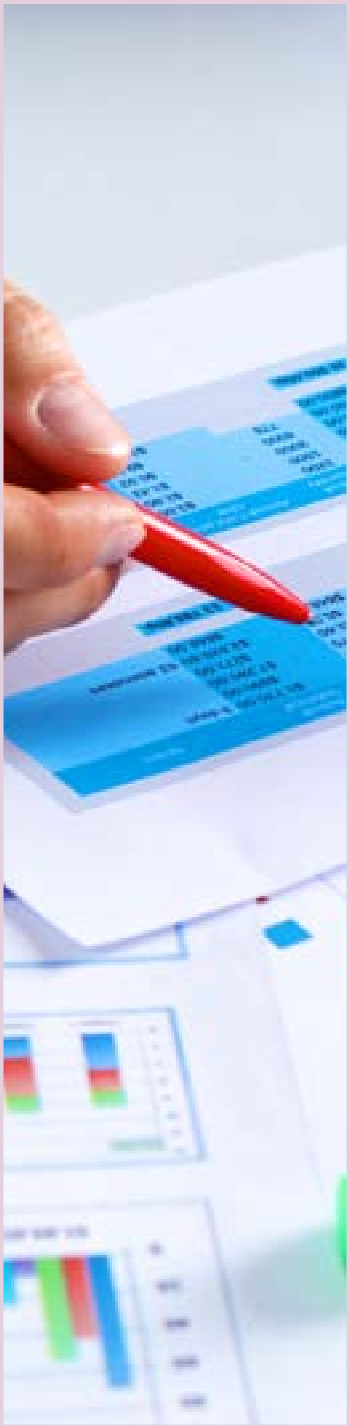
PERSPECTIVE



EFFECTIVE CAPACITY MANAGEMENT WITH MODELING AND SIMULATION - ASSISTED PERFORMANCE TESTING

Abstract

In this competitive marketplace, businesses seeking to maximize profitable outcomes need to ensure their information technology (IT) systems are robust in order to deliver uninterrupted application availability. This is significant for enterprises as high performance of business applications is an underlying driver of business growth. Accordingly, the top priority for system administrators is to maintain the health of their servers in order to meet stringent application performance Service Level Agreements (SLAs). However, several organizations are suffering due to a reactive approach to monitoring servers and applications only in the production phase. Thus, performance issues can be addressed only after they have occurred, often impacting operations and business.

To address this challenge, organizations need performance testing to detect issues before the production phase in the Systems Development Lifecycle (SDLC). However, to be effective, testing needs to be performed continuously and should consider changes in workload and performance owing to business growth and transitions. Organizations need a proactive approach that can deliver a timely and holistic view of an application's performance, thereby mitigating instances of application failure and overcoming performance issues and server downtimes. This paper discusses such a proactive approach that enables organizations to benefit from predictive and continuous performance and capacity management of their applications and infrastructure.

Infosys®
Navigate your next

Introduction

In today's connected world, the performance of information technology (IT) systems has a critical role play in shaping the success of an organization. System administrators are required to execute multitudinous operations to monitor and improve the performance of all individual components of business applications. Some of these activities include monitoring the system performance/workload, overcoming business outages or poor performance, and system/infrastructure upgrades with quick turnaround times1.

In such a situation, performance testing is critical to ensure that application end users do not face performance issues in the production environment. However, typically, performance engineering exercises that identify and resolve performance-related issues are reactive in nature. Further, the scenarios for performance testing change continuously according to business growth and these transitions are not considered during performance testing exercises. Today, performance engineering exercises are evolving from reactive to more holistic and proactive methodologies2. The gradual transition is described in Figure 1 below.

To successfully meet system performance expectations, businesses need to consider future workload through capacity management exercises3. Currently, such exercises are unable to deliver timely inputs regarding deteriorating system performance and infrastructure upgrades. Thus, to sustain optimum growth, businesses need to adopt mechanisms that can continuously monitor capacity management and system performance, identify errors and accordingly take corrective measures. This can be achieved by linking proactive performance testing techniques with continuous capacity management activities in a holistic manner.

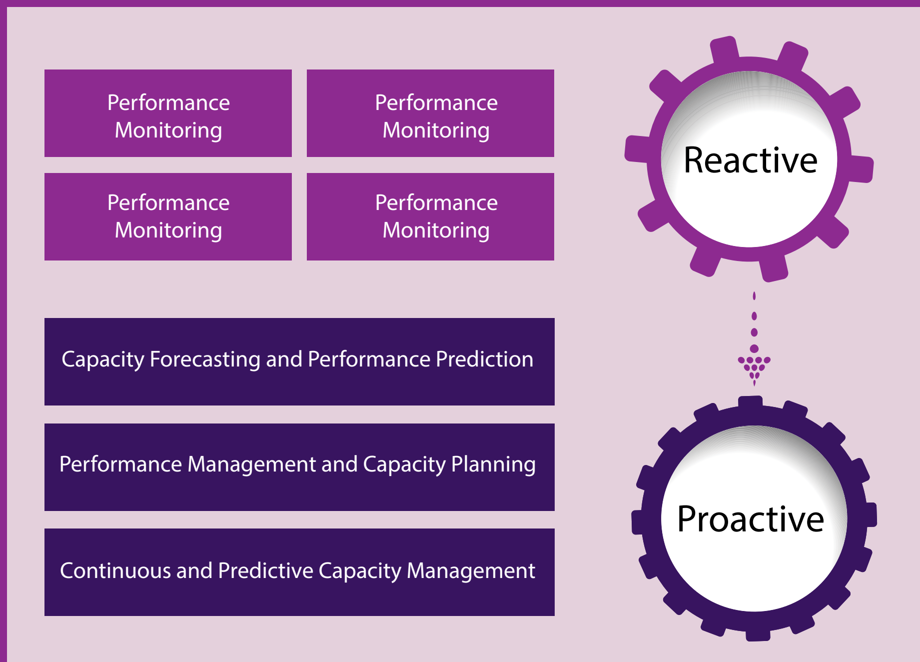


Figure 1: The Reactive to Proactive Transition of Performance Engineering Activities

What is Capacity Management?

A MET A Group study states that capacity planning is "the most important issue for large enterprises with employee strength of more than 1000 people". The study further states that the key goal for Capacity Planning and Management is to provide a holistic and measurable view of the IT environment and to describe its relevance in business terms4. To achieve this goal, organizations need to proactively manage IT resources with clear inputs from application and infrastructure planning groups.

According to the Information Technology Infrastructure Library (ITIL), capacity management is defined as 'the discipline that ensures IT infrastructure is provided at the right time, in the right volume and at the right price, ensuring, at the same time, that IT is used in the most efficient manner'. Additionally, ITIL finds that a typical capacity management exercise consists of application sizing, performance and workload monitoring, resource and demand forecasting, and modeling5. Currently, a typical capacity management process involves several steps as described in Figure 2 below6.

Many enterprises believe that capacity management processes should be run only once a year. However, ITIL states that for capacity management to be effective and successful it should be continuous and involve measurement, analysis, prediction, and tracking of performance and capacity[5, 7].

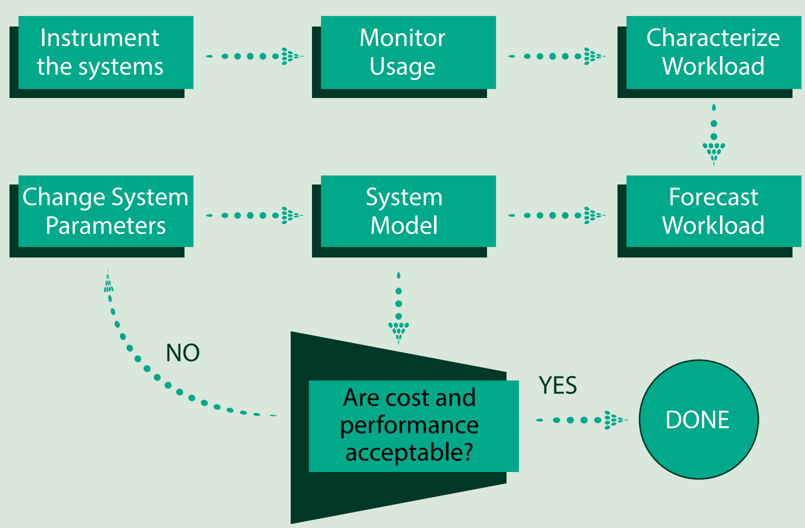


Figure 2: Current Capacity Planning Process

Continuous and Predictive Capacity Management – The Proposed Methodology

Continuous and predictive capacity management enables capacity planners to share inputs for infrastructure upgrades prior to testing. This methodology provides mechanisms that track application use, monitor changes in usage over time, forecast trends, and proactively take corrective actions. The key steps involved are as mentioned in figure 3.

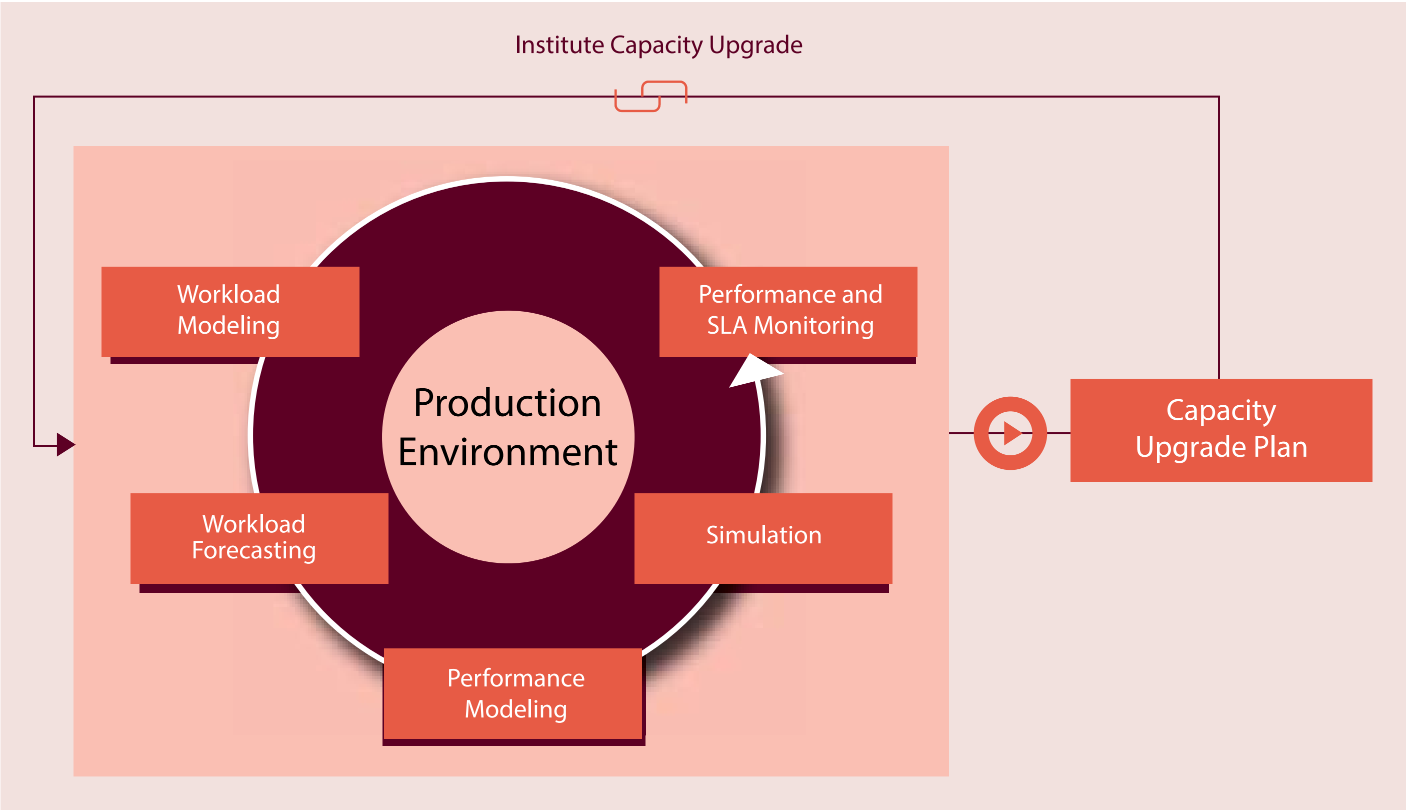


Figure 3: Continuous and Predictive Capacity Management Methodology

- The workload modeling tool monitors the usage of applications in the production environment and tracks the variation in usage over time.The historical usage data collected can be used to predict future changes in the system.
- A simulation engine is used to predict system performance in the current production environment. The engine comprises system performance models and is built using the performance test results and the forecasted workload. In cases where the system performance does not meet the pre-defined Service Level Agreements (SLAs), proactive corrective measures are taken before an issue actually surfaces.
- The performance model of a system is always built in line with the actual system in the production environment. It is designed by continuously monitoring the use of infrastructure resources and the performance of the system. If deviations are found between the performance model predictions and actual system performance, the performance model is modified to mirror the changes in the actual system.

Information Workflow in Proactive Performance Testing

Monitoring the production environment provides data on system workload and performance. The current workload of a system can be obtained by analyzing log files on a web server. A large set of this workload data can then be used to predict workload. Such predictive forecasting also considers possible changes in workload owing to business growth.

Performance testing is conducted before a system moves into the production phase of the Systems Development Lifecycle (SDLC). The results from the performance testing as well as system architecture and infrastructure details are used to build the initial performance models for the test environment. This modeling and simulation-enabled performance testing is further used to strengthen the test results and provide informed recommendations to augment the performance of an application. If required, these models can be revised to overcome the differences between the test and production environments using industry benchmarks.

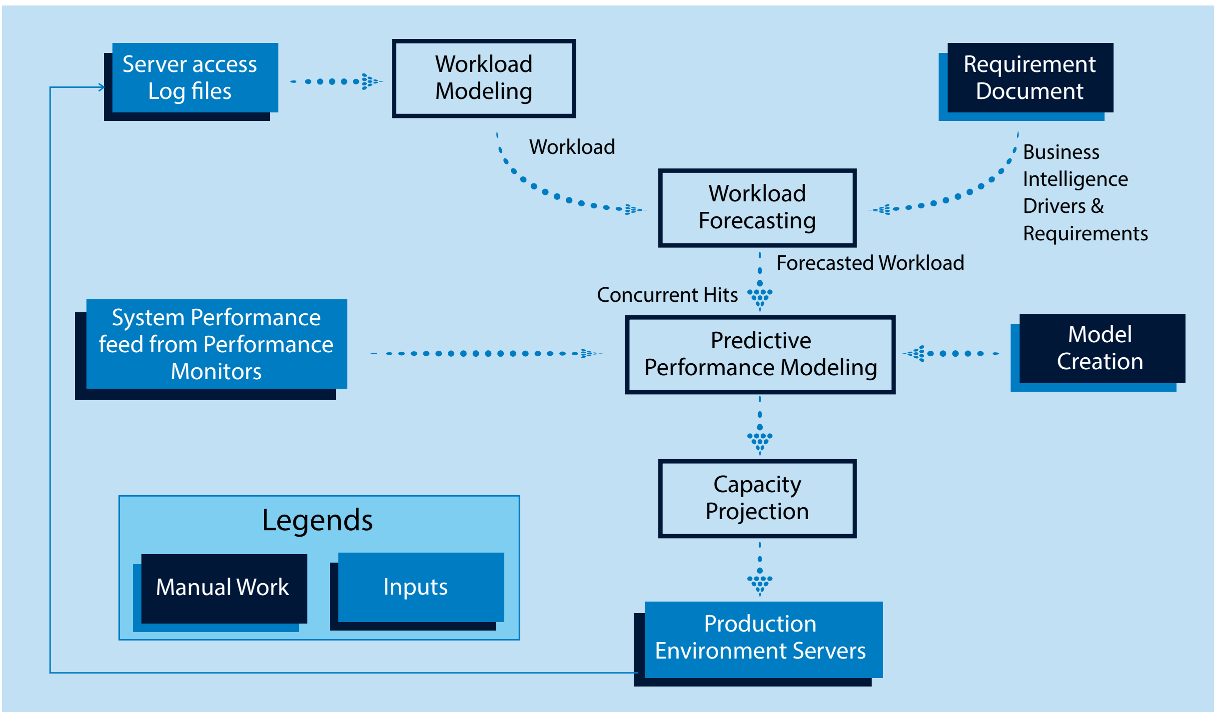


Figure 4: The Information Workflow in a Continuous and Predictive Capacity Management Methodology

The projected workload is an input for the performance model and is used to predict performance under the modified workload. If the hardware and software performance shows deviations from the pre-defined SLAs, the system capacity is upgraded accordingly using data from multiple rounds of performance modeling and several configurations. The system environment can then be continuously monitored to ensure SLA adherence and to correct the performance models.

The proposed solution methodology drives significant benefits and helps organizations to:

- Analyze the effect of external changes such as workload growth on the performance of a business application. This helps minimize risk, improve quality and enhance end user experience.
- Upgrade system infrastructure through a monitoring, forecasting and performance modeling process. It enables system administrators to address infrastructure requirements for consistent system performance over a period of time and deliver upgrades when required.
- Maximize their return on investment (RoI) for infrastructure and ensure uninterrupted application availability, thereby helping them achieve their business objectives.

Conclusion

Organizations today need uninterrupted application availability in order to ensure sustained profitability. They require solutions that can improve their existing performance testing capabilities to ensure superior system and application performance. With the aim of improving testing by ensuring robust predictive and management mechanisms, the continuous and predictive capacity management methodology can be leveraged to deliver advanced capabilities such as timely tracking, measurement and analysis of performance models before migrating to the production environment. Such a proactive methodology enables businesses to accurately predict and adjust system performance in the actual environment with minimum risk of errors, thereby ensuring a better experience for the end user. Further, system administrators benefit from instant visibility into system issues and ongoing requirements for infrastructure upgrades. The solution methodology ensures consistent business application availability, thereby maximizing RoI, enhancing the end-user experience, and ensuring superior system performance.

References

1. "Analyzing Computer System Performance with Perl::PDQ" by N. J. Gunther, Published at Springer-Verlag GmbH, 2005.
2. "Performance Engineering of software systems" by C.U. Smith, Published at Addison Wesley, 1990.
3. "Model Based Performance Prediction in Software Development: A Survey" by Balsamo, S., Di Marco, A., et. al. and published at IEEE Trans. Software Engg., pp. 295-310, Vol 30., No. 5, May 2004.
4. "How Does Your IT Organization Measure Up to Current Industry Wide Spending and Performance Metrics", Meta Study, November 2003.
5. "ITIL Capacity Management", Web Source: http://www.itlibrary.org/index.php?page=Capacity_Management
6. "The Art of Computer System Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling" by Raj Jain, Wiley, April 1991
7. "Information Technology Infrastructure Library" Web Source: http://en.wikipedia.org/wiki/ITIL#Capacity_Management

About the Author

Nidhi Tiwari

is a Senior Technical Architect with Infosys Labs, the research wing of Infosys. She has over ten years of experience in varied software technologies. She has been working in the performance engineering and cloud computing for six years. Her research interests include adoption of cloud computing, cloud databases and performance modeling. She has authored papers for many international conferences, journals and has a granted patent as well.
She can be reached at nidhi_tiwari@infosys.com

Amit Gawande

works as a Technology Lead at Infosys Labs. He has over five years of experience with a focus in the areas of performance engineering methodologies and cloud computing. His research interests include performance modeling and simulation techniques.
He can be reached at amit_gawande@infosys.com

For more information, contact askus@infosys.com

Infosys®
Navigate your next