

Enterprise Application Performance Management: An End-to-End Perspective

By Vishy Narayan

With rapidly evolving technology, continued improvements in performance management and establishing baseline metrics is important to the sustenance of key infrastructure elements

INTRODUCTION

Performance management of enterprise applications is key to achieving business objectives and maximizing returns on IT investments. In addressing application responsiveness and scalability, the goal should be to build predictable, adequate performance into systems by considering quantitative behavior from a system's requirements stage through to its maintenance and enhancement.

Due to increasing system complexity, rapidly evolving platforms, shorter time to market and inadequate quantitative models and tools, performance management often represents the most challenging aspect for enterprise IT everywhere. Continued improvements in performance engineering and establishing baseline metrics is important in these days when enterprises have to move at the speed of business.

THE ECONOMICS OF ENTERPRISE IT INFRASTRUCTURE DOWNTIME

Increasingly, enterprise wide business applications enable the language of commerce for companies of all sizes. While companies invest heavily to implement enterprise software, less attention is paid to its maintenance after deployment. Thus post deployment, performance issues continually affect overall computing and operational efficiency. In fact, application problems are the single largest source of IT downtime. According to the analyst firm Gartner [1] estimates are that 40 percent of unplanned downtime is caused by application issues.

According to market research specialist Infonetics [2], large enterprises typically fritter away anywhere between 2 percent to 16 percent of their annual revenues because

of losses associated with network downtime. This translates into hundreds of millions of dollars annually. In a recent survey conducted by Infonetics in five different verticals (finance, health care, transportation and logistics, manufacturing, retail) researchers found that some are affected more than others. Finance and manufacturing verticals were the most affected, with the average financial institution experiencing 1,180 hours of downtime per year, costing them 16 percent of their annual revenue, or \$222 million, and manufacturers are losing an average of 9 percent of their annual revenue.

END TO END APPLICATION PERFORMANCE

Understanding performance engineering and management is essentially an experimental approach to predicting the likely performance of systems. It can involve building and then monitoring a typical system, under the workloads of interest. System modeling is another method which helps in reproducing the time dependent behavior of an unrealized system. This is either driven by a trace of the expected workload or by a workload represented as a set of random variables also known as stochastic modeling. Different elements of performance engineering practices should be integrated with design methods and demonstrated in operations that performance requirements can be achieved without sacrificing other desirable design qualities such as understandability, maintainability, and reusability.

With the growth in computational complexity and its reliance on high performance platforms and networks, continued improvements in performance engineering and establishing baseline metrics is important to sustenance of key infrastructure elements. Equally important is the recognition of emerging new architectures

for distributed management and control planes. Modeling performance, rather than following a “test and tune” approach, allows an enterprise to reduce IT fire-fighting costs by identifying problems upfront and minimizing the probability of application outages.

Typically, managing application performance is a crisis-driven task. During implementation cycles for packaged and custom developed applications little or no attention is paid to actual end-to-end performance scenarios, taking into consideration only the typical software engineering lifecycle management. Thus, post-deployment performance challenges go largely unanswered due to cost constraints in making changes to the application environment. Hence, focus now shifts to maintenance and incremental functional improvement, as opposed to proactive performance management and optimization.

A system-level view of the enterprise infrastructure encompasses host platforms, which include their hardware, operating system (OS), application software and the connectivity mechanisms, all connected to the external environment by the ubiquitous “communication networks.” The “performance management” triangle [Figure 1] shows the relationship between applications, constraints due business requirements and user expectations and the performance elements that drive information architecture to meet these requirements and business goals.

End-to-end performance involves host system characteristics such as memory, I/O, bandwidth, and CPU speed; the OS; and the application implementation. To maintain throughput levels, the host system must be able to move data from the application buffers, through the kernel, and onto the network interface buffers at a speed faster than that of the network interface.

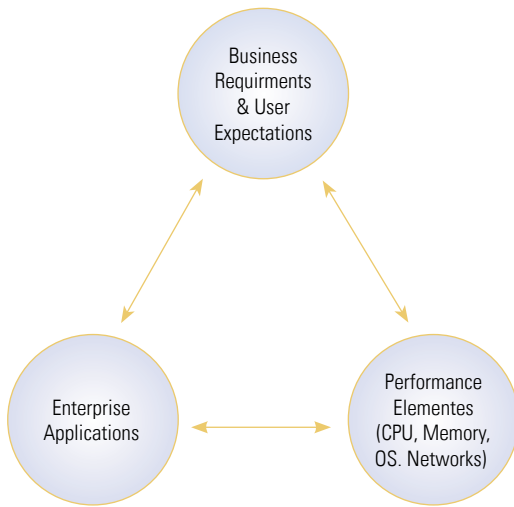


Figure 1: Performance Management Dependencies
Source: Infosys Research

With constant improvements in technology, with faster processors, larger amounts of memory, and accelerated disk transfer rates many computing systems are no longer hardware limited. Most operating systems are gaining the capability to drive networks at speeds of up to several gigabits per second. Increasing complexities in applications, system platforms and networks increase the challenges of understanding and delivering a measurable performance in overall operational throughput.

END TO END PERFORMANCE MANAGEMENT

Companies small and large have long realized the serious problems of applications performance and its effect on the bottom-line. The need of the hour is to have a unified team of application architects, network architects, system administrators working with business managers to understand how the company's technology infrastructure supports the business and to chart out the joint response to

troubleshoot and provide resolution to any performance problems. In addition to providing day-to-day operational solutions, long term proactive approach to address scalability and other issues should be addressed. Capacity planning, network re-engineering, volumetric analysis for enterprise data and storage requirements should be periodically addressed and updated. It is thus evident that enterprise performance management needs to address across the entire technology stack for it to be effective. Hence application performance engineering leading to an optimal performance metric should strike a balance between the business requirements, application architecture and the operating environment to deliver an application landscape that is highly reliable and manageable without compromises.

AN APPROACH TO ENTERPRISE PERFORMANCE ENGINEERING AND MANAGEMENT

The challenge in managing system performance is mainly due to the increase in size and complexity in the enterprise application environment and the need for application integration. User expectations of application response times and quality of service issues are also growing at a non-linear rate. Achieving an optimally performing application environment is a process that addresses performance management proactively as a part of the complete life cycle of application development - during design, development, production and maintenance lifecycle of systems [Figure 2]. At every step in the above stages is a corresponding stage in performance management.

We can define a three step process to effectively design, plan and execute performance engineering and management initiative for an enterprise.

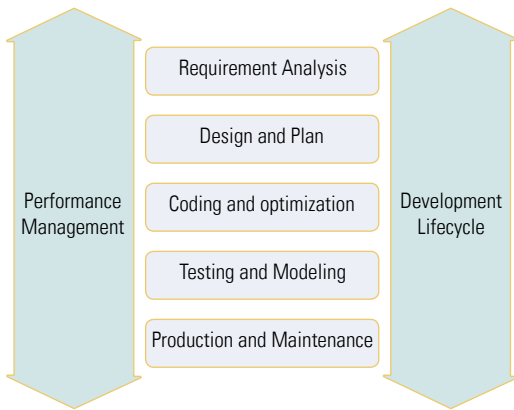


Figure 2: End-to-end view of performance management
Source: Infosys Research

I. BASELINE MEASUREMENTS

Baselining or , “if only we’d known where we were, we’d probably have had an easier time going somewhere else” is key to understanding and defining performance parameters and metrics. A baseline is a profile which can be used for purposes of comparison at those times when performance is not acceptable or is degrading. Networks and critical applications should be baselined both individually and during concurrent operations. Measurements for various scenarios will only strengthen the value of the baseline data.

Baseline measurements require fewer metrics and a longer sampling interval than crisis measurements, because a particular problem is not being investigated. Instead, the goals are to characterize system performance only, and to watch for trends. Baseline measurements can be used to review performance trends over time, compare against current performance when investigating or diagnosing current performance problems, provide data for performance forecasting and to develop and monitor service level agreements. Baseline measurements provide very important data points for capacity

planning and should be archived for historical purposes.

Data points can be established with baseline measurement data for comparisons to crisis hour measurements. While crisis measurements are made for performance problem diagnosis and require much more detail so that performance problems can be determined, broad baseline metrics can help in quickly identifying performance problem areas.

II. UNDERSTANDING APPLICATION DATA FLOWS

Enterprises in general store large amount of data over a period of time. To drive maximum benefit and harness valuable business data that resides in disparate systems across the enterprise, companies need to be able to integrate relevant applications. Such successful integration requires a thorough understanding of existing data structures and data flow throughout the application environment. Visibility into data sources makes data movement visible, starting from the macro view of sources and targets through to designing and creating the data flow to reduce errors and increase productivity. This process is key to application integration efforts.

The above capability can be used in application performance measurements analyzing applications at the flow level and arrive at a set of statistics detailing application and server level performance numbers, round trip times between source and destination data flows and network service related metrics. This helps maximize efficiency and minimize response time by interpreting flow-level data into actionable information.

III. IDENTIFICATION OF THE PROBLEM

When performance issues arise, service

professionals need to determine the nature and cause of the problem. Several systems in the monitoring and management space can aid in identifying the precise area of the problem (e.g., network management systems, network probes etc.). With appropriate network tools one can identify which applications are consuming network resources and when they are being used.

With knowledge of all network and application resources within the IT infrastructure, the next step is to closely monitor the performance of key business-critical applications. Understanding how an application's performance changes over time, under various circumstances, and how it interacts with other applications historically, can help in identifying the source of the problem. Analyzing packet-level network data and flow based application performance data from various points on the network can further drill down the problem focus area.

Most generally the problem can be due to:

- Poorly designed network infrastructure, incorrect configuration or bandwidth bottlenecks
- Inefficient and incorrect use of network protocols e.g., data packet sizes, incorrect window size etc.
- System hardware and/or operating system or database issues.

Any changes in system and/or network configurations must be documented and if there are considerable changes to the above then baseline measurements must be repeated and appropriate correlations noted.

UNDERSTANDING BEHAVIORAL PATTERNS

One other important aspect of performance engineering and management is in understanding the effect of network performance on two broad types of application categories---Communication and Transactional Applications. Knowing how the network affects these applications and under what circumstances, usage and time-of-day patterns, typical bandwidth utilization and geographical scope are key to gaining further insight to performance related issues.

- **TRANSACTIONAL APPLICATIONS:** These are commonly used enterprise wide applications centered around the core business activity. Supply Chain Management (SCM), Enterprise Resource Planning (ERP) and other such end-to-end business process applications fall under this category. These applications typically have request/response data flows and such flows increase as the number of transactions increase. Response times are very crucial in these circumstances as business depends on it.

- **COMMUNICATION APPLICATIONS:** These are typical day-to-day "information exchange" type of applications between employees and external entities across geographical areas crossing the enterprise network boundaries. Some examples are email, instant messaging, voice-over-IP (VoIP). While network bandwidth is a key requirement other network parameters such as delay, packet loss ratios and jitters all lead to poor performance.

Thus taken in totality and adding up instances across all enterprise applications, we can see

that application and network resources need to be properly maintained, balanced and continuously fine tuned to arrive at an optimum level of application performance to achieve the primary goal of meeting business and user requirements.

As the enterprise environment is constantly changing due to dynamic business needs, a rigorous set of best practices needs to be instituted and continuously monitored to obtain a consistent level of performance characteristics.


CONCLUSION

As the enterprise grows so does the complexity, size and reach of the business critical applications. These create new challenges in sustaining an optimum level of performance to meet business requirements and future scalability. Timely investments in an organization wide effort to initiate and institute a performance engineering and management activity will pay rich dividends over time.

An approach of incorporating performance management process continuously through the application development, production and maintenance lifecycle is the most effective

method to achieve and meet performance metrics to sustain the business. Further a three step approach to identify and resolve performance issues enables reaching the goal much easier. With the growth in computational complexity and its reliance on high performance platforms and networks, continued improvements in performance engineering and establishing baseline metrics is most important in these days when enterprises have to move at the speed of business.

REFERENCES

1. Software Quality in a Global Environment: Delivering Business Value", Gartner Report, September 2004.
2. The Costs of Enterprise Downtime: North American Vertical Markets 2005, Infonetics Research, January 2005.
3. Strategies & Issues: Measuring End-to-End Internet Performance, Rich Carlson, et.al, IT Architect, April 2003.
4. Optimizing Linux Environments for Performance and Scalability, IBM Global Services White Paper, November, 2003.
5. Convergence of Fault and Performance Management, Network General White Paper, March 2006. 

Author in this issue

VISHY NARAYAN

Vishy Narayan is a Principal Architect with the System Integration practice, Infosys. He has several years of experience in implementing large scale IT infrastructure projects involving various network technologies, systems and protocols. Prior to Infosys, he spent several years at NASA Ames Research Center, building wide-area communication networks and was a technical lead for the NASA Information Power Grid (IPG), which provided global grid services to the research community. He can be reached at vishwanath_narayan@infosys.com.

For information on obtaining additional copies, reprinting or translating articles, and all other correspondence, please contact:

Telephone : 91-80-41173878

Email: SetlabsBriefings@infosys.com

© SETLabs 2006, Infosys Technologies Limited.

Infosys acknowledges the proprietary rights of the trademarks and product names of the other companies mentioned in this issue of SETLabs Briefings. The information provided in this document is intended for the sole use of the recipient and for educational purposes only. Infosys makes no express or implied warranties relating to the information contained in this document or to any derived results obtained by the recipient from the use of the information in the document. Infosys further does not guarantee the sequence, timeliness, accuracy or completeness of the information and will not be liable in any way to the recipient for any delays, inaccuracies, errors in, or omissions of, any of the information or in the transmission thereof, or for any damages arising there from. Opinions and forecasts constitute our judgment at the time of release and are subject to change without notice. This document does not contain information provided to us in confidence by our clients.

Infosys[®]

POWERED BY INTELLECT
DRIVEN BY VALUES