A Roadmap to Success in Capacity Management

or Why Real-time Monitors are a Waste of time!

White Paper

This paper is based on the Author's experiences in assisting many larger organizations in the establishment of formal Capacity Management (which incorporates the disciplines of Performance Management and Capacity Planning). The paper outlines the objectives of Capacity Management, tells you what you need and what you don't need in order to move forward, describes things that will help and things that will get in your way, and concludes with a roadmap to success.

1 Introduction

This paper outlines the steps that you will need to follow in order to establish formal Capacity Management in your organization. In this context I am using the term 'Capacity Management' to encompass the disciplines of both Performance Management and Capacity Planning.

Section 2 proposes a simple definition of the objectives of the Capacity Management function. Section 3 lists the things that you will need and section 4 the things that (despite common use!) you don't need. Section 5 describes the things that will make things worse for you, and section 6 those things that make things better. The paper concludes with a roadmap to success and some final thoughts and conclusions.

2 Objectives

Over the years there have been many alternative definitions for the objectives of the Capacity Management function within the organization. One of the oldest (and still the best in my humble opinion) is as follows: -

The continuing provision of consistent, acceptable service levels,
at a known and controlled cost.

Let's look in more detail at some of the key phrases in what is seemingly a fairly simple and straightforward definition.

(a) 'Acceptable service levels'

- Given a requirement to support a given workload at a given service level, derive the resources required
- Given a set of resources and a workload level, define the service levels that will be provided
- Given a set of resources and a service level requirement, define the workload levels that can be supported

(b) 'Continuing provision'

This means that it is not sufficient merely to ensure that current performance is satisfactory. A formal and rigorous approach to capacity planning will also be required in order to determine the useful life of the current configuration. This is defined by its ability to support the expected workload and achieve the required service levels. This, of course, requires a clear definition of the workload that the system will be required to support.

(c) 'At a known and controlled cost'

This means that a mechanism must be provided which will define the resources required to support a given workload at a given service level.

3 What you need

Embarking upon the road to formal Capacity Management can be a rather daunting prospect, particularly since the discipline is as much about solving tomorrow's problems, as it is about solving today's. Most senior managers, when they consider the long-term benefits of Capacity Management agree that it is the Right Thing to do. Nevertheless, there are always short-term demands on budgets ("We need to buy that new SAN box right now.") that address today's problems, and will compete for budget allocations. Only the confirmed commitment and support of senior management will ensure that the longer-term issues are given sufficient weight.

Things that you will need to move towards a Capacity Management capability are: -

- A clear set of objectives
- Senior management commitment
- Process/flow definition
- A realistic plan
- The right people

This means that a definition of the required service level (defined in terms of the response times) must be agreed and monitored. It should be borne in mind that such a Service Level Agreement (SLA) is not a 'stick to beat the IS department' with. An SLA defines the obligations both of the provider and of the consumer of information services. It is astonishing how many organizations, without informing their IT department, expect to increase their workloads (often dramatically) and yet expect their current systems to cope.

Capacity Management is all about managing the relationships between three inter-connected variables - resources, workload and service levels, as shown in the diagram below.

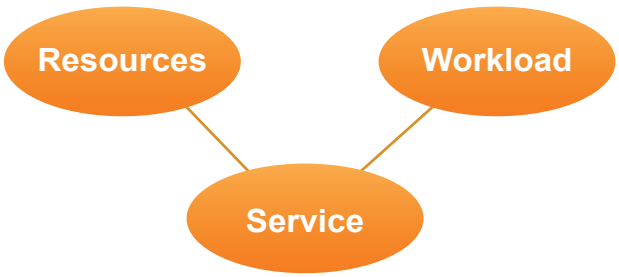


Figure 1. System Variables

It is impossible to change any one of these elements without affecting at least one of the other two. The Capacity Management team must be able to take any pair of variables and derive the third i.e.: -

- The right toolset

The roadmap outlined in section seven of this paper will expand on these topics.

4 What you don't need

"Capacity Managers don't need pagers!"

4.1 Threshold-based Alerting

A lot of organizations these days are relying for their capacity management on threshold-based alerting systems. Many have sophisticated event-tracking systems and e-mail/pager based alerting. Relying upon event management means that you are told when you already have a problem. This is rather like having a device in your car that tells you when you are having an accident! (Important information of course – but the information is just a little too late.)

Contrast this with mechanisms which track behaviour, provide linear extrapolation where such techniques are applicable, and provide performance prediction based on well-known mathematical principles which take into account the non-linear relationship between loading and response times (e.g. queuing network modelling tools). The use of such techniques means that potential problems can be identified and avoided in advance – being proactive rather than re-active.

4.2 'Dashboards'

A senior IS manager once told me that he needed "to know the CPU utilization at any moment in time". When I told him that his request made no sense, things got quite heated! Utilization figures only make sense when considered over some period of time. The definition of Utilization for a device is its busy time divided by the elapsed time. At any one instant the CPU is either busy, or it isn't. So, there could only be two possible answers to the manager's question – either 0% or 100%!

The elapsed time chosen will depend upon a variety of factors, including the degree of volatility of the required measure and the use to which the measurement will be put. CPU and other device utilizations can vary dramatically when viewed over very short (less than one minute) elapsed times. This is why the use of performance 'dashboards', which update their 'dials' every couple of seconds or so, does not provide any useful information to the Capacity Manager. (Such tools can also soak up very considerable resources on the target system.)

The information required by the Capacity Manager will usually relate to the behaviour of the metric, observed over some period of time, characterized by the statistical distribution, including the mean and 95th percentile. The latter is becoming more and more important for response time analysis, since many service level agreements are now based on 95th percentiles rather than arithmetic means.

The importance of this is illustrated in the graph at Figure 2 below, which shows the distribution of a workload's end-to-end response time in a multi-tier client/server environment.

Figure 2. Response Time Distribution

Note the fact that this is nothing like a normal distribution, and the difference between the mean (at 1.88 seconds) and the 95th Percentile (at 3.84 seconds).

This particular graph is taken from the baseline of a multi-tier planning model. Using this sort of

technique, the capacity planning team can derive not only the fact that the workload will not meet the required service levels if expected (or current) growth rates are maintained, they can indicate the 'choke point' i.e. the system or device responsible.

We should bear in mind that the members of the Capacity Management team do not (or, at least, should not) get involved in tuning, optimising, debugging, tracing, or indeed any number of short-term 'fire fighting' activities. Unfortunately the Capacity Management team is, all too often, expected to deal with day-to-day issues as well. When this happens the short-term issues will usually predominate, to the detriment of the Capacity Management process.

4.3 The 'Red Adair' Syndrome

Figure 3. A burning oil rig.

Over 40 years ago an oilman called Red Adair set up a company that specialised in extinguishing oil rig fires. His organization could dispatch a well-equipped team anywhere in the world, at a moment's notice, to put out these dramatic fires. Unfortunately many IS organizations adopt the same approach for managing their systems.

No doubt the individuals concerned are highly motivated and feel that rushing around putting lids on things is a valuable contribution. Unfortunately this reactive approach is a very expensive way of managing anything. Oil rig fires suddenly start with immediate and cataclysmic consequences. By contrast, most computer-related problems (particularly those related to performance and capacity) gradually develop, and should therefore be spotted and fixed before end-user service levels are affected.

"If you are reacting to incoming user complaints about poor performance,
then you have already failed."

5 Things that make things worse

5.1 The e-commerce myth

This can take a variety of forms, but a typical VP in charge of IT infrastructure when asked about his capacity management approach will respond with something like:

"Well, we're opening up a number of our key systems to web-based access. We have no idea what the traffic rate is going to be, or what our service levels are going to be. We will just have to see how it goes and monitor things carefully."

This is a recipe for disaster. There are well-documented instances throughout the e-commerce world of newly available systems being overwhelmed with requests and simply folding under the pressure. This is guaranteed to lose you customers, credibility and money, all of which are hard to win back!

Capacity Management teams, who are equipped with the required monitoring and modelling tools, can define the workload that can be supported, at the required service levels, by the current resources. Routers and load balancers can be configured to limit the traffic passed on to the web and database servers. This approach maximises and maintains the system's throughput, and enables potential workload tracking that will drive forward investment planning.

5.2 Knee-jerk reactions

Most organizations will willingly invest in IS resources, if they can do two things: -

(a) Clearly justify the expenditure

(b) Budget for it in advance

Nothing is more likely to send the average CFO into orbit quicker than an unbudgeted-for emergency hardware upgrade. Such things can throw the financial planning of even quite large companies into chaos. No time is left for research into competitive offerings, or for sensible negotiation. Meanwhile service level targets are not being met, which will also damage the organization's profitability.

5.3 Hardware suppliers

The quality of the capacity management advice provided by suppliers varies widely, both between and within manufacturers. I have seen everything from soundly-based accurate performance predictions, to pure guesswork, recommending whatever machine the hardware sales rep gets most commission on that week. Putting your investment decisions in the hands of your supplier is unlikely to be the most cost-effective approach!

Beware also the "poor performance means you need more horsepower" approach. This may work, up to a point, in cars, but very often does not for computer systems. A common cause of poor performance is the delay contributed by an over-loaded disk I/O subsystem. Buying yourself a more powerful CPU simply gives you the ability to hit the I/O subsystem harder! Degraded service levels are often the result.

As always, it is important to upgrade the right thing. Consider a (not unusual) system where 80% of

your response times are I/O busy and queuing, and 20% are CPU busy and queuing. If you double the power of your CPU, your response times will not improve by much more than 10%. This fairly obvious result has recently resulted in several IS managers unexpectedly being able to spend more time with their families.

5.4 Disconnected Users

This occurs when there is inadequate (or, quite often, non-existent) communication between the suppliers and consumers of the IS service. I have seen a number of IS organizations who were in blissful ignorance, both of the fact that the users had a very low regard for the quality of the service, and that significant business changes were about to dramatically change the workload that the IS systems would be required to support.

There is a strong argument (backed up by successful implementation at a number of very large organizations) that the Capacity Management function should contribute to, or even be part of, the Quality Assurance function. Delivering systems that do what the users want means looking after current and future performance as well as functionality. A performance bug is just as much of a bug as any other sort! This, of course, implies that the Capacity Management team becomes involved in new applications during the design and development process, ensuring at each stage that production performance targets will be met.

This process is sometimes referred to as Performance Engineering (PE) or Software Performance Engineering (SPE) as originally advocated by Connie Smith. [Smith 1990]

Using SPE techniques, the cost of the resources required to support the application (at the required service level) can be derived early in the development process, and refined as that process continues. This allows swift termination of application developments if the cost of their implementation would outweigh their business benefit.

5.5 Disconnected Modelling Gurus

Not so many years ago the typical capacity planning team was composed of frighteningly clever people, often with impressive beards, who worked away in dark corners, muttering incantations and using strange and complex mathematical modelling tools.

The team would occasionally emerge and pronounce on issues that they considered important, and then retreat back to their ivory tower. Due to the speed with which modern systems and workloads change, this traditional approach is increasingly irrelevant to most organizations' needs for management information.

6 Things that make things better

6.1 Track Performance& Resource Consumption

Make the best possible use of the historical information available to you. Historical information is only of value when it can tell you something about the future. Back to my car analogy. You have a great set of mirrors - so you know where you have been. You can look out of the side window – so you know where you are. But the only really important thing is to look out of the front window - so you can see where you are going!

The Performance Database (PDB) will very quickly become a valuable resource, enabling you to build a clear picture of the relationships between business metrics (more of which in the next section) and resource consumption. The data can also be used for trend analysis, particularly in the area of workload growth and seasonality patterns. This can be fed through into capacity planning.

The capacity planning team can then provide performance forecasts on the basis of workload volumes that, in the absence of any other information, can be expected to continue to follow established patterns. Clearly this will then have to be 'overlaid' with additional information on business changes that will cause the workload to deviate from the historical 'norm'.

6.2 Business Information

Clearly the Capacity Management team cannot operate in a vacuum. Business information, on the ways in which the business will develop and require IS support, is vital if accurate capacity plans are to be produced. One significant problem occurs during this process, that of different units of measure.

Business planners will describe future workloads in terms of Natural Forecast Units (NFUs). These will be numbers of cars, trucks, flight bookings, auction listings, invoices etc., etc. The business planners have no interest in (and couldn't care less about) things like CPU utilization, disk service levels, hub throughput capabilities etc. All they are concerned about is that IS performance is adequate to perform the business function.

The Capacity Management team must be able to translate these business metrics into IS equipment resources, and therefore into $ expenditure figures. Capacity Management teams are increasingly focused on the process of incorporating business statistics directly into performance databases, and using correlation analyses and other statistical techniques to understand the relationships between business volumes and resource consumption.

6.3 User Feedback

Whilst system monitors and logs are vital tools in the pursuit of Capacity Management, one of the best sources of information, particularly for users within your own organization, is to go and talk to them! One day spent in a user department watching the way in which the applications are actually being used, and how they perform in real life, is worth a week of poring over device utilization analyses.

6.4 Relevant/Timely Modeling

Modern monitoring and modeling tools are now available which enable the integration of capacity planning with the business planning function. Models and predictions can now be produced quickly enough to be part of the management decision-making process.

6.5 Intranet Publishing

One of the criticisms that organizations have most frequently expressed about their Capacity Management teams was that their work, although generally acknowledged as beneficial, was not very 'visible'. CM teams are now making very

considerable use of Intranet reporting which, with the advent of entirely automated reporting mechanisms, means that comprehensive and accessible html-based reports can be disseminated with the minimum of effort.

7 The Roadmap

This section is entitled 'The Roadmap' which implies that you start from somewhere, follow the map, and end up at your destination. Whilst a number of steps do have to be taken, the journey to effective Capacity Management is not a simple linear process, rather a more iterative learning experience, where the work done at each step may well cause enhancement, revision, or even replacement of previous work.

The recent 'Roadmap' process to promote peace and stability in the Middle East includes a number of important milestones that are equally applicable here: -

(a) Mutual understanding - so that everyone involved appreciates the needs and views of all participants

(b) Boundary definition – defining who is going to be responsible for what, and how the various parties will integrate

(c) Process definition – defining the required steps and timetable.

Commercial organizations, National Governments, and international relations all react badly to sudden and dramatic change. Revolution can only be justified if the evolutionary process is unnaturally and unreasonably blocked in some way. [Machiavelli 1513] The route towards full adoption of the Capacity Management principles outlined in this paper is an evolutionary process, starting with a small application or service subset, and a simple intranet reporting mechanism.

As the benefits of a formalised Capacity Management approach start to become apparent, it will be increasingly easy to obtain commitment from the various participants who provide data and use the CM team reports. After this initial phase, you can then start to extend the range of services to include SLA management and capacity planning.

The sections below outline the Stairway to Heaven Success.

7.1 A clear set of objectives

A suggestion for an overall objective statement was given earlier: 'The continuing provision of consistent, acceptable service levels at a known and controlled cost.' This objective will need to be extended to incorporate the scope of the Capacity Management team activities, in particular, the range of systems and business processes that they will be required to cover.

7.2 Senior management commitment

This is an obvious, but vital step on the road. Without senior management commitment, the required personnel and tool resources will not be provided, and the organizational changes (particularly in terms of business information flow) will not be made.

7.3 Process/flow definition

This is a definition of the way in which the Capacity Management team will interface with the rest of the organization. It must include a definition of:

- All the in-bound information flows that will be required
- The organizational impact and any required changes
- The ways in which the work done by the team will be directed by the organization's needs
- The ways in which the team's findings will be disseminated and used.

7.4 A realistic plan

The key word here is 'realistic'. Whilst it would be nice to have a completely comprehensive, fully integrated Capacity Management function from day one, it just is not going to happen. The required organizational changes may well take a considerable amount of time to define and implement

7.5 Recruit or retrain the right people

Despite the fact that mathematicians are amongst the most important and worthwhile people on the planet, capacity managers do not have to be mathematicians. Although a certain degree of numeric ability will be required, an ability to communicate with the business is equally important. Such abilities will be required not only to derive the required information from the organization, but also to communicate and present findings and recommendations effectively.

An effective capacity management team member really needs to have a foot on both the business and the technical community. Remember that Capacity Management is a business discipline with technical implications – not the other way round!

7.6 Acquire the right toolset

The precise toolset will depend on individual organization's circumstances, but a minimum starter set will include: -

- Performance and resource consumption monitors for all system components which contribute to the delivered service
- A mechanism for storing that data in a central repository
- A mechanism for importing business data, and any other data that would provide further understanding of the relationship between the business workload (measured in Natural Forecast Units) and IS workloads
- A facility to generate regular reports automatically, across the complete range of business and system data in the repository, and to distribute those reports in an appropriate manner to the target audience (typically using html and the corporate intranet)
- A mechanism to produce ad-hoc reports and to 'drill down' into the data for problem diagnosis
- A trending facility for use on those aspects of the data for which linear extrapolation is a valid technique

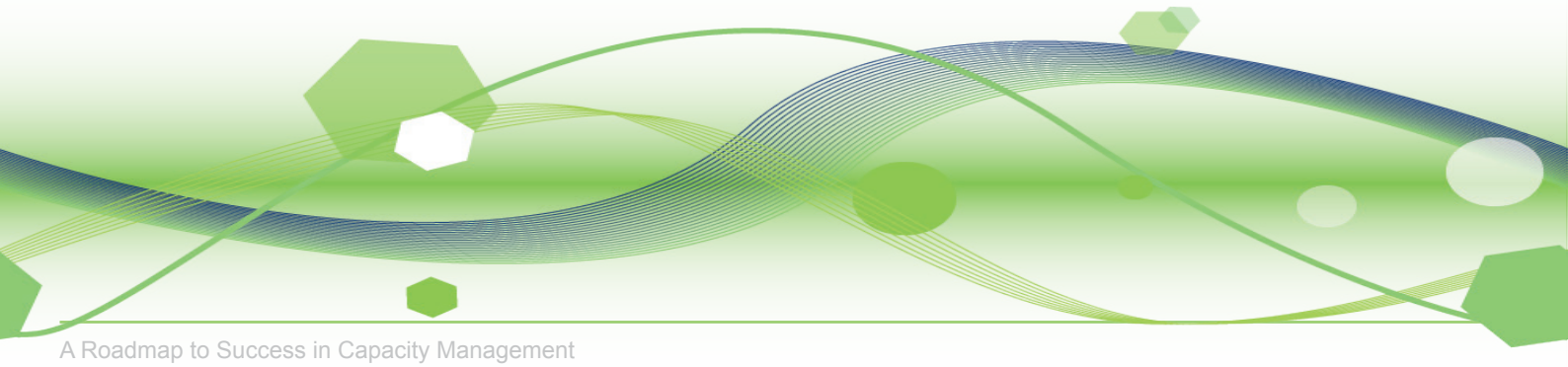
A performance prediction facility that will allow the non-linear behaviour of systems experiencing resource contention to be taken into account. This will usually be based on some form of analytical modelling.

7.7 Walk before you start to run

Avoid the temptation to try and cover too many target systems or applications right from the outset. It will take some time for the organization to adapt to the additional disciplines required for Capacity Management and to start to benefit from its provision

7.8 Iterative evolution

As I mentioned earlier, Capacity Management is and evolving process. Activities and achievements must continually be monitored and checked against the objectives and the plan. This is an ongoing iterative process of refinement and improvement.



8 Conclusion

In order to achieve consistent, acceptable service levels, at a known and controlled cost, you need Capacity Management.

You DO NOT NEED threshold-based alerting, fire-fighters or dashboards, all of which contribute to knee-jerk, event-driven 'management'. Other things to avoid are the 'wait and see' approach (particularly inappropriate for e-commerce systems) disconnected users and modeling gurus, and purchasing decisions based on the recommendations of hardware suppliers.

You DO NEED to track performance and resource consumption, correlate that against business workload measurements, report to (and gain feed back from) your user community through internet publishing of reports at the right level of detail, and relevant, timely modeling to support purchasing decision-making.

No magic or divine intervention is required in order to achieve success in Capacity Management. It is important to avoid the myriad distractions that some eager software tool vendors can strew in your way. Follow the roadmap that I have described, and concentrate on providing (and demonstrating!) your contribution to continuing quality service levels, and to the organization's bottom line.

Bibliography

[Smith 1990] Connie U. Smith, Performance Engineering of Software Systems, Reading, MA, Addison Wesley, 1990.

[Machiavelli 1513] Nicolo Machiavelli, The Prince, 1513.

www.metron-athene.com