The Top 10 Myths of Performance Analysis and Capacity Planning

Adam Grummit

White Paper

1 Introduction

This White Paper reviews Performance Management as well as Performance Engineering, Service Level Planning, Network Planning and Performance Audit. These disciplines are all considered and discussed in terms of the top ten perceived problems in their implementation, defined as dilemmas in terms of “myths” and “realities”.

A large commercial organization may devote 6% - 8% of its total budget to Information Technology (IT). The sheer size of this expenditure emphasises the crucial importance of effective planning. Furthermore the quality of many IT systems whether in the public or private sector impinges directly on our lives.

As IT has developed so have the associated functions of providing the hardware configuration and resource management, systems software environment, control of application software, technical support, network management, and end-user support. This supporting infrastructure is important and it requires formal planning, known as IT Infrastructure Planning (ITIP).

ITIP breaks down into Performance Engineering and Performance Management for both Computers and Networks. Performance Management in turn is split into Performance Analysis (including monitoring, tuning and optimization as well as bottleneck analysis) and Capacity Planning (embracing Capacity Forecasting). Performance Engineering incorporates both software performance prediction (sometimes called Software Performance Engineering or application sizing) and performance testing. ITIP Audit links and monitors the entire process with feedback to control the loop.

The ITIP framework is well established in traditional mainframe environments and third party tools have been developed for most major operating systems to assist in the tasks defined. However, the level and degree of their applicability to distributed systems is less clear.

The distinctions between multi-tier solutions with PC’s running Windows, workstations under UNIX and minis or mainframes under UNIX, LINUX or POSIX-compliant proprietary operating systems are becoming less clear. With the implementation of e-commerce distributed systems with RDBMS, 4GL and GUIs under client/server architecture increases, so finding the appropriate level of ITIP for distributed systems becomes more important.

The core problem is that many attitudes are based on the perception that it is more effective to just add more low-cost boxes than to create a proper plan. Experience suggests that this is fallacious: the true cost of a simple upgrade is not very low and the cost of many spread over a network is large. In these days of trying to get more from less, the hard financial case tends to be the key issue, cost-benefit studies need to be applied to distributed systems as to any other corporate investment.

Clearly, the existence of well defined methods and tools demonstrates that PM is an active, recognized discipline. The relationship between capacity planning and the need for software performance engineering to size new applications is recognized. The requirement for an effective interface between capacity planners and application designers is documented. The practical problems in establishing a performance measurement regime and the difficulties of performance trials are well known .

Cost-effective planning of IT Infrastructure is thus seen as a “Good Thing”. Yet proven cost-benefit justifications seem hard to find. Why is this so? Probably because of the potential damage to an organization publicly admitting to disasters or bad practice.

Perhaps there is also an effect due to the “insurer’s dilemma”. Establishing the conditions where you can introduce a good insurance policy, may well lead to the avoidance of the worst of the risks involved. But that does not invalidate the policy, it merely indicates how many risks are easily circumnavigated.

The exercise of PM requires both a high degree of technical expertise and also an awareness of the business needs and objectives of the organization. Thus the performance analyst is not merely a systems software specialist but also a business analyst with the necessary skills to liaise across the boundaries of IT and management.

The level of detail required to carry out successful planning may well be different from that required to solve a performance problem.

Performance Management is the computer-assisted control of all aspects of the performance of a computer system. Along with other functions, such as resource management, event management and security management, it forms a key part of System Management.

The performance analyst may spend a certain amount of time tuning a system to overcome problems. Equally, rogue modules in application software can be optimized to avoid unnecessary resource demands. But there comes a point where these activities become self-defeating, with diminishing returns because the system itself is saturated. The performance analyst's role is to evaluate hardware configurations, analyze workload throughput, transaction response times and device utilizations, tune for optimal performance, define configuration updates to meet any immediate problems and to plan for potential expansion (or contraction) of workload or configuration in the light of business forecasts.

There are three sub-functions within Performance Management. They are Performance Analysis, Capacity Forecasting and Capacity Planning. These can be seen as different views of the same problem, requiring different levels of details in different time-scales.

Performance Analysis

- Compare actual versus planned performance, using as input Performance Data (from Operations) and Performance Plans (from Capacity Planning)
- Analyze performance problems and report on any major deviations from plan
- Decide on short-term tuning actions and produce detailed instructions for Operations

Capacity Forecasting

- Analyze current workloads
- Measure new applications
- Forecast resource usage of new applications
- Predict the performance of future workloads
- Continuously revise all measurements and predictions

Capacity Planning

- Be aware of current and future organizational objectives
- Determine required service levels, consult with business planners and system users
- Maintain Performance Plans, which deal with expected performance variations through time
- Generate and report Action Plans to maintain adequate service, in terms of performance, reliability and resilience
- Revise all plans as forecasts change

2 Structures & Myths of ITIP

The functions of ITIP can be summarized as:

- Maintain Machine & Network Service Levels
- Support business decisions and plans
- Avoid performance bottlenecks
- Optimize resource management & tuning
- Size new applications and their impact
- Predict future workloads and their impact
- Control procurement by effective planning
- Ensure cost-effective and targeted upgrades

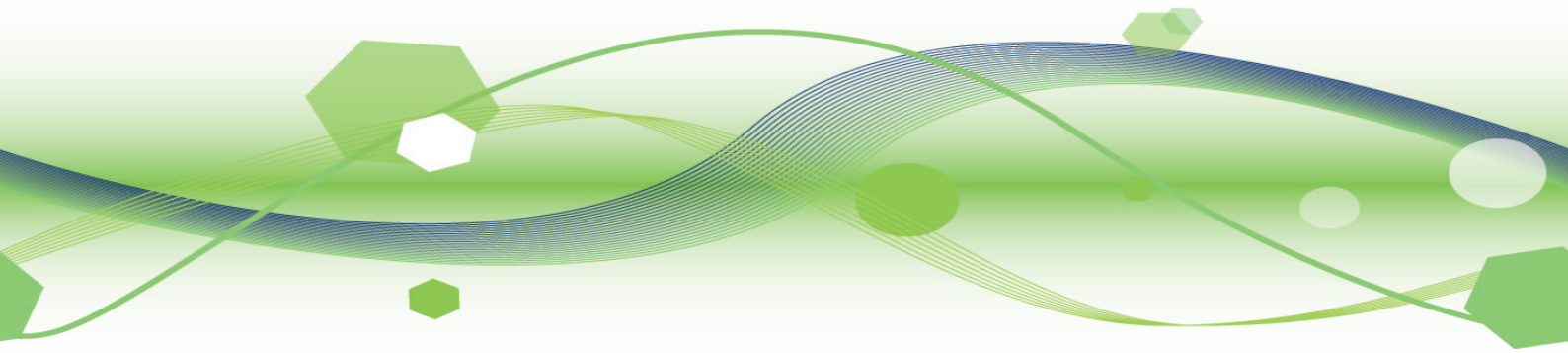
The benefits of improved Service Levels, better throughput and faster responses vary with applications. Similarly, the benefit of avoiding extended debates with end-user departments about performance can be significant in cost and time. The commercial costs of failure due to performance disaster need assessment in each environment.

The costs of ITIP are also well known, at least in terms of hardware and software. The total cost has to take account of the entire infrastructure:

- Hardware
- System & application software
- Infrastructure tools
- Management control
- Technical support
- Network
- User help & User time

In many sites, the purchase costs of the hardware platform and software are the criteria for selection, and yet these typically form less than half of the total annual cost. The size of the investment is thus underestimated and so the advantages in tight control not fully appreciated. The impact of disruption due to multiple incremental updates to configurations, software and applications is great not only within the IT department but sometimes for the entire business.

Any resistance to the adoption of ITIP may lie in the ITIP modules themselves. Each of these has its own myths and realities resulting in problematical dilemmas which are discussed below.



3 Performance Management

Performance Management embraces activities such as system measurement, monitoring, analysis, resource accounting, tuning and optimization. It is a continuing process in any well run data center. It can only be avoided by having a totally static workload, or a grossly over-configured machine. In the first case there is no need for future planning, and so long as the current service is acceptable, the future is secure. This is a rare set of circumstances.

In these economically stressful times, the second case is also rare, where an organization views its IT Service as so critical that it will throw money at it to ensure that all is well, rather than investing the funds in its core business.

Nonetheless a number of popular myths seem to apply in some cases.

Myths about Performance Management

MYTH 1: I don't need to do PM until I have a problem

REALITY: Better to be proactive to predict the problem and implement a solution before it occurs

The basic goal of Performance Management is to anticipate resource requirements, to identify difficulties while they are still potential problems and to implement the appropriate solution before a failure occurs.

MYTH 2: Hardware is so cheap that Performance Management and Capacity Planning is unnecessary

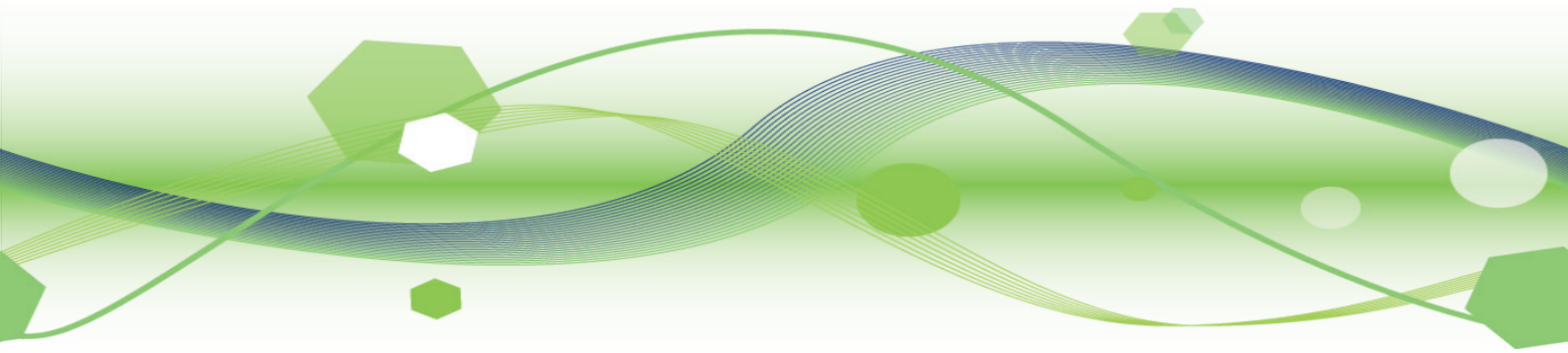
REALITY: Total IT expenditure is still rising so that there is increased cost justification leverage

Hardware price performance is improving at around 40% per annum. The demand for IT resources is increasing at around 60% per annum. Furthermore although each component may have a lower unit price IT systems are increasing in complexity and size to the point where informal, ad hoc, planning methods are totally inadequate.

MYTH 3: Performance Management only needs to be done once a year

REALITY: It is best done as a continuous process of measuring, analysing, predicting and tracking

Performance Management and Capacity Planning is a continuous process in which the projection of future requirements is based on monitoring and analysis that take place regularly during development and production. If you don't understand both the current situation and how it developed you will not be able to forecast your future requirements. Without continuity, performance becomes event driven.



MYTH 4: Real time monitoring and tuning/ optimization is all you need

REALITY: You need a combination of detailed monitoring for analysis and aggregation for planning

The level of real time that is applicable to systems depends on their timeframe. Many commercial systems can be managed on five minute snapshots so long as the system can be seen as of five minutes ago at any time on a browser. Finer granularity for the purposes of cockpit style displays are not effective for solving real problems where patterns of behavior of large populations of users are involved.

MYTH 5: I can't do Performance Management until I have tuned my system first

REALITY: The two processes work best together

There is no such thing as a fully tuned system. Fortunately the PM process highlights system bottlenecks and simplifies the choice of remedial actions. A good PM model will allow you to identify the benefits to be obtained from particular tuning actions. Clearly, the PM model is more reliable once the worst excesses of poor tuning are removed, but such a model will nonetheless highlight the unnecesary bottlenecks.

MYTH 6: Management reporting to the web on a regular basis takes up too much time

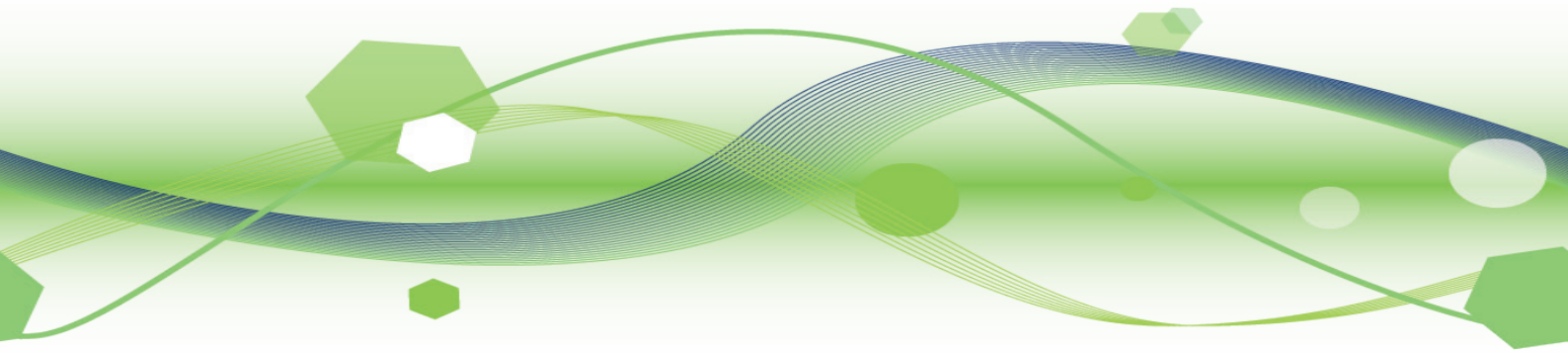
REALITY: With the right tools this can be automated

The need to provide regular management reports increases with business criticality of systems. Rather than issue piles of paper, or stick colored plots on the wall, most sites now want automated dynamic reports on the web showing the status of any node using a browser.

MYTH 7: Analysis and interpretation of performance reports is too complex

REALITY: Automatic advice and exception reporting makes the data easy to understand

The growth of distributed systems with large numbers of nodes requires that management reports are exception based and can also incorporate some intelligent interpretation automatically.



4 Performance engineering

Performance Engineering (PE) is the process of identifying the resources likely to be required to run a new application at a required level of performance. This "application sizing" is best performed as early as possible in the development life-cycle, the sooner the more effective. It becomes more difficult and more expensive to affect changes to a specification as it develops from a functional outline to a detailed specification and mountain of coding.

By the very act of undertaking PE the problems are avoided. Thus it is Good Practice, but it is difficult to find a "before and after" since this is tantamount to asking a designer to admit to being ineffective prior to doing PE.

Performance Engineering tools enable the resource demands of a new application to be defined from the logical specification. There are only a few Performance Engineering tools available which reflects the degree of effort involved in their implementation. This is due to the traditional approach of requiring the user to write the application in a high level pseudo-language to define all the significant processes and accesses to data. These tools tend to be used only on major developments. New tools exploit rule-based systems which incorporate the expert knowledge required to size a logical definition at a more accessible, lower level of detail and are run on workstations.

5 Capacity Planning

Capacity Planning is the “establishment of a regime where computing resources are supplied as necessary for the consistent provision of a required service at a known and controlled cost”. It is conerned with both business issues and technical detail. It is about:

- Effective Planning
- Business Decisions
- Service Levels
- Workload Prediction
- Analytical Modeling

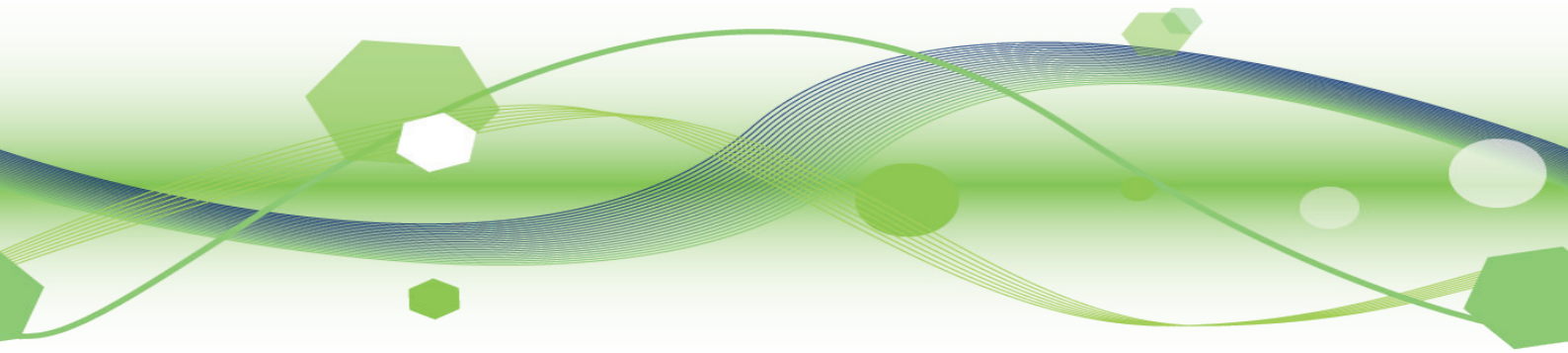
Effective planning means producing plans which allow business decisions to be backed up by the right quality of computer support to users. This process involves predicting the workloads which will result from new and changed business directions, and using analytic models to determine the future performance of IT.

However, few sites seem to undertake regular capacity plans without the stimulus of a planned upgrade.

MYTH 8: Money is wasted on redundant or irrelevant equipment

REALITY: Excessive spending remains undetected until measured by capacity planning

The solution lies in establishing and maintaining a well controlled and timely procurement plan to the benefit of the enterprise. The alternative is the problem of trying to resolve performance panics and shorten procurement cycles.



Network Capacity Planning

Network Capacity Planning is necessary to establish the network delay and hence the end-to-end user response time. As more and more users are accessing the network, the need increases to establish the required network bandwidth to support the traffic.

However, this has been found to require excessive work and additional network bandwidth is relatively cheap. The result is a tendency to over-configure so that the network is not a critical part of the system.

MYTH 9: Network Capacity Plans need too much time to define traffic and workloads

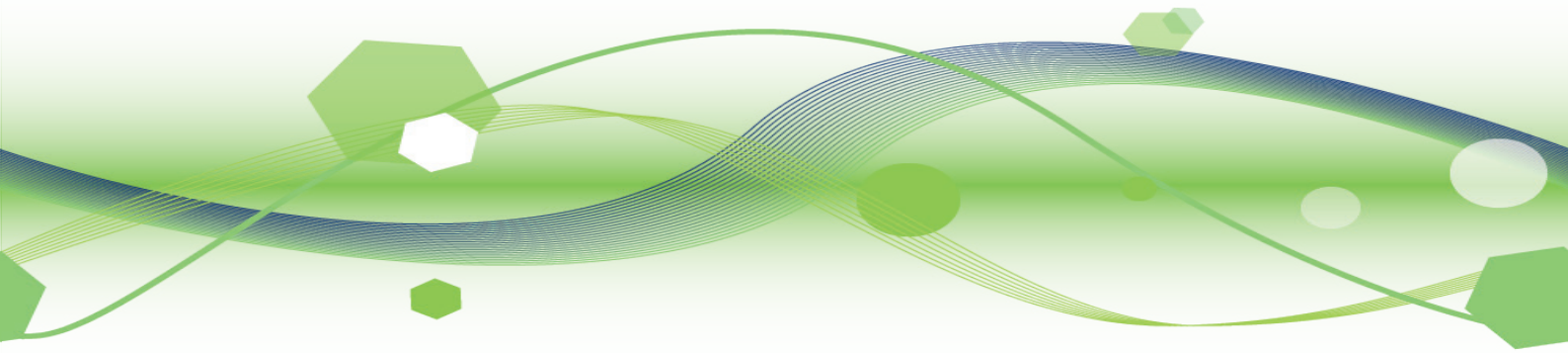
REALITY: Networks are readily incremented and utilization can be assessed by bandwidth

The optimal solution lies in automatic collection of a standard set of performance data (not yet defined) by the network manager for input to new tools for effective planning of networks.

Network Capacity Planning tools model the behaviour of a network to predict the delay in response time due to the network. Network Capacity Planning has traditionally been a discipline that requires significant effort to collect traffic statistics and relate it to workloads. Typical utilizations on networks were low when most communications were point-to-point and in batches.

The emergence of e-commerce, electronic mail, image processing, the increased distribution of computers and the introduction of graphical user interfaces (all of which increase traffic), have all led to concerns about network saturation. The major cost of implementing network planning tools lies in the expertise required to use them and the time required to characterise the workload traffic. So far, this has been seen only worthwhile for major networks where formal Service Level Agreements exist.

However, increasing network loadings and the close integration of networks and processing nodes, in, for example, client/server systems, may force a reappraisal of attitudes. A new generation of network management tools and planning tools that integrate networks and computers in a single model may be required. Although the incremental cost of upgrading networks may be less than that for mainframes, and the procurement cycle shorter, the inter-dependence of networks and open systems will lead to greater adoption of such packages.



6 Service Level Agreements

Service Level Agreements are a written contract setting out in quantified terms the obligations of the provider and user of a computer-based service. They must be realistic so that they meet the business need, are achievable, measurable and verifiable. They are the basis for a Partnership and impose limits on the user as well as targets for the service provider.

The points to be stressed here are that:

- To a user, an acceptable level of service is that which enables him to do his job properly;
- The user's "job" is dictated by the aims and objectives of the organization;
- Almost any level of service can be provided if money is no object;
- Since very few organizations have bottomless wallets, what constitutes "acceptable service" becomes a matter of discussion and negotiation between several interested parties, not only IT.

They are usually implemented where the service level is mission-critical or where previous problems have led to more formal control. In either case, the implementation of agreements is a major exercise and is only practicable with the right tools, of which there are few.

They are based on a marriage between real time monitors, capacity planning packages and systems to help arbitrate on violations.

MYTH 10: Service Level Agreements need too much effort to establish and track

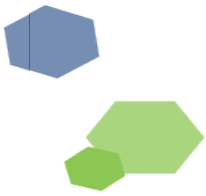
REALITY: SLAs are only agreed when the service provider knows he's safe

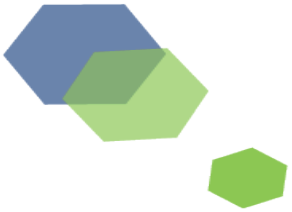
Service quality has several dimensions, including:

- functionality
- ease of use
- performance
- availability and reliability

Traditionally IT managers have concentrated on functionality, usability and reliability and have assumed that performance is something that the technicians will put right once they have made the system work. The results of this approach are readily apparent from a quick scan of the computer press and it is sure that applications have been delayed or abandoned because they couldn't be made to perform and computers and networks have had to have unscheduled upgrades to handle their work. Although this policy may have sufficed when IT systems were relatively simple it is fraught with danger as systems have grown in size, complexity and criticality. The importance of defining and managing service levels effectively is now widely recognised, especially in terms of e-commerce.

These "top ten" myths are intended to be thought provoking rather than being presented as major obstacles to the adoption of the ITIP process. However, it is often some of these negative ideas that have to be overcome at new sites in order to establish the need for ITIP. Only then can a formal cost-benefit model be used to confirm the value of an ITIP project.





www.metron-athene.com