Section VIII -- Capacity Planning

Why do capacity planning?

In some sense, every computer site does some form of capacity planning, even if it is only an occasional acknowledgement that the machine is getting busier and maybe it's time to do something about it.

Effective planning efforts are a hallmark of successful businesses. Important internal services must be available as the costs of disruptiuons can be dramatic and exponential. Over capacity can be equally damaging due the lost opportunities caused by the devotion of capital to unused resources. Expenditures must be planned, both as a mechanism to insure adequate review for cost justification purposes, and so that financing can be obtained in the most economical manner.

Computers should be treated no differently than any other expensive resource, but often are. Too often, management doesn't understand them. Too often the technical people managing them, either deliberately or through lack of business sense, do not properly assist management. Data center managers, capacity planners, or performance analysts who do not put their technical activities in a proper business millieu are doing their employers a disservice. Proper planning is a must for data centers.

Goals and Objectives

The goal for the capacity planner is to insure that adequate computer capacity is available to meet the organization's needs.

A number of objectives, constraints and considerations relate to this goal:

> The capacity planner must have an understanding of these needs, but this is not accomplished by simply knowing that applications A, B and C have to be run. The needs assessment must be qualified by defining service levels appropriate to those needs. Service levels are statements about response times and turnaround times.

Planning for service levels includes negotiating with user groups about what service objectives the computer center will be expected to meet. Often written agreements between the computer center and the user group serve to quickly arbitrate problems and result in better service delivery in the future.

The acquisition of capacity must be economic. The planning process serves as a vehicle for:

- cost benefit analysis
- identification of fruitful areas for service improvement
- timely consideration of alternate strategies

The planning must result in timely acquisitions of capacity. Obviously, running out of capacity is a problem. Having it too soon is equally damaging:

- carrying costs
- unreasonable reliance on exceptional service levels develops
- it invites inefficient usage

"Nothing consumes computer capacity more than unused capacity."

Functions of Capacity Planning

Long Term Planning

The capacity planner must play a central role in decisions concerning major hardware and software alternatives. Critical to any decision will be performance (capacity) relative to cost.

Hardware comparisons can be deceptive. Comparisons of published CPU cycle times are thoroughly misleading. Many factors are not included:

- relative instruction power
- hardware overheads
- operating system overhead
- layered software performance
- load degradation
- compiler performance
- I/O device capacities

Capacity planners must develop appropriate measurement tools based on the uses to which the machine will be put.

Short and Mid Term Planning

The capacity planner -- like the performance analyst -- must be continually obtaining information about utilization:

What is the current demand?
- By user and application
- By time (loading patterns)

What is the current capacity vis-a-vis the demand?
- % utilizalation, average and peak
- service level being delivered
- safety margins

What growth or shrinkage can be forecast?
- Change in user workload by application, user and time
- Change in software -- performance changes, new versions, etc.

What capacity will be needed at any given future date?
- Must acheive desired service levels
- Must reflect workload mix
- Changes in peak loading?

What alternatives to simple addition of hardware exist?

The Capacity Planner's Tools

- cost prohibitive

Hardware Monitors

Devices which monitor hardware operation to measure frequency of various operations -- instructions counts, idle time, memory references, cache hits, etc.

Advantages:

- Accurate -- no impact on machine operation
- Flexible -- if it can be electrically sensed it can be measured
- Transportable -- a hardware monitor works on any machine

Disadvantages:

- Expensive -- simple monitors cost $30,000 and up; comprehensive ones are $100,000+ } actually quite higher than Stat's here.
- Cannot distinguish logical load components
- Requires skilled operators who understand hardware logic, software, and can interpret the results
- Use of the monitor is labor intensive
- With modern VLSI, there are many fewer access points to the CPU logic, and, therefore, many fewer functions which can be monitored.

Generally only useful to large organizations which can utilize the device on a number of computers and devote a full time analyst to it.

Software Monitors

a good alternative if you find a good one.

Software which runs on the system being monitored and gathers and analyzes performance data.

All variations of effectiveness, accuracy and efficiency can be found among the available products.

Advantages

- May be easy to use and understand as they are targeted at a specific architecture.
- May be able to segragate load components
- Can look at a wide range of variables
- Relatively inexpensive

Disadvantages

- Can impose severe loading on the machine, possibly becoming a capacity problem in and of itself and skewing the results
- Accuracy may be questionable, depending on measurement technique and coverage of all aspects of resource utilization. "Capture Ratio" may or may not be calcuable.
- May "interpret" the data in a manner not supported by logic or fact
- May focus on irrelevant data or statistics
- Generally not able to look at detailed hardware measures, such as instruction counts and distributions
- Not system independent

Benchmarks

Useful for machine to machine comparisons. May be actual or synthetic.

Actual benchmarks are a representative sample of an actual workload. While most accurate, they are very difficult and most time consuming to set up.

Synthetic benchmarks are created to simulate an actual workload yet be easily transported from machine to machine. To be of use, they must accurately reflect the character of the subject workload:

- instruction mix
- quantity and type of I/O
- memory utilization
- use of and dependicies on the results of system components such as layered software, service routines and compilers.

Stress Tests

A type of benchmark used to predict response behavior at various loadings of a multi-user application. The procedure is to run the application with increasing numbers of users, each following a carefully crafted "script", measuring response and machine loading at each usage level.

Rules of thumb for typical transaction or data entry type applications on the VAX:

- after 1 or 2 users, machine usage will increase linearly
- response times will increase slowly and evenly until 90% (+ or - 5%) CPU loading, at which point it will increase rapidly. This assumes I/O device capacities are not fully loaded at any usage level. The 90% figure is exclusive of low priority (batch) loads.

(In short, because of the efficiency of VMS -- assuming adequate memory, properly tuned -- the load behavior of the VAX under VMS follows a linear trend, and simple extrapolations are meaningful and reliable except where I/O device capacities will be taxed. This exception is, however, rather rare among VAX software.)

Simulation and Modeling

These are methods which use statistical procedures to represent a load. To be effective requires intensive analysis and appropriate representation of all load variables, interelationships, and queueing patterns in mathematical terms.

The analysis of empirical data or tests is generally much easier and equally or more accurate.

Procedures for Capacity Planning

The most important part of capacity planning is effective forecasting of workloads.

Top down forecasting, based on trend analysis, is unsuitable for multi-application loads. Real world change is seldom constant over time or among various users.

- time series proj. based on statistical data in general be careful when using stats as they assume certain assumptions that are very misleading in real-world computer applications.

* The forecast must be bottom up -- application by application, even user by user. This forecast must be based on communication with the user groups and their management. What and when are their expectations of usage changes.

Discussions with users must be in terms of "natural forecasting units" such as number of reports, inquiries, customers, payees, invoices, etc., as opposed to computer technical terms such as CPU seconds, disk blocks, etc.

Current usage levels, in natural units, must also be known.

Based on current usage, accurate assessments of resource units -- CPU seconds, I/O's, disk blocks -- per natural unit must be calculated. Be careful to allocate overhead to users responsible for it -- those doing large amounts of terminal I/O and those using clustered disks -- especially if shared file applications are being run.

An accurate demand load can be now calculated for any point in the future.

Capacities of equipment should be apparent based on technical data -- maximum I/O rates and disk sizes -- and general monitoring or testing, observing inallocable overheads and loading limits for acceptable response.

Care must be taken to account for DP center management functions. For example, as disk space utilization grows, so will resources devoted to backups and reorganization runs.

When determining capacity adequacy make sure peak loading is considered. Demands of applications must be considered by when used -- time of day, day of week or date in the month.

