special category, not to expire until a specified date or until deleted.

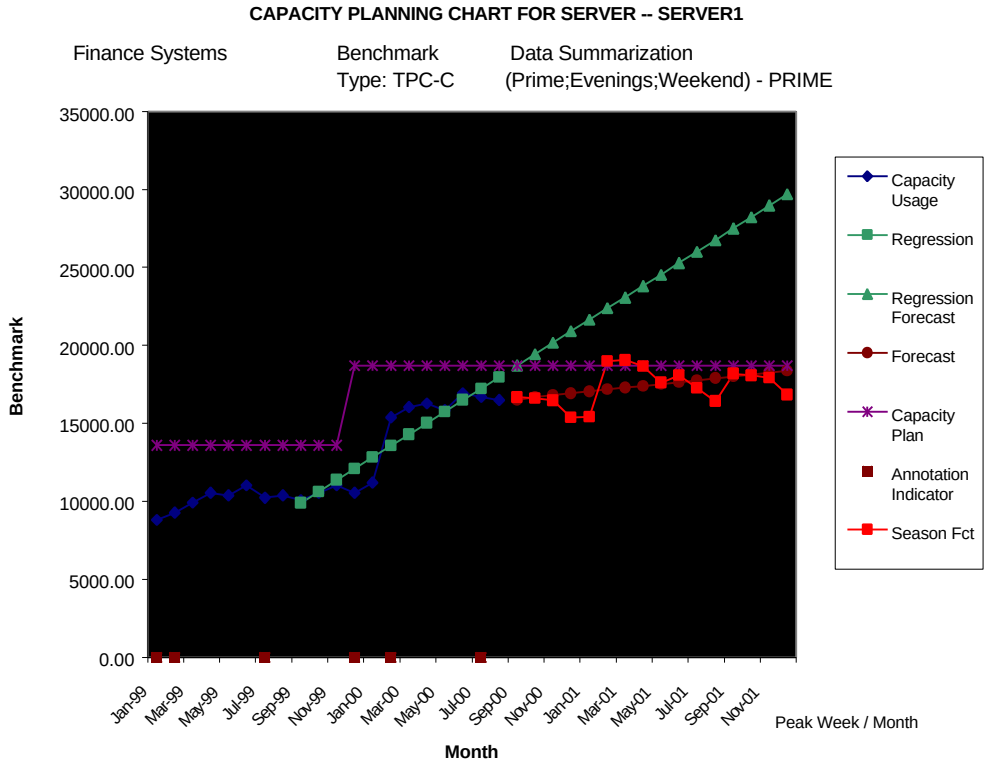
12) The forecast system would consist of a forecast component and a user information component. The user information component should be web based and provide a summary and drill-down capability about the actual utilization, the utilization forecasts, and capacity plans. The forecast component would preferably be PC based so the analyst could use the system during CM team reviews to present updated information, suggest changes in the forecast, and provide alternative suggestions dynamically during the team meeting.

13) Options should be available to use a monthly, weekly, or the monthly peak weak as the peak period statistic.

Other resources such as memory, disk storage, and network adapters could be included in a CM program. Capacity planning resource charts for memory and network adapters could be handled in a

similar manner to CPU except the criterion would be some selected paging or memory usage statistic for memory and a the rate of peak data traffic for network adapters.

A bit more complicated approach could be used for tracking and planning for disk storage requirements. The sample below in Figure 2 uses gigabyte as the unit of resource usage. The problem is more complicated for storage as there may be many layers of "virtual" versus "actual" and "allocated" versus "used" regarding the storage measurements. Different operating systems use different categorizations. So each environment has to be analyzed separately for available / appropriate measurements. In some cases the measurements may require the use of a data base monitor.



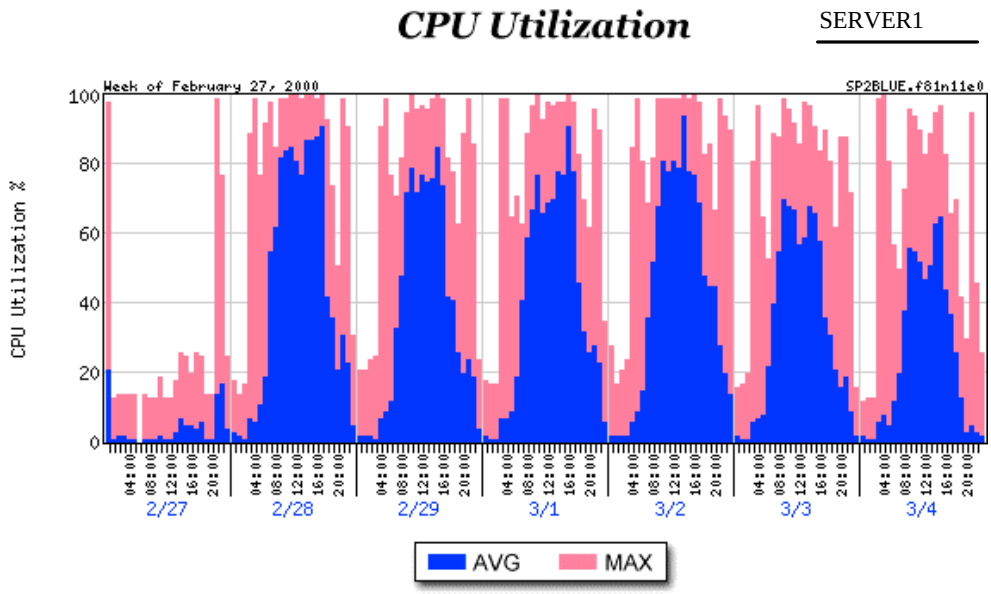
1/99-11/99: NCR 5100 - 8W Pentium II 200MHz.; 12/99+: NCR 5200 - 4W Pentium II 450MHz

Upgrade Criteria is 85%. Current Benchmark Rating is 22000.

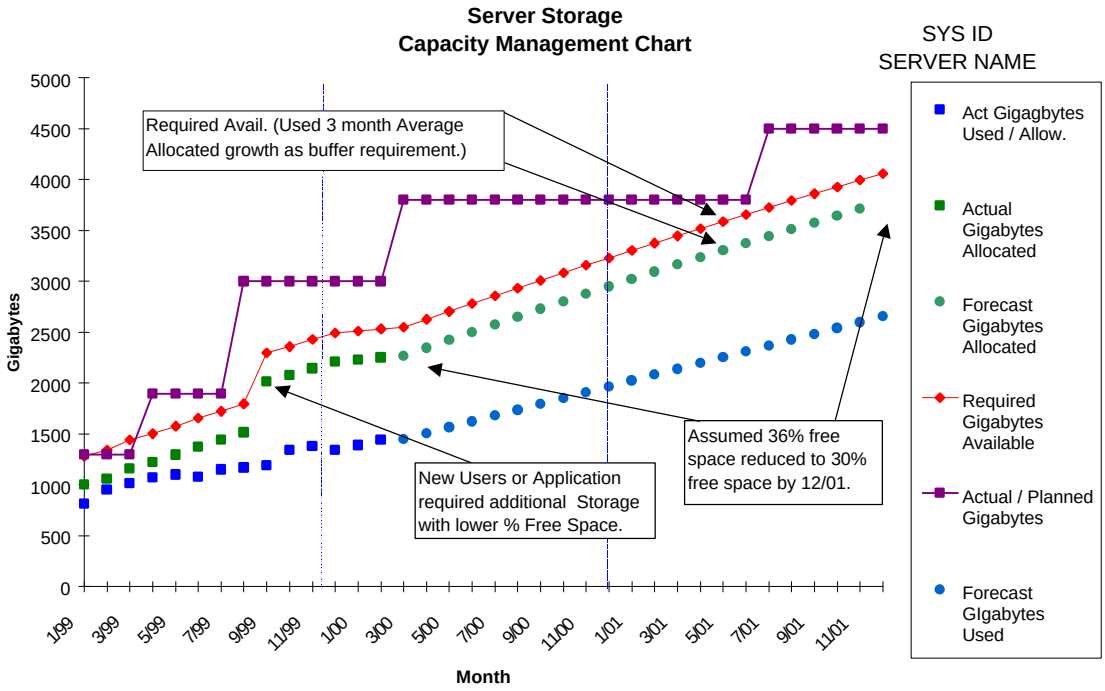
Server Remarks

12/99 Upgrade for Human Resource System Implementation on 2/00 planned to last through 12/00.

Month	Annotations
7/1/00	7/1/00 Added 2 tape drives
2/1/00	Added new HR Application & 200 new users
12/1/99	Upgraded Faster Processors
7/1/99	G/L / PO new Releases Implemented
2/1/99	2/99 G/L Fix Implemented
1/1/99	1/99 Upgraded G/L Reporting System



CPU "Level 1" CM Plan
Figure 1



Storage "Level 1" CM Plan
Figure 2

But for any environment the problem can be generically described as in Figure 2, with a "used" measurement that determines the growth trend and an "allocated" measurement that determines the amount of "free" space retained in the allocated space. Other adjustments may be added to account for reserved storage, other storage overhead, or to allow for time to plan an upgrade.

Whatever criteria are used, they should be well documented for each server, as each may require a different approach or different measurement. Of course it may be possible to adopt some space allocation con-

ventions and policies so the approach is consistent across common types of servers.

Level 2 Accurately partition resource utilization by meaningful workloads. Integrate data with other IT ERM Components.

If process or transaction measurements are available, analysis of resource usage can be substantially improved, particularly if meaningful criteria can be collected with the transaction measurements and translated to even more meaningful attributes.

Workload measurements may be "characterized" by summarizing the data using multiple attributes including company/department, application, and process type categories. A criterion is selected for each resource process or sub-type to assign the usage of the resource to an attribute. See Figure 3 for an example of a possible SAP R3 CPU workload characterization attribute matrix[6]. Notice that criteria captured with transaction measurements such as job name or transaction name may be used and translated to more general criteria such as a specific application.

Naming standards may imbed attributes codes in the criteria values. More often translation tables may need to be derived to map the criteria to the appropriate attribute, particularly for a "Heavy" Level 2 characterization. For instance, there may already be a mapping of user id to company / department code in a human resources or security system which can be accessed. As the data is collected and loaded into a measurement data repository attribute values are derived from the measurement criteria and then may be used as 'keys' in the data repository summarization scheme (See Figures 3 & 4). A 2-tier logic structure often may be used. Both logic tiers may use a translation table and / or logic based on naming conventions or known relationships between the criteria and attributes. The first tier of logic checks for known exceptions to the naming standards. The second tier of logic checks for standard translations. If a measurement value can't be translated a "measurement workload characterization" exception is generated resulting in an exception report so the new values can be updated in the appropriate translation tables. Of course a change or system update process should exist to prevent the exceptions from occurring in the first place when new applications are added to the system.

Workload Resource Usage Measurement Characterization

SAP R3 Workload Characterization Planning Matrix

Resource	Processing Type / Resource Sub-Type	Attribute Criteria" for Company/Dept	Attribute Criteria For Application
Application Server CPU	Production / On-Request Batch	Job Name	Job / Tran / Program / Pgm Std/
	Prod. Batch Daughter Jobs	Job Name	Job / Tran /Program / Pgm Std/
	User Batch	User ID	Tran -Program / Pgm Std
	User Batch - Daughter	User ID	Tran -Program / Pgm Std
	Dialog / Update	User ID	Program Name / TCODE

Sample Level 2 "Heavy" Workload Characterization Chart

Figure 3

Sample SAP R3 Attribute Data Repository Matrix

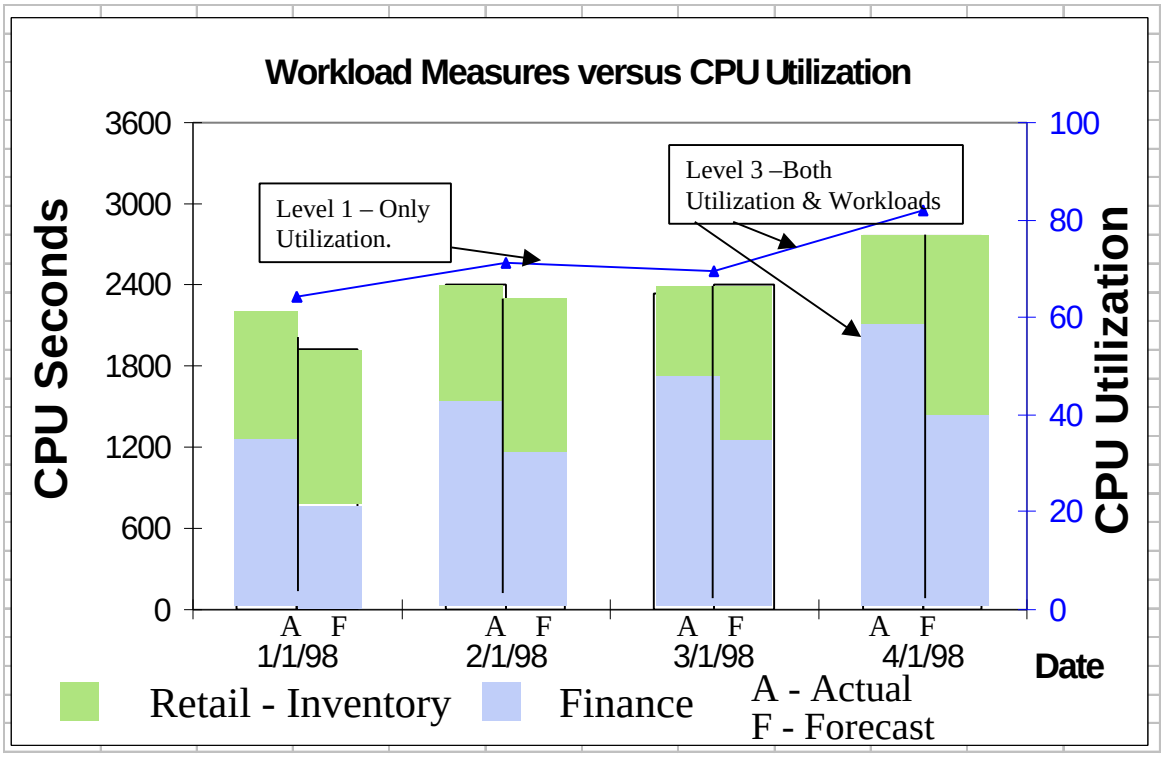
Retention Time	Retention Type	Attributes					
7 Days	Detail	Sap System	Server	SAP Instance	Period	Location	Company Dept / Code
		Process Type	User ID	Application	TCODE	Job name	Program Name
		Date	Hour				
10 Days	Day	Sap System	Server	SAP Instance	Period	Location	Company Dept / Code
		User ID	Application	TCODE	Job Name	Process Type	
		Date	Day	Week of Year	Month		
14 Weeks	Week	Sap System	Server	SAP Instance	Period	Process Type	Company Dept / Code
		User ID	Application	TCODE	Job Name		
		Year	Month	Week of Year			
25 Months	Week	Sap System	Server	Period	Process Type	Application	Company Dept / Code
		Year	Month	Week of Year			

Sample Level 2 “Heavy” Workload Attribute Data Repository Matrix
Figure 4

For the greatest effect, workload data is stored in a data repository and summarized by day, week, and month to provide both detailed and summarized information. Notice in Figure 4, which suggests a possible SAP R/3 data repository attribute summary and retention plan, how fewer attributes are retained for longer time periods so that the data repository isn’t excessively large.

The data repository is designed for Capacity Management and other IT ERM components to provide for data integration. For instance, CPU seconds could be chosen as the CPU measurement criteria and monthly period trends could be available using different combinations and granulation of multiple attributes including peak, prime/non-prime, Company/Department, application, sub-application component, and process type. Data integration may exist between the IT ERM components – the same data summarized in different ways may be used by different system components.

Simple Illustration on How Workload Measurements Improve Forecast Analysis
Figure 5



- Level 2 "Light"

A "Light" Level 2 implementation would consist of a single level workload categorization for the server resource. Examples for CPU usage include RMF Performance Groups in MVS and process level summarizations in UNIX.

- Level 2 "Heavy"

A "Heavy" Level 2 would consist of a multiple level workload categorization for the server resource. This categorization would preferably include multiple attributes such as peak period, company/ department, process type or resource type, application, and possibly other attributes depending on the platform and the resource. Generally the multiple levels of attributes would provide for all IT ERM components to use the same data, usually summarized in different ways in the same data repository for different purposes. For instance the same data used for charge-back could be used for CPU forecasting and performance analysis. Company/ department, application, and process type attributes are required to establish resource levels and service levels separately by customer and application. Without workload usage data characterized at this level of granularity it much more difficult to accurately correlate business application information with resource usage and establish meaningful baselines for forecasting business applications.

- Improved Analysis Capabilities

Often not enough emphasis is placed on developing "well-characterized" workload data from measurement data. Figure 5 illustrates that reviewing how utilization forecast matches up to the actual resource usage is much less meaningful without supporting workload data. In the illustration it would appear without the workload data that the forecast is on "target". However, reviewing the workloads the Finance System is significantly over forecast and the Retail System has an offsetting under forecast usage.

Trending and forecasting these workloads separately as shown in figure 6 would illustrate the variances even better. Had the total utilization had a significant variance in figure 5, the workload data would immediately identify what workloads caused the variances.

So although it may take a significant investment to initially develop a "Heavy" Level 2 approach, once it is accomplished the ongoing cost is rather minimal and the resulting information available for analysis significantly reduces the amount of time needed to determine which user and application is causing forecast variances. This

significantly may save staff time spent and use it more efficiently since detailed reports could immediately identify particular jobs and transactions or users who contribute to variances. But without Level 2 measurements and reports, the team may be reduced to determining relationships by spending much time analyzing the timing of events in detail and then only guessing what caused the variances. This takes time and resources much better suited to be used with Level 2 reports to analyze why the particular applications had variances in more detail.

- Unit of Work / Transaction Definition

Transaction measurements may occur at different levels of granularity, but generally one settles with the type that is available for the application or operating system in use. A standard does exist called ARM[7] to support the generation of workload transactions, but not all applications support it.

Transactions defined at different process levels may have significant differences in granularity and be useful for different purposes. For instance using a telephone call center application as an example, a support telephone call could be considered a business transaction and be very meaningful to a capacity planner. It could also be business level criteria that are used by the Call Center Supervisor in evaluating support representatives and their average call volume, call turn-around time, etc.

The "business" transaction could consist of several "application" transactions – check an account balance, check a particular invoice, make a price quote, check inventory availability, and take an order. Each of these in turn could consist of several "program" transactions, which consist in turn of a series of "data base commit" transactions and also many "screen or dialog" response time transactions. The low level very granular screen transactions are what would be considered important by a performance analyst in evaluating if the IT service levels were being achieved or not. They would also be important to the capacity planner in modeling capacity requirements. But the higher-level units of work would also be very useful for correlating resource requirements with business criteria volume usage.

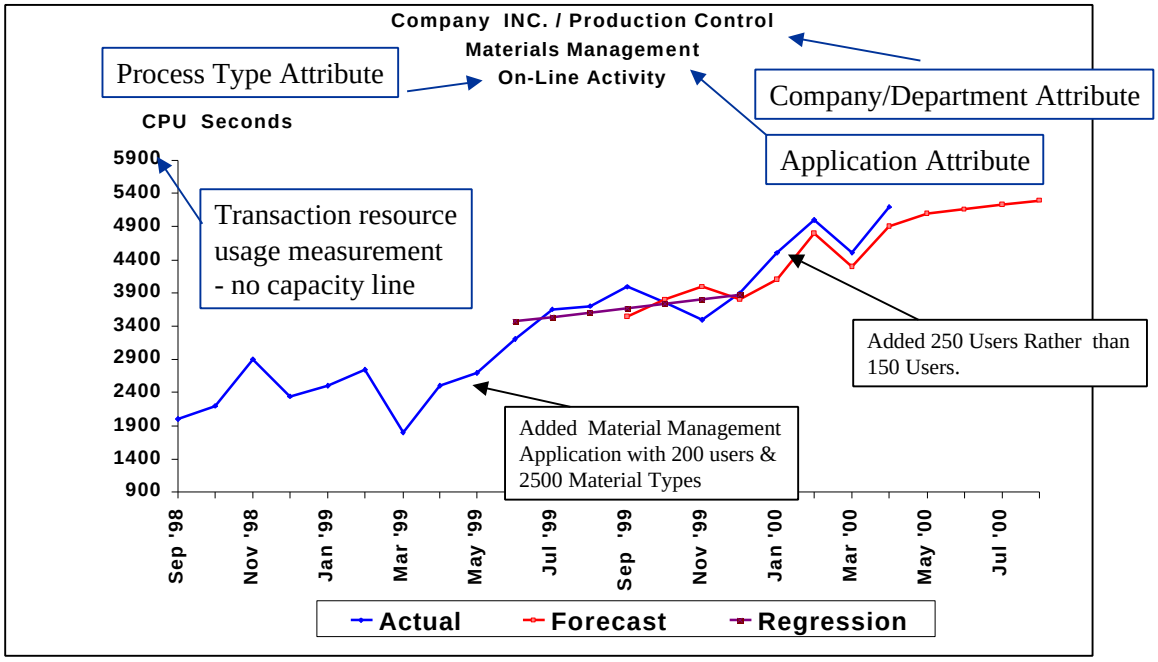
So the concept of "transaction" or "unit of work" should not be considered monolithic and is more complicated than generally realized. For instance, if transactions are normally collected at a very low level of unit of work a great amount of data may be unnecessarily generated creating a lot of unnecessary overhead. One of the most popular Business UNIX ERM systems collects data at the dialog-step or interactive-screen level creating much more overhead than would otherwise be necessary.[8] A preferred approach would be to collect summarized accumulation statistics at a much higher level of "unit of work" such as the application transaction (unit of work) level, but provide for the lower level transaction collection as a monitor "trace" option, as the detail would only really be useful when debugging a performance or program problem.

Level 3 Include an automated workload forecast system.

At Level 3, the workload data serves as input to a trend-oriented forecast system. Then in Level 4, the separate forecasts may be summed and combined after adjusting for measurement capture ratios to estimate total resource requirements which achieve adequate service levels. The system should provide for automated default forecasts extending historical actual trend data into the future.

Workload forecasts do not include upgrade criteria since each workload is considered separately and may (optionally) include adjustments to accurately approximate raw resource utilization. But the forecasts in total may be used in Level 4 as input to a modeling tool or other methodology to estimate forecast utilization, upgrade requirements, and possibly service levels by workload. A level 3 forecast system will include a capability to support follow-up month-to-month variance analysis and forecast revisions with the capability to retain the initial budget based forecasts. A "default" forecast should be optionally generated for all workloads using either an average or a regression statistic.

"Level 3" Workload Forecast Chart



Sample "Heavy" Level 3 Workload Forecast Chart
Figure 6

- Level 3 "Light"

A "Light" Level 3 forecast system implementation would consist of developing forecasts for a single level workload categorization for the server resource to be forecast. Examples for CPU usage include RMF performance groups in MVS and process level summarizations in UNIX.

- Level 3 "Heavy"

A "Heavy" Level 3 forecast system would consist of a multiple level workload categorization for the server resource. This would have to include peak period, company/ department, process type or resource subtype, and application attributes among others depending on the platform and the resource. The forecast system would allow for forecasts input at high levels of summarization (using only one or two attributes) to be automatically distributed to all "baseline" attribute levels using alternative algorithms. One algorithm would be based on splitting the forecasts based on fractional percentages calculated from the actual historic monthly usage percentages using the other attribute value combinations by month. Then high-usage workloads could be forecast using company, department, application, and process type attributes. The mid-usage workloads could be forecast using the company/ department and application attributes letting the process type forecast be less accurately automatically generated using historical statistics. The default forecast would typically be used for low-usage workloads.

Once finished developing forecasts by company/ department/ application categories using business oriented input, the forecast could be automatically re-summarized for only the peak period using a single operational attribute (such as RMF performance group for MVS) to be used as input for configuration modeling tools (level 4). Or, more granular levels of Company organizational categorization could be used for charge-back purposes – company/ department/ division / work team even though the original forecasts were accomplished at a higher level of summarization!

Level 4 Predict service levels & capacity requirements using (level 3) workload forecasts.

At Level 4 processes exist to provide a means to combine the workload forecasts and model them to determine the proper capacity requirement to minimize costs and

yet obtain service levels. This could be with a very sophisticated modeling tool or simply by using industry rules of thumb. At Level 1, the upgrade criteria is used rather than a more accurate configuration-modeling tool.

Using workload forecasts from Level 3, an automated process is used to estimate capacity requirements necessary to satisfy service levels. For CPU, detailed data for resource usage by process type may be used in a multiple linear regression to calculate capture ratios for each process type. Then the resource could be optionally translated to units of utilization or benchmark values rather than CPU seconds.

Process integration as well as data integration may occur at Level 4 with the other IT ERM components. Predicting capacity could be integrated with business cycle budgeting, performance management, and charge-back processes. For example,

1) Simulation or analytical modeling tools are used for performance tuning;
2) Forecasts are useful for cost recovery;
3) Yearly financial planning and client budgeting processes could be conducted in conjunction with a yearly capacity forecast where optionally the following year's charge-back could be fixed based on the capacity forecast, rather than depend on the actual usage results. Although policy would have to allow for unexpected large variances, in general the forecast accuracy would allow this to occur. Also unexpected large variances could more likely be traced to variances in business application volume assumptions, rather than assumed to be estimation error -- placing the responsibility for cost variances back to the client.

- Level 4 "Light"

A "Light" Level 4 implementation could be as simple as applying measurement capture-ratios based on process type to the workload forecasts, combining the forecasts to estimate future utilization, and using 'rule of thumb' criteria or past experience for capacity upgrade requirements. Usually the "light" level 2 workloads would not be suitable for cost recovery or client charge-back unless they naturally break out by customer and application.

- Level 4 "Heavy"

A "Heavy" Level 4 approach would be more sophisticated using workload resource forecasts as input to a configuration modeling tool to provide estimated service levels and capacity requirements, possibly using a what-if iterative analysis. In addition a "Heavy" Level 2 workload characterization would provide for client cost recovery or charge-back and other types of process integration with performance, financial, and client management.

Level 5 Use business application criteria with an application model to predict service levels and forecast resource usage requirements of the most resource intensive client applications.

Relating business criteria to specific company/ department application components increases the sophistication and accuracy of a workload forecast.

For the most resource intensive existing applications and particularly for new applications with no information regarding baseline resource requirements an application workload requirement estimate methodology would be used to translate known application business volume criteria to workload usage demand estimates. Benchmarks may be used as the sole methodology, to derive baseline usage data, or as verification of an application workload estimation method. Vendor application modeling tools may be utilized.

Workload data available from Level 2 may also be utilized to baseline the model or to approximate similar applications. The output from the application-modeling tool may be use as input to the Level 4 configuration-modeling tool, if the application shares resources with other applications. With more emphasis towards web enable E-commerce service levels, application modeling may become more important for business clients.

- Level 5 "Light"'

Actual business criteria are not tracked and reported as part of the program. On a project or action item basis, simple correlations are analyzed and proportional relationships derived between key business criteria volumes and resource usage.

Summary of Six Levels of Capacity Management Sophistication

Level	Description	Focus / Objective
Level 0	No Ongoing Program	Informal / Project Approach / Reactive
Level 1	Forecast Utilization Trend	Insure Capacity Availability
Level 2	Interpret & Manage Workloads	Workload Definition & Data Integration
Level 3	Forecast Workloads	Trend Workload Forecasts
Level 4	Predict Resource Requirements / Service Levels	Predict Requirements / Integrate System Management Processes
Level 5	Forecast Business Applications	Base Forecast on Business Information

Figure 7

- Level 5 "Heavy"

Actual business company / department, application criteria are tracked and correlated with resource usage as predicted by a formal modeling method or tool. The model may be automated and may be continually refined.

Using The Framework as an Audit Tool

The six levels should really be defined differently when viewed as audit standards rather than as different sets of expanding CM processes. This paper has focused on a definition consisting of expanding sets of automated processes (See Figure 7), and leaves the audit definition to your intuition; realizing that the levels themselves are never mutually exclusive, since all the levels represent common CM sub-processes, only accomplishing the processes at different degrees of detail, sophistication, data/ process integration and accuracy. Perhaps several audit ratings would be applicable to any environment – a) what is technically achievable, b) what level is most cost effective, and c) what level is actually achieved.

Conclusion

From the author's experience, most vendor offerings automate processes related to the more sophisticated levels such as 4 & 5. Less vendor capabilities exist at Level 1, 2, & 3, perhaps because significant customization may be required to fully develop well-characterized workload measurements and vendors prefer to provide turnkey solutions. Professional capacity planners tend to develop their own custom solutions. The lack of well-characterized transaction measurements based on appropriate units of work in most IT environments and limitations in vendor software integration at all levels tend to limit the ability to accurately and cost-effectively implement high-level Capacity Management and related IT ERM programs. But it is also the author's opinion that with sufficient improvements in IT industry measurements, standard practices, and vendor software, then more, higher-level, cost-effective CM programs could exist in the future.[9]

References

[GTJR99] Bob Gutjahr, "Tutorial on an Integrated Performance & Capacity Management Process", CMG 99
[CHNY99] Bob Chaney, "The Capacity Performance Council – Start Yours Today!", CMG 99
[THOM96-1] George I. Thompson, "A Manager's Framework for Enterprise Resource Management", CMG 96
[THOM96-2] George I. Thompson, "The Need for an Enterprise Resource Management / Forecasting Infrastructure", CMG 96
[THOM97] G. Thompson, J. Munoz, K. DeBruhl, "The Availability & Quality of SAP R3 Workload Data for Performance / Capacity Management Requirements", CMG 97

[1] As examples see references [GTJR99],[THOM96-1],[THOM96-2],[CHNY99]
[2] This is supported anecdotally by many CMG papers – For Instance reference [CHNY99]
[3] See following section and references for explanation of "IT ERM".
[4] Specific suggestions regarding automated IT ERM requirements are presented in more detail in the references [THOM96-1], [THOM96-2]
[5] Some large Corporations may have a centralized PM/CM Group. Most relevant is having a Sr. Management commitment to an ongoing Capacity Management Program. See Ref. [CHNY99]
[6] See Reference [THOM96-2] for a Mainframe Example
[7] See Computer Measurement Group Web page about ARM: http://www.cmg.org/regions/cmgarmw/
[8] See Reference [THOM, 97]
[9] See Reference [CHNY99] for an example of an effort to improve capacity planning processes at

an enterprise wide level across platforms on an ongoing basis. Also see end-notes [4] & [8].