META Practice

Practice 2100 **26 September 2003**

Optimizing Infrastructure and Application Performance

Infrastructure Strategies, Operations Strategies, Service Management Strategies

Glenn O'Donnell

FOCAL POINT

For many organizations, performance of IT infrastructure and applications is a troublesome mystery. For others, such performance is well understood and maintained under rigorous control by a pervasive culture of discipline and structure. Unfortunately, most organizations align more with the first scenario, since IT practices have evolved largely without coordinated guidance. For this group, performance is more of a black art than a planned business enabler. Thankfully, this is changing at an accelerating pace, with broader adoption of capacity planning processes and well-engineered infrastructure and applications.

CONTEXT

Performance management is receiving increased attention within IT organizations, because it offers a means to measure infrastructure and application health as well as the foundation on which we can measure business value and impact. This business-value theme is causing IT organizations to refocus performance management efforts. Since monitoring the performance of underlying infrastructure components is a process that is mostly mature, concentration has moved to applications and optimizing overall performance across both infrastructure and applications. The enigmatic end-to-end perspective is finally coming within reach. It is this perspective that facilitates the measurement of true business value.

Availability Improves, But Performance Lags

In many IT infrastructure areas, fault-tolerant capabilities are now becoming ubiquitous (e.g., switched environments, load-balanced applications, clustered databases). Redundant systems can absorb the impact of failures with little or no perceptible impact on end-user service availability, which is a change from past problem-resolution processes, which were optimized with the expectation that a failed device would be the most significant problem. Performance is not such a simple issue, however. When a slowdown impedes business processes, the root cause cannot be easily masked by redundancy nor can it be easily detected.

Merely increasing capacity in an attempt to conceal performance issues is no longer a viable strategy. In some cases, additional capacity can alleviate the pain, but increasingly, other factors beyond simple capacity are involved, not to mention cost issues. A classic example is a networked application. This obsolete approach mandates more network bandwidth to solve performance problems, but “chatty” applications that involve frequent protocol exchanges are more severely impacted by network latency and other delay contributors.

Chattiness is increasing as applications and IT services become more complex. A new approach to pre-emptive avoidance of performance issues is needed. In the development stage, a rich variety of real-world conditions that can impair performance of information technology, and thus the business, must be anticipated. A methodical and disciplined engineering attitude, enabled by strong performance planning tools, is the solution.

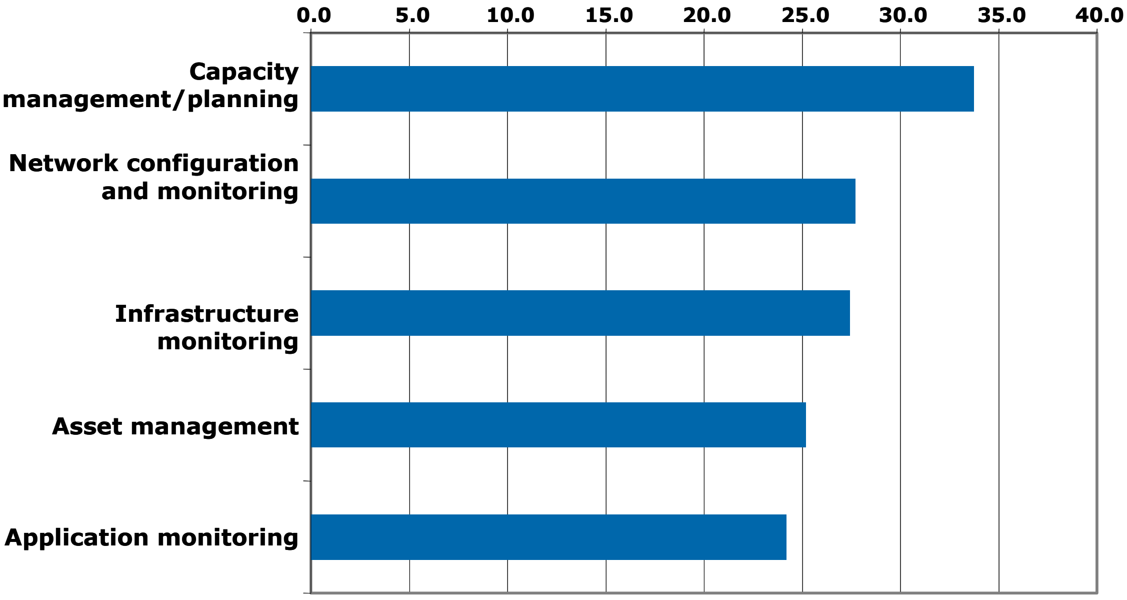
Performance Optimization Business Drivers

Engineering IT systems for optimum performance is not a new idea, but it has taken on new significance as economic conditions tighten. Performance optimization is a form of capacity planning (see SMS Delta 1098), but with a special focus on the resulting infrastructure and application characteristics and the relationships between the two. It is these relationships (see SMS Delta 1146) that bring cohesion to both domains for superior quality of the resulting end-to-end system. Planning endeavors (e.g., infrastructure, applications, capacity) must incorporate these relationships.

META Trend: Through 2008, IT operations groups seeking to effectively develop and enhance their operational processes must formalize their efforts, focusing on process definitions, performance measurement, and analysis of potential refinements — ultimately creating a culture that embraces continuous improvement. Although most IT operations groups' efforts are still in their infancy, significant gains will be made by leveraging the process refinement practices experienced by both IT (e.g., ITIL) and non-IT oriented (e.g., Six Sigma) organizations.

A recent META Group infrastructure and application management study (see Practice 2007) reveals that capacity planning is the top critical issue for large enterprises, with 33.8% of respondents identifying this as a critical issue for 2003. This high priority will continue through 2005/06, escalating consolidation and efficiency demands. The figure below (see Figure 1) shows the top five critical issues for large enterprises (i.e., those with >1,000 people). It is noteworthy that four of the top five issues involve performance concerns.

Figure 1 — The Top Five Critical Management Issues for Large Enterprises



Source: META Group

The various performance-oriented efforts are changing to include deeper analysis of performance data (see Delta 2259) and to trigger actions based on the data (see SMS Delta 1069), causing basic collection of monitoring data to be commoditized. One actionable use of this data that is gaining momentum is pre-emptive optimization of performance characteristics. There are several reasons for this heightened awareness of performance optimization (and also of general capacity planning). Some key drivers are:

- ***The IT organization needing to quantifiably and reliably demonstrate business value:*** The IT organization exists to serve its employer's business requirements, with business leaders understandably requiring proof of a return on their IT investment. Several metrics can be captured to quantify this value. Optimized performance directly impacts development cycles, accuracy, and a reduction in adverse incidents, which all can be easily translated into cost savings, and in some cases, even revenue increases. Structured processes with measurable business value convey the IT organization's contribution to improving the company's overall business performance. A pre-emptive approach to performance optimization bolsters structured processes that have been long neglected in the industry.
- ***Painstaking spending justification:*** Spending for initiatives such as infrastructure expansion, application development, and consolidation now require more meticulous validation than ever before. Careful planning prevents unnecessary spending through appraisal of resource requirements and anticipation of performance against available existing resources. This becomes a proactive expense control, while still maintaining a high standard for performance expectations. This effort is straightforward in infrastructure development, since capital expenditures are tangible, but it may not be an immediately obvious step for application development, since cost savings come from reduced application development cycles. Without proper controls, these costs are less predictable because the number of repetitive cycles cannot be known.

- ***A need for pre-emption to enable anticipation of potential obstacles:*** All possible scenarios that could potentially scuttle IT projects must be planned for. Using pre-emptive approaches, applications and infrastructure can be designed to prevent the occurrence of obstacles or to adapt to circumvent obstacles that may arise. Anticipating and planning for failure modes is a common engineering practice used to ensure strong systems with a minimum number of design iterations. Ideally, organizations want to "get it right the first time." The cost of multiple development cycles to achieve functional systems is unacceptable.
- ***Traditional performance efforts being insufficient:*** Some level of planning performance is now common in IT organizations. Although the industry refers to this as capacity management or performance management, little actual management is performed. Performance data is collected and reports are generated, but the resulting action is still a tedious, manual response. Performance "management" implies some assisted or automated action resulting from the measured data. Optimizing performance is difficult using this simple view and simple management technologies.
- ***Optimization being an endless pursuit:*** Focusing on optimization in system development and then forgetting about the system's performance until a crisis happens is the wrong approach. Changing conditions and demands require optimization to be a continual effort. This applies, of course, to maintaining performance levels, but profound benefits emerge as the system continues to be fine-tuned and additional performance increases are extracted (e.g., Six Sigma). These iterative processes yield profound benefits to the IT organization and the business due to operational efficiency being maximized (and costs minimized) with this approach. The same principles are used to squeeze out margin and market advantage in any highly competitive business (e.g., discount retail, fast food).
- ***Excessive complexity leading to difficult prioritization:*** IT environments are now so complex that companies are inundated with excessive alerts, which make noise more than provide actionable information. In addition, due to the extensive infrastructure in place, it is often difficult to identify which alarms are affecting the business in the most significant way and therefore require the most resources.

Any one of these drivers alone would compel performance optimization ventures. Together, they form an imperative that cannot be ignored. Significant value is achieved by injecting stronger discipline and more intelligent analysis into any IT planning and development exercise and any ongoing operations.

A common theme among all of these drivers is economic benefit. As we strive to operate the IT organization as a business unto itself, leaders must consider their responsibility to provide effective yet economical services to their customers. This fiscal prudence is motivating a focal shift from technology to operational processes and best practices aimed at business value. Of course, technology is still the enabling end product, but with a process focus, the operational processes are tuned and then technology is sought out to accelerate execution of those processes. The reverse (i.e., building processes around predetermined technology solutions) leads to almost certain failure, since processes are then limited by the scope and functional limitations of the technology.

Performance Optimization vs. Performance Monitoring

Performance management is a commonly accepted operational process intended to bolster business relevance. In some ways, it is a mature market, given the long history of performance management products and the pervasive use of these tools. Yet most actual process implementations are immature. There are many facets of performance management, including monitoring and optimization subprocesses, and many dimensions represent an inherent complexity of the process, which impedes high maturity levels. Process simplification is achieved through the use of industry best practices, common technology (where possible), structured data flows, and clear roles and responsibilities, as well as by dismantling technology silos. It is also helpful to understand important performance management subprocesses. Performance monitoring and performance optimization are related subprocesses. Both are necessary and present unique business value to the organization, but each presents diverse operational and technical issues. Although there are similarities across the two, we can identify some significant operational differences (see Figure 2).

Performance Management

Performance management remains a reactive process aimed at performance-incident resolution. Some vendors promote proactive performance management, but the "proactive" qualifier is often stretched for marketing purposes. With performance optimization, performance management truly becomes proactive, since performance issues are anticipated and solutions are pre-emptively engineered.

As the market drives performance management to become proactive, we will see the line blur between management and optimization. Eventually, it will be all management and optimization will be implied. This drive toward optimization must come from users, since most vendors continue to struggle in marketing an optimization message. Because much of this trouble is the result of general operational maturity (i.e., How can one optimize that which can't even be measured?), increasing maturity will help the convergence toward robust performance management that includes optimization.

Figure 2 — Comparing Performance Monitoring and Optimization

Issue	Performance Monitoring	Performance Optimization
Organizational driver	Production/operations	Development/engineering
Timeliness and frequency	Real-time notification of performance events is performed continually	Performed during development, QA, and predeployment, and periodically afterward for performance refinement
Current adoption	Medium to high	Low
Objective	Rapid identification of performance bottlenecks to reduce mean time to replacement	More robust infrastructure and applications to meet business requirements
Growth rate (2003-05)	10%-15% annually	30%-50% annually

Source: META Group

Performance Optimization

Performance optimization fine-tunes infrastructure and applications for anticipated business requirements. In the purest sense, it is synonymous with capacity planning. In practice, most capacity planning efforts are partial realizations of performance optimization, but the two are converging. Capacity planning is usually a postdeployment process, whereas performance optimization occurs throughout the entire system life cycle. Predeployment optimization is mainly application testing, which has progressed well, but inclusion of real-world conditions in application testing has lagged. Similar fundamental performance technologies can be used in all stages of optimization, so gradual convergence of performance-oriented products (both predeployment and post-deployment) is certain through 2007. Also worthy of consideration in performance optimization is the notion and practice of tuning. Tuning is a form of performance optimization that is most common in application development.

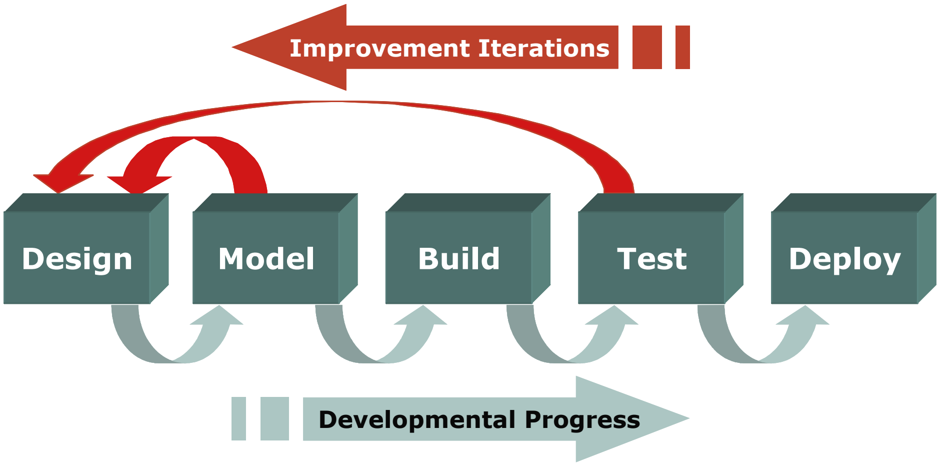
Performance Optimization in the Development Process

The first step toward performance optimization is to embrace rigorous methodologies for infrastructure and application development. Haphazard assembly of these critical business elements is becoming intolerable. True engineering principles of design, modeling, and testing/verification (see Figure 3) are necessary to ensure accurate performance expectations and minimize inopportune surprises. The following guidelines apply for both applications and infrastructure projects, with some noted differences.

All complex systems (e.g., aerospace, automobiles, semiconductors, bridges, skyscrapers, global economics) are developed following these proven engineering principles. In many of these cases, a failure to follow such intensive development can be catastrophic (e.g., the dot-com economic bubble collapse, the Three-Mile Island nuclear disaster). IT failures may not cause such spectacular tragedies, but the impact to a company can be crippling or even terminal. Failures are inevitable in extremely complex systems, but good planning discipline significantly minimizes such failures.

Although the chain is sequential, feedback loops provide a means to assess success and retreat to redesign until the desired results are met. The model stage is important because it uses simulation and emulation (smaller, targeted simulations) to speed the development. Simulators (and emulators) use software to mimic real-world conditions. Modeling tends to be done mostly within the second stage, and application development often uses modeling throughout the design, build, and test stages.

Figure 3 — Engineered Development Flow

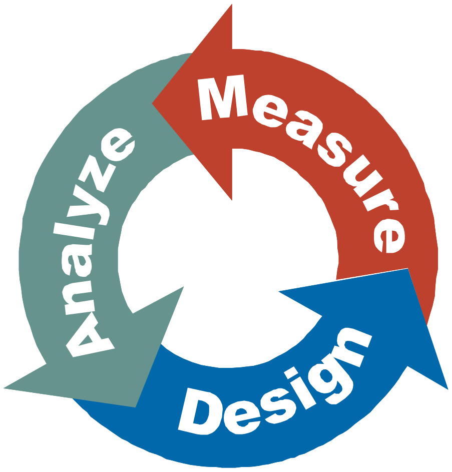


Source: META Group

Performance Optimization in Operations

Operational strategies for performance optimization involve relentless analysis of performance data and then action being taken on the results. This action will include feedback to infrastructure developers or application developers. Architecture is also affected by these optimization exercises. The operational aspects of optimization are embodied in capacity planning, which drives adjustments in capacity to regulate performance by comparing performance needs to the ability to meet those needs. Performance metrics are collected and tracked against improvement goals (see Figure 4). Analysis of new methods or technology enhancements should be performed to determine the feasibility of approaching these goals and eventually reaching them.

Figure 4 — The Optimization Cycle



Source: META Group

Another useful method to coerce performance optimization seems obvious, but it is frequently ignored. Common performance problems can be analyzed to identify optimization opportunities. By eliminating the root cause of a performance headache, overall performance improves.

Full performance optimization goes a step beyond traditional capacity planning by modifying not only infrastructure and applications, but also the very processes we employ for infrastructure and application development. Process flows, individual tasks within the processes, and organizational culture all profit as opportunities for incremental improvements are discovered. Over time, these “tweaks” add up to substantial efficiency gains and cost savings.

If IT organizations truly wish to demonstrate business value (and all IT organizations should), it is critical to record, track, and advertise the success of each performance optimization victory. Most increments will appear insignificant, but the cumulative rewards will be hefty. Trends will prove IT value by categorically displaying forward progress. The secret to successful operations-driven performance optimization is culture. The technology is the easy hurdle. Entrenched methods and organizational politics are the bane of operational efficiency and overall performance of infrastructure, applications, and the IT organization itself.

Collaboration Across Organizational Boundaries

A traditional wall exists between development and operations, especially for applications. IT organizations must dismantle this barrier. Parties on each side of the wall have not communicated well and a culture of animosity has hampered system releases. This wall must be torn down to provide effective applications and services through an amicable collaboration between system developers and system operations staff. The industry is rich with rhetoric about collaboration, but truly collaborative organizations have been elusive. As IT organizations mature, collaboration is achieved in small steps. Each small step accelerates the cultural shift toward a streamlined, collaborative environment that improves quality, enhances business value, and saves money.

The removal of organizational barriers is being driven by new technologies that demand more cooperation. J2EE application environments are a good example, where accelerating development cycles are stressing the ability of operations to manage the ever-increasing flurry of complex systems. The operations staff typically does not possess sufficiently deep knowledge of the J2EE applications and must depend on developers to assist in the operations processes. In addition, business dependence on these J2EE applications is escalating. Business requirements mandate tighter cooperation between IT organizational entities. This collaboration between development and operations is a good catalyst for acceleration, since necessary technology investments are low. Simple changes in processes compel changes in the human element of the organization, and this profoundly stimulates the cultural shift to organizational excellence.

Dealing With the Future Complexity Explosion

Today’s complexity will explode as IT systems become even more dynamic. Already, the impact of multi-tiered, Web-based applications involving complex assemblies of distributed software and hardware components is clear. These systems are becoming more intertwined with business execution and automation, increasing a company’s dependence on them. Due to the enormous risk exposure to the company, extreme care is required in development to minimize this risk.

Two technology developments now in their early stages are triggering the next phase of complexity escalation. Web services and the various adaptive organization initiatives (e.g., IBM’s autonomic computing, HP’s adaptive enterprise) promise to increase system distribution, but even more notably, they will result in highly dynamic environments that morph and adapt internal relationships to changing conditions. The fluid nature of these relationships will change how we plan, develop, manage, and operate future IT systems. Web services will also necessitate the implementation of these processes beyond the single organization and to trading partners; this already occurs, but at the level of message passing and interfaces.

Fortunately, some time remains for organizations to prepare for this onslaught of complexity. By streamlining processes and organizational culture now, IT organizations can adapt to dynamic system shifts incrementally. Organizations that wait will be ambushed by these developments, and trying to make adjustments too late will be agonizing or impossible. Organizations must prepare now or become irrelevant tomorrow.

Bottom Line

IT infrastructure and application development must follow structured engineering processes to instill discipline in the organization and optimize performance, and therefore value to the business. Performance optimization is the responsibility of everyone in the IT organization, from the architecture and development teams to the operations group.

Business Impact: Organizations must employ structured discipline to accumulate improvements in technology and in IT operational investments.