The Application Of Modeling And Simulation In Capacity Management Within The ITIL Framework

Sonya Rahmani; Otto von der Hoff
SRA International, Inc.
sonya_rahmani@sra.com, otto_von_der_hoff@sra.com

Abstract. Tightly integrating modeling and simulation techniques into Information Technology Infrastructure Library (ITIL) practices can be one of the driving factors behind a successful and cost-effective capacity management effort for any Information Technology (IT) system.

ITIL is a best practices framework for managing IT infrastructure, development and operations. Translating ITIL theory into operational reality can be a challenge. This paper aims to highlight how to best integrate modeling and simulation into an ITIL implementation.

For cases where the project team initially has difficulty gaining consensus on investing in modeling and simulation resources, a clear definition for M&S implementation into the ITIL framework, specifically its role in supporting Capacity Management, is critical to gaining the support required to garner these resources. This implementation should also help to clearly define M&S support to the overall system mission.

This paper will describe the development of an integrated modeling approach and how best to tie M&S to definitive goals for evaluating system capacity and performance requirements. Specifically the paper will discuss best practices for implementing modeling and simulation into ITIL. These practices hinge on implementing integrated M&S methods that 1) encompass at least two or more predictive modeling techniques, 2) complement each one's respective strengths and weaknesses to support the validation of predicted results, and 3) are tied to the system's performance and workload monitoring efforts. How to structure two forms of modeling: statistical and simulation in the development of "As Is" and "To Be" efforts will be used to exemplify the integrated M&S methods. The paper will show how these methods can better support the project's overall capacity management efforts.

1. Introduction

ITIL is a best practices framework and set of guidelines that define an integrated, process-based approach for managing information technology services. Translating the ITIL theory into operational reality can be a challenge. Methods of implementation and best practices using ITIL principles are out of scope for this paper. Rather, this discussion aims to highlight how best to integrate modeling and simulation into ITIL implementations.

A clear definition for M&S implementation into the ITIL framework especially its role in supporting Capacity Management is critical to gaining customer and stakeholder buy-in. In the case example, discussed later in this paper, the team had difficulty gaining consensus on investing in modeling and simulation resources. The benefits of modeling and simulation were unclear to the project's overall mission and as a result there was insufficient modeling resource allocation. However, once M&S was tied directly to the system's Capacity Management activities as part of ITIL, the M&S efforts gained traction. Lessons learned from this case example have been leveraged as part of developing this paper's thesis.

The successful implementation of M&S within ITIL will encompass the following characteristics: 1) use of at least two or more predictive modeling techniques, 2) methods complement each one's respective strengths and weaknesses to support the validation of predicted results, and 3) techniques are tied to the system's performance and workload monitoring efforts.

2. ITIL BACKGROUND

ITIL encompasses a set of concepts and policies for managing information technology infrastructure, development and operations. ITIL consists of the following five disciplines (illustrated in Figure 1):

- Service Strategy
- Service Design

- Service Transition
- Service Operation
- Continual Service Improvement

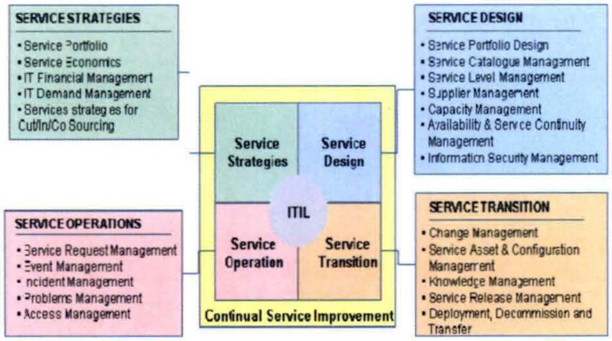
SERVICE STRATEGIES
• Service Portfolio
• Service Economics
• IT Financial Managemert
• IT Demand Management
• Services strateg es for Cut/In/Co Sourcing
SERVICE DESIGN
• Service Portfolio Design
• Service Catalogue Management
• Service Level Management
• Supplier Management
• Capacity Management
• Availability & Service Continuity Management
• Information Security Management
Service Strategies
Service Design
ITIL
Service Operation
Service Transition
Continual Service Improvement
SERVICE OPERATIONS
• Service Request Management
• Event Management
• Incident Management
• Problems Management
• Access Management
SERVICE TRANSITION
• Change Management
• Service Asset & Configuration Managerrent
• Knowledge Management
• Service Release Management
• Deployment, Decommission and Transfer

Figure 1: M&S Integration into ITIL Framework

2.1 M&S and the ITIL Framework

The scope of Service Design includes the design of new services, as well as changes and improvements to existing ones. Service Design consists of several areas; however, for purposes of this discussion, the focus will be on the Capacity Management area.

2.2 Implementing M&S Using ITIL Framework

Capacity Management is the discipline that ensures IT infrastructure is provided at the *right time* in the *right volume* at the *right price* and is used in the most efficient manner.[1] The real success lies in implementing an integrated M&S approach that 1) encompasses at least two or more predictive modeling techniques, 2) complement each techniques' respective strengths and weaknesses to support the validation of predicted results, and 3) is tied to the system's performance and workload monitoring efforts.

For system development and deployment projects that are still in early operational stages, additional model validation challenges may arise from the lack of a scalable Performance Test environment or a full system monitoring solution thereby limiting access to actual performance data. Using at least two types of modeling techniques can help to overcome this early validation challenge by raising confidence in model results where general agreement is obtained using disparate modeling techniques. In addition, the combination of M&S methods can successfully deliver capacity forecasting flexibility for both large and small scale projects.

Projects with the following characteristics will most benefit from an M&S implementation tied to ITIL principles:

- Clear-cut performance analysis goals
- Strict Service Level Agreements (SLAs) or Operational Level Agreements (OLAs)
- Enterprise class applications
- Volumes experiencing significant growth
- Time-based mission critical or real-time systems
- Lack of a full-scale performance test environment (need for alternative system evaluation techniques)
- Cost sensitive capacity requirements
- Long lead-time resource acquisition

The M&S implementation should be driven by definitive goals for evaluating system capacity and behavior given clearly stated performance requirements. The M&S implementation team likewise needs to be equipped with performance analysis and engineering expertise together with target system subject matter knowledge. Furthermore, the project's ITIL framework should be tailored to tie M&S to the following ITIL activities: Monitoring, Demand Management, Performance Tuning and Application Sizing.

3. A CASE STUDY

A case study on a federal IT system is used below as an example to illustrate M&S implementation in ITIL's Capacity Management processes. The federal system contains over 100 million records and processes close to 50 million requests annually. In addition, the system specifically meets the program characteristics in Section 2.2 mentioned above.

All these factors underscored the need for a robust and flexible capacity management program. As a result, a formal Capacity Management Process was created using the ITIL framework. The ITIL framework was tailored to support the federal system's overall Service Delivery and Service Support functions. In creating the Capacity Management processes, the project implemented modeling and simulation activities as a set of integrated activities. Figure 2 illustrates the M&S relationship central to Capacity Management within the program's ITIL process framework:

[1] ITIL Open Guide. March 2, 2009. < http://www.itlibrary.org>

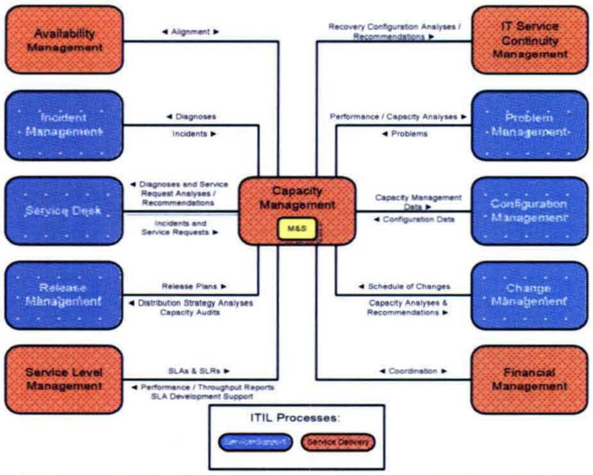
Availability Management
◄ Alignment ►
Recovery Configuration Analyses / Recommendations ►
IT Service Continuity Management
Incident Management
◄ Diagnoses
Incidents ►
Performance / Capacity Analyses ►
◄ Problems
Problem Management
Service Desk
◄ Diagnoses and Service Request Analyses / Recommendations
Incidents and Service Requests ►
Capacity Management
M&S
Capacity Management Data ►
◄ Configuration Data
Configuration Management
Release Management
Release Plans ►
◄ Distribution Strategy Analyses Capacity Audits
◄ Schedule of Changes
Capacity Analyses & Recommendations ►
Change Management
Service Level Management
SLAs & SLRs ►
◄ Performance / Throughput Reports SLA Development Support
◄ Coordination ►
Financial Management
ITIL Processes:
Service Delivery

Figure 2: M&S Central Relation to Capacity Management within Enterprise ITIL Framework

As part of this implementation, M&S activities were joined to several ITIL activities (as described below):

- Monitoring – system performance data (e.g., resource utilization metrics, response times, throughput, etc.) and workload monitoring (e.g., arrival patterns, transaction volume, etc.) were collected and analyzed from both the Production and Test environments. M&S uses these data to build and update the models.
- Demand Management – M&S applies stochastic abstractions and transaction volume models to workload impact analyses.
- Performance Tuning – M&S supports project efforts to identify steps required to handle current and/or new workloads to optimize system performance or operational policy.
- Application Sizing – M&S supports identification of resources needed for a new system application or a change to existing application. For example, model results provide input into hardware acquisitions required for new system deployments.

One of the most significant factors that contributed to the success of the program's Capacity Management Process was the tightly integrated M&S implementation within the project's overall ITIL framework.

The ITIL framework references four modeling techniques and includes Trends Analysis, Analytical Modeling, Simulation Models and Baseline Models. This paper classifies both Trends Analysis and Analytical Modeling as forms of statistical techniques. In addition, Baseline Models are defined in the context of a simulation model, and defined as a "benchmark" of the current ("As Is") system performance.

This case example illustrates that it is the combination of both statistical and simulation modeling techniques that directly support making the program's Capacity Management Process a success.

3.1 M&S Techniques in Case Example

A combination of statistical and simulation model techniques were used to quantify performance, estimate capacity, provide subject matter input, and afford validation to the overall modeling activities. Statistical techniques included:

- Trending using ARIMA (Auto-Regressive Integrated Moving Average) models for time series data – these methods were used to support characterization of existing system workloads and forecasting of future growth patterns based on historical volumes.
- Analytical model development efforts – these were used for several different needs including deriving mathematical expressions of system workloads to characterize workload arrival patterns and critical resource capacity models. In addition, historical transaction data were also analyzed to identify key performance factors and develop reusable statistical descriptions of the system's behavior.

Figure 4 illustrates typical transaction workload regression trending models for two classes of system transactions. The blue line depicts historical data whereas the red line represents the regression predictions. The use of ARIMA modeling techniques suitably captures the temporal characteristics of workload seasonality as well as year over year background growth where present.

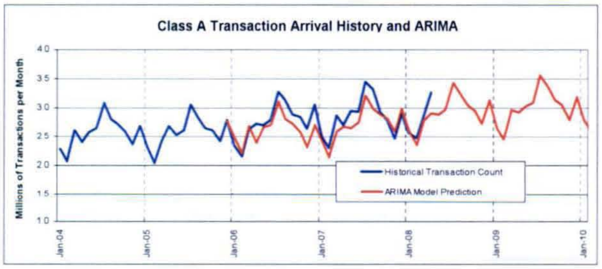

Figure 3: Transaction Arrival Trending Models

Statistical modeling strengths includes the ability to apply relatively simple methods that require shorter turnarounds to getting answers as well as requiring less detailed input data; weaknesses include a higher risk of being less accurate for predicting response times and throughput, loss of predictive accuracy where future behavioral patterns vary substantially relative to historical patterns, and an inability to deal with queuing and resource contention analysis.

Simulation modeling is used to gain more accurate predictive results for response time, throughput and resource consumption. The simulation modeling techniques included:

- "As Is" simulation model development efforts that craft simulation models of the existing systems and validate against the performance in the production environment (baseline model which "benchmarks" the current system).
- "To Be" simulation models that leverage the "As Is" models to develop the anticipated views (i.e. future operating conditions).

Simulation modeling strengths include providing the capabilities for more accurate projections of system throughput and response times in support of hardware acquisition estimates and architecture validation efforts and the ability to predict and analyze dynamic queuing properties and resource contention conditions. Simulation modeling weaknesses can include requiring a longer turn around time and large volumes of detailed output performance data. Valid use of the simulation model results will depend on the accuracy of the performance data used to develop the models.

However, where used in collaboration, the two differing modeling techniques can be combined to generally support a broader set of performance analysis needs and introduce flexibility in satisfying the project's capacity management objectives.

3.2 Developing the "As Is" Models

Early on, one of the biggest challenges was lack of production monitoring on the legacy system components. The project had an urgent need for precise simulation model results; however, most of the legacy system lacked any performance monitoring tools that would correlate workload to resource consumption (e.g., CPU, disk reads/writes, etc). As a result, the integrated M&S methods were tailored to tackle these challenges by modeling parts of the system as a "black box" and using a combination of statistical and simulation techniques.

The statistical analysis encompassed evaluating historical performance data (such as response time and throughput) to characterize statistical latency distributions under no queuing conditions. These techniques were used to combat the lack of instrumented performance data on specific pieces of the system. For these components, historical response data were analyzed to identify a time where there was little or no queuing in the system. During these periods, the start and finish times of each transaction were collected and used to create a histogram illustrating resulting service times. The histogram data were used to build the best fit curve characterized as a probability distribution. Thereafter, the team used the distribution to represent the system service time in the simulation model. Figures 4 and 5 below illustrate the histogram of response times under no queuing conditions for Production and the Simulation Model.

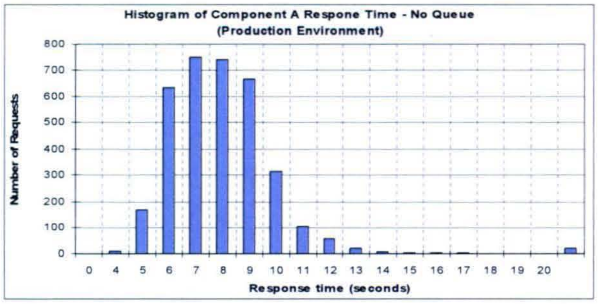

Figure 4: Histogram of Component A Response Time - No Queue (Production)

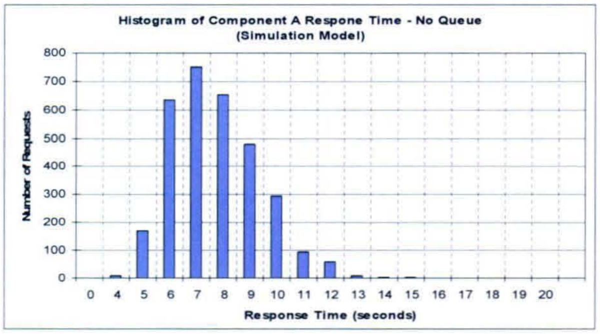

Figure 5: Histogram of Component A Response Time - No Queue (Simulation Model)

Although, the model simulation was not able to entirely capture the long tail observed in Production data, the associated statistical data demonstrated that there was little difference in overall response time between the simulation and production data results (see Tables 1 and 2 below).

Table 1: Production Statistics

Mean	7.87
Median	8
Mode	7

Standard Deviation	1.79

Table 2: Simulation Model Statistics

Mean	7.72
Median	8
Mode	7
Standard Deviation	1.65

After validating service times, the queuing behavior was analyzed using a time period starting with an empty queue that gradually built over time. The service request arrival times were also assessed for that period. The simulation model was run with the statistically derived service and arrival time models. Figure 6 depicts the validated simulation results:

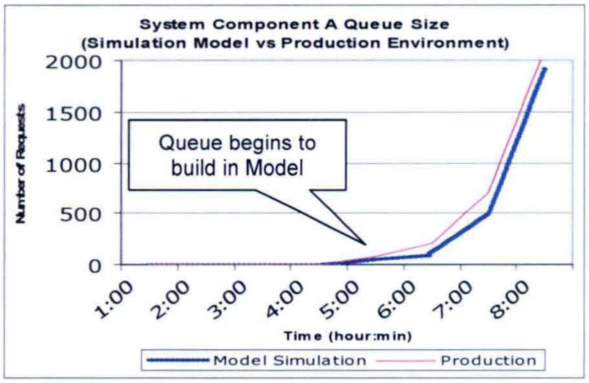

Figure 6: Simulation Model vs. Production Environment Queue Validation

The team compared the model's simulation results to production data in order to successfully validate against the true system performance. In this manner, the team was able to leverage two different modeling techniques to successfully build the "As Is" simulation model. The statistical analysis facilitated service time characterization in a manner that could then be applied in the simulation models. This would not have been possible without these statistical models due to lack of production performance data. In addition, if we had used statistical techniques in isolation, we would not have been able to vary response time and correlate this to queuing behavior over the course of a day.

Simulation models were subsequently updated once production monitoring tools had later been deployed. Collected performance data were evaluated using analytical techniques to associate resource consumption with the workload executed (viz. CPU, database reads/writes, etc). The simulation model was validated under full workload conditions by comparing results (response time, throughput, and CPU consumption) to the production environment. Production changes (e.g., new code deployed, architecture or platform changes, etc.) could then be quickly rendered in the simulation environment by leveraging monitored data against the validated baseline "As Is" model.

Implementing two different modeling techniques therefore proved critical to performing capacity management early in the system development lifecycle when performance data were not yet available. Model accuracy was in turn improved after production data became available.

3.3 Leveraging "As Is" to Forecast Impact of New Workloads

The program's Capacity Management forecasting responsibilities includes regular engagement with the system stakeholders to identify workload changes that may impact the IT system's performance and computational resource needs.

A recent workload addition of several million records exemplifies the important role M&S played in the Capacity Management process. The M&S team worked closely with the Demand Management office to characterize the new workload's yearly demand based on historical behavior of similar historical service request types. The team used statistical regression models to predict future seasonal arrival patterns and adapted existing workload distributions into daily workload arrival patterns for the new transactions. Finally, the "As Is" model was simulated with the new workloads. The team provided analysis on expected response times, throughput, and resource utilization plus impacts anticipated to existing workloads.

Figure 7 below illustrates an example of resource utilization forecasted data.

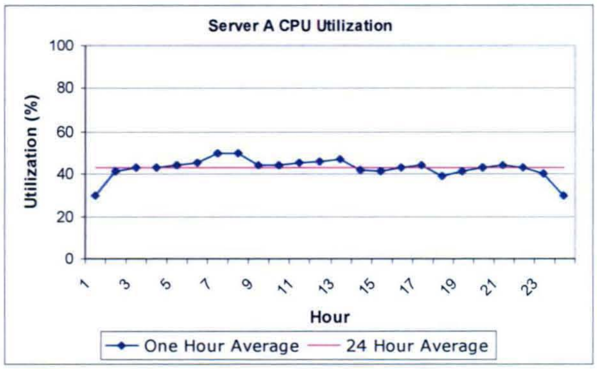

Figure 7: Simulation Model – Forecasted Server A Utilization

Figure 8 below illustrates an example of forecasted system response times. Adherence to SLA response times were of critical importance to the customer and program.

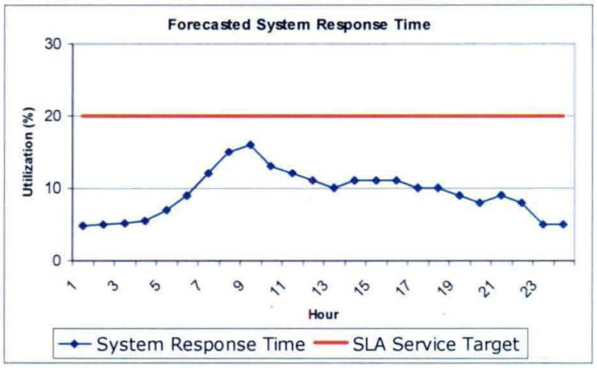

Figure 8: Simulation Model – Forecasted System Response Times

3.4 Leveraging "As Is" to Develop "To Be" Models

Recently, the government system went through a massive modernization effort that upgraded both its hardware and software components. The customer expressed several concerns on how this would impact operations and most specifically SLA adherence. An M&S Tiger Team was therefore tasked to develop simulation models that would help forecast computational resource requirements to deliver needed capacity and to justify capital equipment acquisitions. Of additional concern were possible impacts to the front-end business processes and wide area network performance.

The M&S Tiger Team's objective was to develop an end-to-end analysis solution that would provide an impact analysis on all three aspects of the business. On the back-end system, specific questions were raised on identifying impacts to resource consumption and response times. For the latter, the back-end "To Be" system model was built leveraging the "As Is" simulation model described in Section 3.2 above. The resulting analysis assembled a comprehensive picture of the new system deployment impacts.

Performance analysis helped to proactively identify specific impacts and areas for operational improvement to ensure a smooth transition during system modernization. This was one of the most successful initiatives on the project demonstrating the critical insight that can be gleaned from using a combination of modeling techniques.

4. CONCLUSION

In conclusion, the development of an integrated modeling approach can significantly impact the success of the project's overall capacity management efforts. The M&S implementation should encompass two or more predictive modeling techniques, complement each one's respective strengths and weaknesses to support the validation of predicted results, and be tied directly to system performance and workload monitoring efforts.

The implementation should include evaluation of the "As Is" system as well as forecasting techniques. The models developed in support of the latter's analysis should provide estimates for response times, throughput, and resource utilization for the "To Be" system. Furthermore, models should be designed to guide the project's hardware acquisition and architecture validation efforts. From the beginning, the ITIL framework should be tailored to implement M&S within Capacity Management processes and relate to the following activities: Monitoring, Demand Management, Performance Tuning and Application Sizing activities.

Following these high level guidelines will establish and promote a successful Capacity Management Program for a broad array of enterprise IT application systems.