META Group

META Delta

22 Month 2002 **File: SMS 1098**

Capacity Planning in a Distributed World

Service Management Strategies

Glenn O'Donnell

Infrastructure capacity planning has intensified due to renewed focus on capital expenditure justifications. This trend will continue to accelerate through 2003, coinciding with growing operational maturity among IT organizations. The deluge of variables, the lack of simple, comprehensive tools, and generally low operational maturity have hampered such efforts to date. Vendors are rapidly evolving tools, though full-featured, user-friendly solutions will be atypical and complex through 2005.

IT organizations are charged with accurately assessing infrastructure capacity and planning for future needs (see SMS Delta 883). This capacity planning/capacity management (CP/CM) process is common in mainframe environments but is missing or inept in most distributed systems. Mainframe operations are more mature, and most variables in the capacity equation are well understood, closely contained, and tightly controlled. Distributed systems are more dispersed, less predictable, and usually under the control of several parties.

Operational maturity is now infiltrating distributed IT operations, driven mainly by economic pragmatism. Comprehensive, distributed CP/CM endeavors are currently rare (2%-4% of Global 2000 organizations), but 20% are now beginning to address this issue. By 2004, 60% will be attempting some form of methodical distributed CP/CM. Success will be a struggle, with only 15%-20% achieving good planning processes by 2004 and 30%-40% by 2006. These early attempts will often be focused on specific technology silos (e.g., WAN, storage, Unix servers); however, such targeted ventures will offer limited rewards. Distributed capacity planning will not be robust or ubiquitous until 2008/09. Through this period, CP/CM will be somewhat tedious, but even limited processes will yield significant benefits.

CP/CM serves many purposes. The most familiar is to predict future infrastructure requirements for efficient planning of capital expenditures to match business needs. Others serve consolidation and reduction to avoid or diminish wasted overcapacity. Chronic performance problems may require detailed analysis that can be accomplished in the CP/CM process. Such analysis is done in conjunction with the problem-management process.

Although CP/CM is closely related to performance management, the two are operationally quite different. Both relate component and application performance data to business needs, but performance management is a more real-time, tactical process, and CP/CM is a more intensive, longer-term strategic function. CP/CM supports an engineering attitude by enforcing a systematic process with tight controls and deterministic results.

Considerable stranded capacity exists in many enterprises, with past capital spending resulting in underutilized resources. Capacity assessments can identify wasted resources and result in cost savings through consolidation or reductions in future purchases. (In one case, a company was able to reduce 80 Web servers to 18 while supporting a moderate increase in service demand.) Initial assessments yield the highest returns for those most negligent in their current operational processes.

Four common methods are employed for CP/CM: exception triggers, trending, modeling, and lab testing. CP/CM efforts will center on processing cycles and memory (for servers), storage (for storage devices), and network performance (bandwidth, latency, throughput, etc.). The appropriate method for any particular resource depends on the performance characteristics of that resource (stable utilization with a gradual increase in demand, stepwise demand increases/decreases, etc.).

META Trend: During 2002/03, IT groups will assess operational process maturity and formalize process models. Through 2004/05, IT efficiency growth will focus on process integration, measurement, and aggregation of synergistic process groupings (i.e., centers of excellence). Through 2006, change, configuration, and asset management process automation will remain high-cost options.

Exception triggers base capacity needs on utilization threshold violations. This method is highly reactionary, often causes wasteful firefighting if triggers are not set appropriately, and delays adaptation to changing business requirements. Triggers are useful for performance management, if managed correctly, but are often inadequate for robust CP/CM. However, one trigger technique does work to identify underutilized infrastructure. Low performance thresholds are set, and components that consistently fail to exceed these thresholds are underutilized.

Trending applies simple extrapolations (e.g., linear regression) to performance data to predict future values. Trending is usually an inaccurate prediction of capacity because linear trending erroneously assumes future growth will continue along the same patterns of the sample period, giving a false sense of predictability. Changing demands can cause large performance swings that are missed by this approach. Advanced statistical methods show promise beyond simple linear regression. Most of this work is still confined to academic research, but some commercial solutions (e.g., from ProactiveNet) are beginning to exploit such algorithms. Statistical predictive tools offer value, but without a means to integrate future business requirements, their utility for CP/CM is limited.

Modeling is a more intensive, but often more accurate method of CP/CM. Modeling incorporates software descriptions of infrastructure and application signatures for simulation. These inputs, with the rules of the simulator, emulate a planned environment where scenarios (i.e., “what if” analysis) can be tested well before actual infrastructure or application development commences and capital expenditures are disbursed.

Most modeling vendors (e.g., Opnet, TeamQuest, HyPerformix, BMC Software, Analytical Engines) emerged from silo-centric markets (e.g., Opnet’s networking roots, TeamQuest’s server roots), but all are broadening their scope to encompass more sophisticated application behavior and end-to-end infrastructure. Tool innovations will accelerate through 2003; however, users should expect some limitations through 2005. Modeling embryonic technologies (e.g., Web services) will be difficult until well after these technologies stabilize.

The most crucial input to the modeling process is an accurate assessment of future business requirements. The software model captures the technical infrastructure composition, but requirements manifest the ultimate realism of authentic business projections.

Active stress testing is similar to modeling in its execution, but it uses actual infrastructure instead of software models. Initial stress testing is usually done in a controlled test lab and not the production environment, to minimize the risk of service interference by an unruly test. Stress testing is a logical follow-up to modeling as a means to validate the results of modeled simulations. Stress testing hardware (e.g., from Antara.net, Caw Networks, and Spirent) and software (e.g., from Mercury Interactive and Empirix) inject application transactions into the infrastructure at high rates to determine performance limits.

Despite the costs and difficulties associated with modeling, it remains (if available) the most effective technology solution for CP/CM. Ease of use has improved dramatically, and costs are dropping. Any complex system, including airliners, automobiles, integrated circuits, skyscrapers, and IT infrastructures, must be properly engineered. Carelessness and poor planning do not scale. The results are failure-prone, costly, and cannot sufficiently support the needs of the business. Engineering discipline ensures predictability, reliability, and high performance at an optimum balance of costs and capability.

Bottom Line

Capacity planning and capacity management are gaining importance to understand infrastructure utilization and to objectively justify future capital expenditures. Modeling is maturing as an effective technology solution, but fragmented methods will moderate execution through 2005.

Business Impact: IT organizations should implement engineering disciplines in IT capacity planning to balance optimized return on capital expenditures with maximized service expectations.