

One Size Fits All:

Is there a universal metric for IT success¹?

Simplicity is the ultimate sophistication.

Leonardo da Vinci

For every complex problem there is an answer that is clear, simple, and wrong.

H. L. Mencken

Wednesday, June 07, 2006

Alistair A. Croll

Coradiant



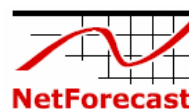
Dennis Drogseth

Enterprise Management
Associates



Peter Sevcik

NetForecast



Eric Siegel

Burton Group



¹ Or, as Dennis would prefer it, “a universal metric for effective application performance.” As this paper recognizes, the scope of IT challenges is too broad and we should limit ourselves to the topic of application performance measurement here.

Introduction

Measuring the performance of mission-critical applications is an increasingly challenging proposition. On the one hand, many complex factors influence whether an application is working correctly, and summarizing the overall health of a system inevitably reduces detail and makes the system harder to tune. On the other hand, non-technical people such as those in marketing and enterprise executives are more interested in an end user's view of applications that represent important sources of revenue or significant cost centers to their business. They don't care about internal technical details; they just want to know if the applications' users are satisfied with performance.

So IT teams face a challenge: Summarize the health of an application in a simple, easy-to-grasp manner while still retaining enough detail to make the information meaningful. One multi-vendor initiative, Apdex, tries to "roll up" service quality metrics into a single number that represents user satisfaction with a particular service when measured against expectations. Apdex can be a valuable tool, especially when combined with thoughtful weighting of the relative importance of various transactions to the enterprise.

The contributors

Recently², a group of industry analysts got together to discuss the viability of a single number and the reporting of IT metrics in general.

- **Alistair Croll** is co-founder and VP of Products for Coradiant, a vendor of web performance monitoring software. He is also the Interop faculty chair for the Application Networks and Webops tracks. He was a principal analyst at Networkshop, where he pioneered studies on SSL performance and load-balancing platforms. He is the author of numerous articles on performance, availability, and web technologies, as well as the book *Managing Bandwidth: Deploying QOS in Enterprise Networks* for Prentice-Hall. Alistair holds a B.Comm with honors and an advanced major in strategic marketing.
- **Dennis Drogseth** is a Vice President with Enterprise Management Associates where he focuses on Networked Services Management practice areas including performance availability and service management across enterprise and telecommunication markets. At EMA, Dennis researched performance/availability, integrated security, changing organizational dynamics in IT, and management issues shared between the enterprise and the service provider communities. With over 24 years of hands-on experience in a variety of leading networking firms, Dennis is also a featured columnist of the *Network Systems Management* newsletter for Network World Fusion and an author of featured articles in *Network Magazine* and *Business Communication Review*.
- **Peter Sevcik** is president of NetForecast, founder of the Apdex Alliance, 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. His Net Forecasts column in *Business Communications Review* magazine, tries to separate fact from hype and describe a vision of the networked future. Peter is a senior member of the IEEE and is also on the program advisory boards of the Interop and Next Generation Networks conferences.
- **Eric Siegel**, a Senior Analyst at the Burton Group, covers web and network performance optimization, measurement, management, service level

² At Interop Las Vegas 2006, <http://www.interop.com>

agreements, and QoS. He is the author of *Designing Quality of Service Solutions for the Enterprise* (John Wiley & Sons) and *Practical Service Level Management: Delivering High-Quality Web-Based Services* (John McConnell with Eric Siegel; Cisco Press) and a frequent speaker. Prior to Burton Group, Eric was Principal Internet Consultant at Keynote Systems; Senior Network Analyst at NetReference, Inc.; a Senior Network Architect with Tandem Computers; and held positions with Network Strategies, Inc. and the MITRE Corporation. Eric received his B.S. and M.Engr. degrees in Electrical Engineering from Cornell University and was elected to the Electrical Engineering honor society.

This joint article summarizes the elements of that discussion at Interop, with room for dissenting and often complementary viewpoints.

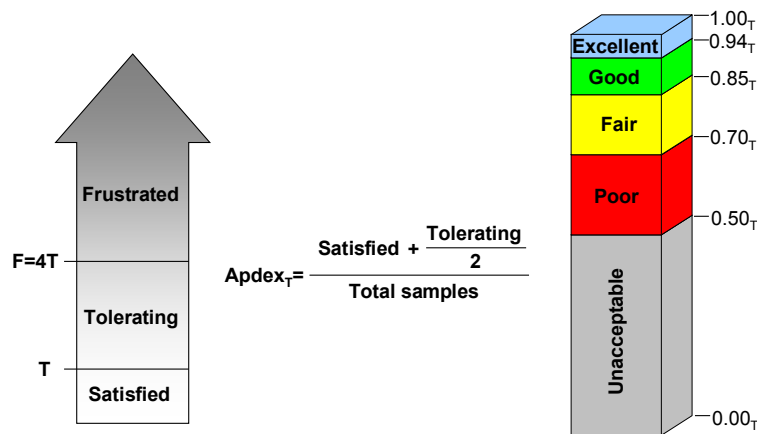
The big question

There tend to be three different requirements for any metric that a business wants to use. It must be **simple enough** to be understood by managers; **accurate enough** to withstand scrutiny; and both **usable and relevant enough** to be an effective agent of change.

Simple


A simple metric is one that doesn't require a lot of thought to understand correctly. Imagine, for example, a traffic light. It is quickly understood and conveys a meaning to most people. And it is only through easily communicated measurements of IT health that we can hope to set and meet targets for service quality.

One such approach is the Apdex index³ (illustrated below). This metric is derived from the number of satisfied, tolerable, and frustrating transactions that an application delivers to its end users. Apdex makes the assumption that the only real benchmark of health is end user experience; in this respect, it is a *symptomatic*, rather than a diagnostic, metric.



The Apdex formula is simple enough to be understood by non-technical managers and those responsible for arbitrating in disputes around service delivery. The manager's intuitive understanding of the Apdex number—that it represents the degree of user satisfaction—is correct, even if the manager doesn't know the precise details of how its internal weighting is

³ See <http://www.apdex.org> for more information



Croll: These three requirements—simplicity, accuracy, and relevance—are often in conflict. Accuracy undermines simplicity; and what's relevant to one person is incomprehensible to another because of the underlying context needed to understand it.

constructed. Its simple zero-to-one scale and standardized design helps managers compare performance across different transactions and systems.

But simple metrics like Apdex may not be as simple as they seem. They require that people understand the nature of the service; the tolerance definitions; and the way in which the measurements were collected. Small adjustments to tolerance definitions and to the data collection and processing system can dramatically affect the overall ranking of a service. Some have argued that single-number, formula-derived metrics are prone to “adjustment” by managers who want to provide a positive message and hide problems.

Therefore, simple metrics usually require a context; they’re not sufficient by themselves. A driver who sees a red light has a great deal of context about what it might mean. The light conveys meaning quickly and universally, but only to those who understand traffic laws and how drivers behave. Metrics need careful design and precise definition if they are to be useful and accepted by all. Stating the actual underlying metrics, while more complex than a single number, doesn’t leave any room for error.

Accurate

Metrics must be accurate; they must correctly portray the true situation without distortion by incorrect measurements or misleading statistical treatments. An incorrectly-scripted measurement probe may indicate that the measured application has failed, but that’s the fault of the measurement system, not of the application. Simple arithmetic averages of application response time or network delay are massively distorted by just a few outlying measurements; the situation for standard deviations is even worse. Measurement systems must be able to detect and tag questionable measurements for investigation. Medians and 85th percentiles, or the Apdex, are far better statistical treatments than simple averages.

A metric, whether simple or complex, is made up of many components, each of which has its own measurements of performance and availability. Summarizing these into a single metric without distorting the relative importance of each contributing component can be a challenging task, and doing so often requires statistics that are specific to a particular service—so they aren’t comparable across services or tools. It would seem, at least on the surface, that accuracy and simplicity are at odds. Nevertheless, each organization must do this work if the metrics are to be credible.

Usable and relevant

A third requirement is that no metric—no matter how simple or accurate—will produce a business benefit unless it is usable by the organization as an agent of change. And the only way to bring about that change is to educate the organization about the importance of performance.



Siegel: We need simple metrics, with complex underlying design. We don't want the measurements to be misleading and to impel an erroneous change.

The design should correctly weight the factors important to the enterprise and use appropriate statistical treatments.



Sevcik: Apdex also benefits users because the metric, generated by multiple systems, will produce the same result as long as the service being evaluated and the definitions of “tolerable” performance are consistent. In that respect, some of its simplicity comes from the fact that it can be compared across services and tools.



Siegel: “Accurate” isn’t the same as “relevant” or “not misleading.” It just means that the metric accurately portrays the underlying condition, even if what you’re measuring is actually irrelevant.

As more and more companies achieve nearly constant uptime (through load-balancing and redundancy) performance becomes the focus of most IT efforts. But less than half of companies measure end-to-end performance of their applications. Surely measuring relevant transactions and convincing the organization that improving them is worthy should be the most important goal of IT.

Knowing which transactions are relevant to the organization can be difficult. Customer-facing, revenue-generating applications might matter to revenues; but back-end operating systems might affect costs. Getting consensus among various divisions about what's most important is difficult. And each tier of an organization, from the low-level operator to the manager, director, or CIO, needs a different level of visibility. So any measurements of IT success must be tailored to the responsibility and role of the person that uses them.

Clearly, even "simple" metrics need complex, painstaking design, and even with much effort, it may be impractical or misleading to reduce all performance to one single metric. This makes them neither simple nor accurate across the organization.

Redefining the question

To decide whether there's a single metric for IT success, we need to understand *what success is*. Indeed, that's so broad a topic it's probably not a useful question. So let's look instead—as Dennis suggests—at whether there can be a single metric for effective application performance.

IT systems are ultimately in the service of end users. One way to determine success is to measure the number of users that are able to achieve their goals or complete their tasks in a satisfactory model. But there are other constraints we need to place on ourselves if we are to find a single, simple metric. Here are some of them.

Measure end-user quality of experience

Focusing on end-user QoE for specific transactions substantially reduces the scope of the question, as it limits us to the symptomatic measurements described above. The number of queries in a queue on a database isn't a measure of service quality. It's a diagnostic (for repair or tuning) or, at most, a measure of how well implemented and cost-effective something is.

Choose relevant metrics

Whatever transactions we're measuring, they should be relevant to the business. This seems fairly trivial; but many IT organizations watch diagnostic or implementation efficiency metrics that matter to *them*, not to the business. It's common to measure server reachability, or the responsiveness of a database to a particular query. But unless we measure the right things—such as the success of an application that delivers revenue to the business—we're collecting end-user experience that's not relevant. The result is miscommunication and misrepresentation about the health of the IT systems.

Weight metrics accordingly

Not all transactions are the same, and not all hours during the day or month have equal importance to the business. Some applications have backups: A customer support website



Drogseth: We didn't agree that there can be a single metric for IT success—the title of the original Interop session. Instead, we agreed that there can be a single metric for effective application performance. We also agreed that performance in the sense of response is just one of multiple metrics for measuring QoE.



Croll: End-user QoE can be derived from many sources, including sniffers, web logs, synthetic testing, and Real User Monitoring. The better the measurements reflect what actually happened to the customers of the service, the more relevant and politically powerful they will be.

might be accompanied by e-mail and telephone support. On the other hand, others—like a stock trading console—might be irreplaceable. Some downtime on a weekend for a particular set of applications may not greatly affect business, but other applications may be crippled if they can't run at that time.

Getting organizations to agree on these weightings is a challenge. But without a proper weighting of all the measurements we collect, we can't summarize properly and we can't help the operations staff prioritize incident management when multiple systems fail concurrently.

Do the right math

Once we've chosen the right metrics, properly weighted, we need to consolidate them into a single number. This requires the right math. Averages, for example, are notoriously bad at summarization. A few outlying measurements, which might be caused by a couple of lost data packets out of a million, will greatly distort averages for metrics that are usually in the tenths of a second, such as network round-trip delay. And a day with an hour of downtime—during which few transactions occurred—might seem, on average, to be a success; while a day of uptime with a few very slow transactions might seem like a failure.

Percentiles and counts of success such as those employed in the Apdex algorithm are good examples of proper summarization.

Define “acceptable” performance beforehand

The key to summarizing using a count- or Apdex-like model is to classify relevant, weighted transactions according to their success. And to do this, we need to define and communicate what “success” entails. In the case of Apdex, a satisfying transaction is one that completes successfully in a certain amount of time; a tolerable one completes successfully but takes too long; and a frustrating one is either incomplete or takes four times the tolerable threshold.

Explicitly stated and agreed-upon definitions of acceptable service are essential. But they're surprisingly hard to achieve. Often, it's more practical to deploy a service and measure it for a while before deciding what's a practical threshold.

Other factors, such as worker productivity or customer abandonment, may also drive a sense of what's acceptable. Users may also have psychological expectations of acceptable performance—a login page should be speedy, for example, but a detailed report or search might take a long time.

Summarize up

Different tiers of an organization have different responsibilities, and so need different data. A CIO might want to see a single number that indicates the health of all the IT systems; the web operations team might care more about the responsiveness of the web front-end. As we consolidate health and performance data within an organization, we need to be able to “summarize up” the metrics of subordinate organizations.

This means that the person in charge of the corporate website might have a single metric for that, with drill-down into individual pages. The director of web operations might get a single number for the health of the web system; with perhaps the ability to drill into the individual web properties that make up that number (such as the corporate website). But the CIO might have a single number which is composed, in part, of the weighted health of the web application.



Drogseth: There is a “double-edged sword” behind a single metric, as well, in that it can create a false complacency. In other words, masking technical complexity to the business consumer is good. Hiding human complexity from the IT service planner is bad.

Other factors

Consider also that when summarizing upwards, there needs to be some measure of how responsive the operational teams are. A CIO shouldn't worry if he or she knows that technical teams are addressing problems; but an unanswered issue is cause for concern. This suggests that a metric of unacknowledged problems, or of unacceptably long outages, should factor into summarization.

Conclusions

Our panel conceded that there can be a single, authoritative measure of IT success, as long as everyone understands that the measure will vary according to the role of the recipient within the organization and that it is a reflection of end-user health, rather than of the components responsible for delivering that health. The specific constraints are:

- That we're talking about end-user quality of experience as a definition of success
- That what we're measuring is relevant to the business
- That we're weighting the various measurements
- That our calculations are statistically accurate
- That we all agree what "acceptable" is ahead of time
- That we summarize and simplify up the organizational chart

IT operators need to engage the business in a conversation about service quality measurement and learn which transactions affect business health and how much. They need to get consensus around a definition of success, and they need to define how metrics will be simplified up the org chart.

As Dennis explains:

"IT's success is dependent on a multi-dimensional approach to understanding its business and customer needs. This requires, among other things, a dialog with the customer/consumer, whether through formally or informally structured programs for outreach to service consumers. Requirements such as flexibility of services in terms of mobility, or service choice, as well as cost effectiveness and guaranteed security levels also need to be brought into account. As a subset of that dialog, a single metric for application performance can be valuable by masking unnecessary technical detail in business-to-IT evaluations.

However, setting accurate and consistent metrics for interpreting 'satisfied' or 'frustrated' is where the real challenge is. Experience is by definition subjective and often difficult to communicate."

All this work can be a daunting task. But whether or not a company is ultimately able to achieve a single-metric approach, discussions about acceptable IT service performance are much needed.



Sevcik: The conversation is best performed as a formal "success definition" process with active participation by all the stakeholders of IT success.

The process must clarify the issues discussed in this paper, define success metrics, and set in motion procedures to gather data and report on the metrics.

However, this is not a one-time affair. The reports must be periodically reviewed and the process should be revisited annually. Real success depends on a continuous improvement mindset.