Capacity Planning:

How Accurately Can You Predict the Future?

BY CHRIS FLYNN AND JIM FOXWORTHY

Effective planning involves predicting the workloads resulting from new and changed business directions and using the proper analytic models to determine the future performance of IT.

IN the '80s and early '90s, the discipline of capacity planning was well established because the incremental cost of new hardware was so large — sometimes in the millions of dollars — that it was unhealthy for data processing management to go to the corporate financial management and announce: "Surprise, we need another million dollars next month." (Financial professionals frown on that kind of surprise.) Today, incremental costs are considerably lower, but the need for change is far more frequent. (Financial professionals are very good at adding up 100 requests for $10,000.) The problems of capacity planning are the same these days, but there are just more of them.

Capacity planning is a process of managing three variables (see Figure 1):

- the present and future applications that must be run to meet an organization's business objectives
- the levels of service that are required by the users of these applications
- the hardware and system software on which these applications are to run

These variables are interdependent. In particular, the combination of applications that run on specific hardware configurations will determine what service level the end user will receive. In most organizations, the use of computer systems will be the only way to achieve certain major objectives. It follows that the ability to manage and predict service levels is very important because end users often can do their job properly only if the levels of supplied service are adequate.

Such organizations must be able to predict the levels of service that can be supplied when a given group of applications are implemented on a particular hardware configuration. Reality, however, usually isn't that predictable. Users develop new applications, existing applications are used in different ways, hardware and software technology is in almost continuous flux, and the needs and expectations of users vary as new facilities are exploited. It is essential to predict service levels not only of existing applications, but also of new applications that may run in the future on processors whose power can only be estimated.

What forecasting methods are available for the prediction of service levels?

- rules of thumb
- linear projection
- simulation
- benchmarks
- analytical models

Rules of Thumb: These may be reasonably accurate or wildly inaccurate. Such rules are often useless unless one understands the reasoning behind them. A more formal method is necessary for long-term use.

Linear Projection: This is satisfactory for predicting device utilization (which can be important) but is incapable of dealing with contention for resources between different applications. It

Figure 1: The Process of Capacity Planning

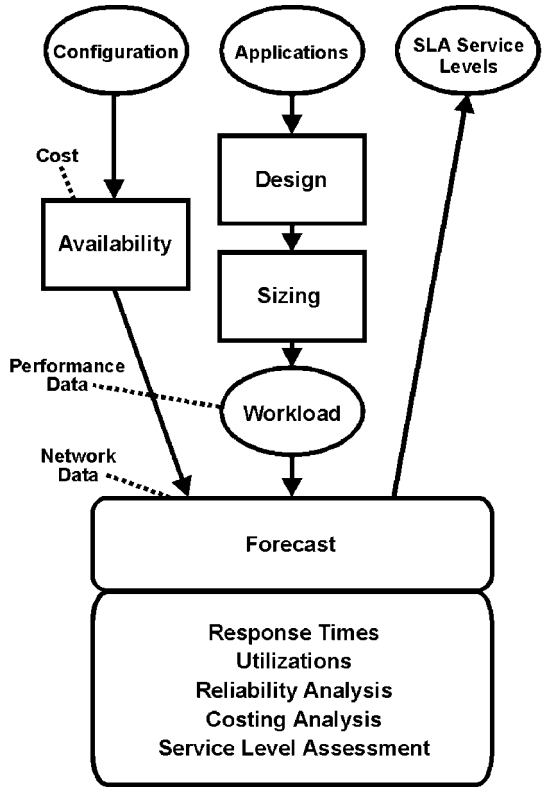
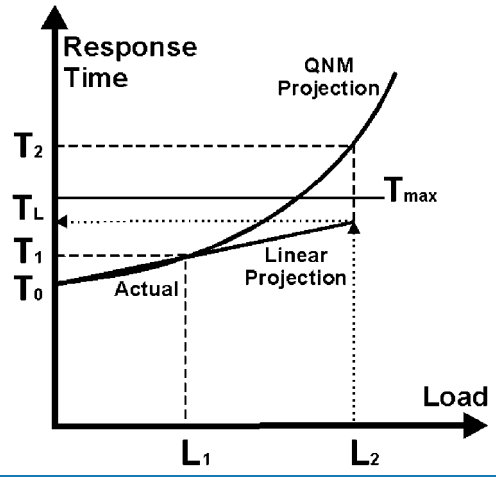


Figure 2: An Example of Linear Projection



may give dangerously misleading results if used to forecast service levels.

In the example in Figure 2, the actual response time observed at Load L1 is T1. A linear projection based on this would suggest that at Load L2 the response time would be TL. A queuing network model (QNM) projection, on the other hand, would predict T2, outside the required Tmax. This would correspond to within 10 percent of the actual behavior of the real system.

Simulation: This is potentially very accurate but ends up being quite expensive, however, as it requires considerable skill and expertise to produce good results and can be very time consuming (typically taking months or even years).

Figure 3: An Analytical Modeling Example

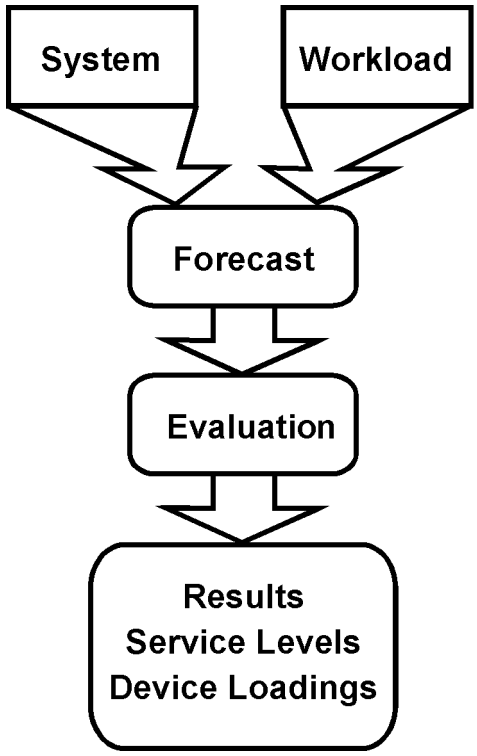
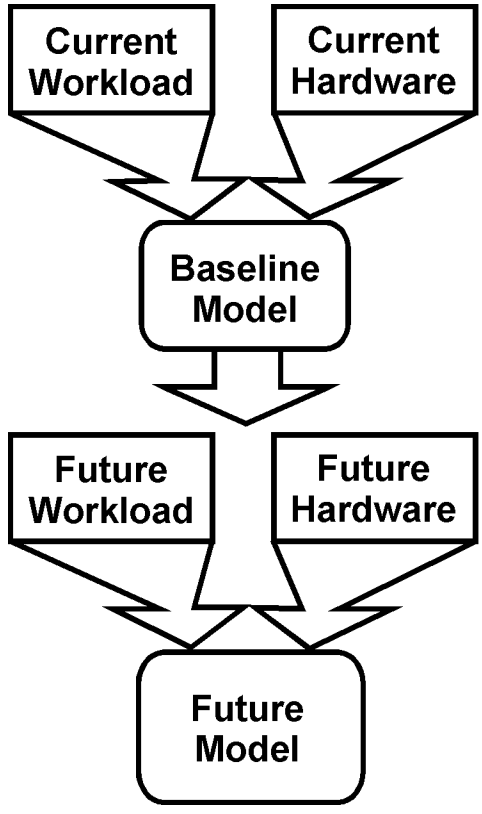


Figure 4: Using a Baseline Model



Benchmarks: These are excellent for comparing the performance of existing applications on currently available hardware but they can also be very expensive to set up. Additionally, the future aspect is missing.

Analytical Modeling: This handles future applications and systems as easily as current ones and is sufficiently accurate for most purposes. It can predict service directly and produce results in minutes. It uses everyday descriptions of available data, so it requires no specialist skill.

For these reasons, analytical modeling is generally accepted as the best way to predict the performance of computer applications.

An analytic model has two main components (see Figure 3):

- The system, which is a simplified description of the major hardware items (processors, controllers, discs, tapes) on a configuration (real or potential).
- The workload, which describes the number of times each device in the system is used by different types of transactions or jobs.

A well-designed analytic modeling tool will have data collection and analysis facilities, specific to one or more operating systems, which will as far as possible automate the process of describing these system and workload components. This is important because time can be wasted if the tool does not provide this support.

The combination of a system and a workload is referred to as a forecast. The modeling tool evaluates the forecast and predicts what service levels will be provided by that combination of hardware and software. Hence those three variables — the applications, the hardware and the service levels — can be integrated into successful plans.

Naturally, the IT manager should be wary of unsubstantiated predictions, however theoretically sound the method of arriving at those conclusions. For this reason, the methodology of analytic modeling is extended to introduce the idea of a baseline model (see Figure 4).

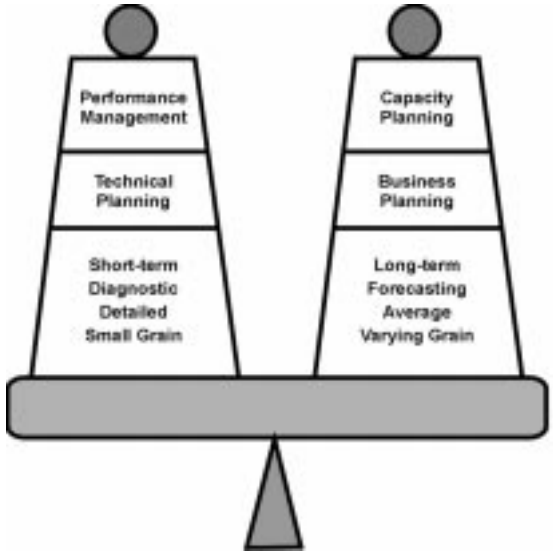
A baseline model is an analytical model built to represent the current state of the system and workload. The model is built from standard system and workload monitoring sources, appropriate to the hardware and operating system in use. As one would expect, the model will predict service levels. In this case, the predicted service levels can be verified and tested by comparing them with the actual service levels measured by the system monitoring tools. This allows planners to verify the accuracy and applicability of the model before it is used to predict the future. Once the baseline has been established, the IT manager can be confident that predictions about the behavior of future applications will be correct.

By modifying a verified baseline model, one can represent changes to the workload and to the hardware. The changed model will predict the levels of service that can be provided under these new circumstances. Modifying models in this way with a well-designed tool is so quick that literally dozens of scenarios can be explored in just a few hours. This ensures that plans will be complete and that potential avenues are not left unexplored for lack of time or effort.

Planners must strike a balance between the appropriate level of detailed information needed and the effort expended to obtain it (see Figure 5). In the field of performance management, this is most true of the trade-off between the very detailed and specific information that may be required to resolve a particular performance problem and the more general average and trend data that characterize capacity planning.

The level of detail required for successful planning might well be different from that required to solve a performance problem. Some confusion can arise because both activities use largely the same data. While it is possible to use analytic models to resolve detailed application-specific performance problems, it is their long-term impact that is more important and which is practically impossible to achieve by any alternative approach.

Figure 5: Capacity Planning vs. Performance Management



Capacity planning, then, is primarily a business activity that happens to have technical content. It is the means by which service levels can be protected in the future.

- effective planning
- business decisions
- service levels
- workload prediction
- analytical modeling

The fruit of effective planning will back business operations with the right quality of computer support to users. This process involves predicting the workloads resulting from new and changed business directions and using the proper analytic models to determine the future performance of IT. ts

---

Chris Flynn is the director of distributed product development for Landmark Systems Corporation. Jim Foxworthy is product manager for Landmark Systems Corporation, and is responsible for the management of the company's client/server solutions. He oversees all aspects of the product life cycle, from new product requirements through product launch and the eventual withdrawal of products from the market. He has more than 20 years experience in the computing industry.