A Framework for Enterprise Application Performance

Net Forecasts – Peter J. Sevcik
BCR Volume 33, Number 11
November 2003

With computer and network applications operating or providing critical support to nearly every aspect of an enterprise, an important subset of the enterprise IT market, called Enterprise Performance Management (EPM), has emerged to ensure that all the aspects of a business perform properly. The EPM market contains four categories of business intelligence and analytics technologies: financial, sales and marketing, supply chain management and human resources. EPM is a $15 billion market in 2003, with about $2 billion spent on software alone.

But, if EPM watches the enterprise, what watches EPM? The cold reality is that the applications that monitor business performance must themselves be monitored for performance, and I'm not the only person who sees the need for this capability. Roughly 30 companies have products to measure application performance and another 70 offer products to improve it. Moreover, nearly all the suppliers of information software, hardware and systems claim to supply a capability that improves performance in one aspect or another; some claim to be a complete performance management or improvement elixir.

However, such claims are often confusing, inconsistent, conflicting and incomplete, and those are the good ones! Thus the need for a framework to organize solutions and properly understand claims.

Goals of a Framework

There are two approaches to defining performance: Create a narrow definition that includes some approaches/functions and excludes others, or take a broader view, with a definition that gives many if not all players a role. I prefer the latter, as my objective is to provide a way to understand the many views of performance and to help enterprises make more informed choices.

To meet these objectives, however, it is important to lay out some goals for the framework. It should be:

Comprehensive Cover all the aspects of performance; each major aspect is a performance function that is correlated to an application.
Clear Define the functions without using the word "performance"; the sum of all functions equals performance.
Uniform Define metrics for each function/application pair that are appropriate and normalized so they can be compared across functions. In other words, the same score in two functions should be equivalent.
Useful The metrics must account for the different needs of different applications and be able to be measured. Define methodologies that can help make practical decisions.
Valuable The framework should ensure that information technology is supporting the business. Methodologies using the framework should show how performance is linked to business goals.

Performance Framework

There are two types of performance attributes: Asset management, which is concerned with improving the effectiveness of the assets that supply the application, and Experience management, which aims to improve the user experience with the application.

Asset management metrics deal with getting users onto the service and determining how many users the enterprise can support without either wasting resources or causing the application to fail. These metrics are associated with scaling the system, and often their analysis revolves around ROI for a new technology.

Experience management metrics, on the other hand, make the user want to use the system and be more productive. In this context, an analysis for new technology would focus on benefits like increasing global reach, speed to market, partner retention, lowering customer churn, real-time operations and increasing sales.

Clearly many factors address these benefits, but this discussion is limited to those that can be changed by technology. For example, within asset

management is a group of techniques that are used for deciding whether to buy, lease or outsource a technology. Although this is an important aspect of business performance, it is not a technology choice. Similarly, intangibles, like product quality and brand, which can't be changed with information technology, are outside this framework.

There is a soft boundary between understanding and rating what the application does to support a business goal versus how well it performs that duty. Application performance is anchored in the "how well" and approaches the "what."

Performance Functions and Metrics
Each function can be subdivided into aspects of performance, and each must be defined and measured. Then, they all must be integrated into a single function metric that will be meaningful to management and useful in performing technical tradeoffs. Some functions have generally accepted metrics, but most are still in development and none is defined in a uniform way so that they can be related across functions. The following are descriptions along with few definitions for metrics that operate on a scale of 0 (complete failure) to 1 (perfection).

I'll begin by describing asset management that requires no user input -- if all the boxes are working and connectivity exists, the system is operating properly and there's no need to ask the users if they agree.

Provisioning The system's ability to establish new service or recover failed service. This function includes discovery, topology maps, alarms, uptime, routing stability and fail-over. The traditional system availability percentages that many management systems calculate can be used.
Provisioning = min (availability of each asset used by a specific application), represented on 0-1 scale

Efficiency The system's ability to make the best utilization of the assets that provide the service. These include aggregate traffic, asset utilization, users per server and users per Mbps. Current utilization reports can be applied.
Efficiency = ave (utilization of each asset used by a specific application), represented on 0-1 scale

Protection The system's ability to protect itself from malicious or unauthorized use that would degrade the performance of the asset. The function includes the effectiveness of technologies such as firewalls, denial of service protection and VPNs. The devices that protect the system are in fact insuring performance continuity. The metric for this function will require converting actuarial risk assessment into an index of 0-1.

The experience management functions are application- and user-specific, so any rating must relate to an end-user's view of the application. A user group and application pair, like a remote office staff using a CRM application, define performance requirements. Therefore, talking to users is essential to any methodologies regarding these functions.

Accessibility The system ability's to provide access to authorized users. This is a measure of the user impact from issues such as local access, user connectivity and the effects of content filtering. Very little is known about the probability that a specific user population has connectivity to the services they want, when they want them. This will require new types of instrumentation and measurement, and the measures will need to be converted to a 0-1 metric.

Quality The quality of the technical aspects of the user's experience with the system. This function has different metrics that are unique to application classes. At this time, the only accepted standard is the Mean Opinion Score (MOS) for voice calls. This is a good metric that must be converted to 0-1 but it's probably not a linear translation as shown in [1].

Quality of transaction applications is the user's task response time. But acceptable response time is a highly variable and not bound to a 0-1 scale [2]. We propose a new Uniform Response Time Index (URTI) that converts the highly diverse time values into a single metric. The URTI would be determined based on the following equation:

URTI = (number of satisfied user + ½ number of tolerating users)/total user population

Users that see frustrating performance are not in the numerator but are counted in total user population. To learn more about defining satisfied, tolerating and frustrated performance, see references 3-5.

Safety The level of comfort that the user has when interacting with the system. For example, spam interferes with the email experience, popups interfere with the application experience, viruses disable the user's machine, keystroke recording changes the user's behavior and privacy concerns hinder the user's ability to enter data. Defining a metric for this new but important function is an open research topic.

Application Classes

Since the framework is application specific, one needs to list the applications that are under consideration by each function. For example, some performance tools are very narrow in scope and apply to only a few applications.

The pace (ping-pong or continuous) and direction (one-way or two-way) of information transfer are a good way to define fundamental application behavior. This yields to four applications groups: real-time, transactional, data feed and bulk data. Other fundamental factors like protocols and traffic volume further break the groups into 12 distinct application classes. The performance framework currently defines six functions that can be applied across the dozen application classes as shown in Table 1.

Table – 1 Application Performance Framework

		Asset Management			Experience Management		
		Provisioning	Efficiency	Protection	Accessibility	Quality	Safety
Real Time	Voice over IP						
	Video Conference						
Transactional	Terminal-Host						
	Client-Server						
	Web						
	Web Services						
Data Feed	Streaming Audio						
	Streaming Video						
	Telemetry						
Bulk Data	Email						
	Peer-Peer						
	File Transfer						

Using the Framework

The performance functions can conflict with one another, i.e., improving one often degrades others. With 30 interdependencies in the framework, optimizing the whole is a complex challenge. Methodologies for optimizing performance by balancing the metrics need to be developed, but they're more likely to be defined by MBAs than engineers.

Imagine the following scenario. An enterprise wants to improve its asset protection with a new security device (e.g., firewall, proxy gateway) to protect Web applications. A thorough review of the transactional Web performance functions will show that some functions improve while others degrade. Table 2 shows such a hypothetical analysis using uniform performance metrics for all the functions. The value of the protection index improved (higher) while four other indexes degraded (lower) and one remained unchanged. This gives the

enterprise a complete view of performance impacts when deciding on the new device.

While the total performance score in Table 2 is nearly the same with or without the new device, the protection score significantly improved and so the device should be purchased. But, if the total score were lower, the enterprise should look for a device with less adverse impacts on the non-protection functions.

Even though some metrics of the framework are not yet defined we can start using it, albeit with a simpler methodology. The next time a performance enhancement vendor makes a presentation, get them to tell you which of the 36 cells in the framework they are talking about. Once they select a few rows and one key column, challenge them to fill in the effects of their technology on the rest of the columns for the rows (applications) they selected. The simplest entry is a plus for improves, minus for degrades and zero for no effect. Of course, they will get a plus for the column they chose, but push them to explain all the aspects for how the technology works in order to understand the effects in the other columns.

Table 2 – Applying the Framework to a Security Change

	Before	After	Observation
Provisioning	.98	.95	adding another box lowered it
Efficiency	.60	.55	lowered user traffic
Protection	.80	.97	improved as planned
Accessibility	.96	.80	users have a harder time getting in
Quality (URTI)	.92	.85	response time suffers due to encryption
Safety	.80	.80	users get no security benefit

Invitation

This framework and metrics are a work in progress. I invite the performance community to a joint effort to make this approach a reality. There is a lot of work yet to be done on metrics and methodologies, and once a uniform set of metrics is defined, it will be necessary to develop models of business requirements and technology effects using the metrics. So the research work will continue for some time.

I am chairing the new performance track at the N+I Conference, May 9-14, in Las Vegas (www.interop.com), and plan to apply this framework in the technical program and live demos. I invite enterprises, vendors, academics, consultants and service suppliers to contact me regarding participation (peter@netforecast.com).

References

All of the references were published in *Business Communications Review* (BCR) magazine and are also available at the NetForecast Web site.

1. Sevcik, "The Pitfalls of Scaling VOIP," BCR, March 2002
2. Sevcik, "Web Performance – Not a Simple Number," BCR, January 2003
3. Sevcik, "Understanding How Users View Application Performance," BCR, July 2002
4. Christy, "This Is Your Father's Performance After All!" BCR, November 2002
5. Sevcik, "How Fast is Fast Enough?" BCR, March 2003

Peter Sevcik is president of NetForecast and is a leading authority on Internet traffic, performance and technology. Peter has contributed to the design of more than 100 networks, including the Internet, and holds the patent on application response-time prediction. He can be reached at peter@netforecast.com.