E-BUSINESS PERFORMANCE METRICS AND CAPACITY PLANNING

(CORRELATION OF BUSINESS ACTIVITY & PERFORMANCE METRICS WITH SYSTEMS METRICS IN E-BUSINESS ENVIRONMENTS)

Chris Leowski, Ph.D.
University of Toronto
Toronto, Ontario, Canada

May 2006

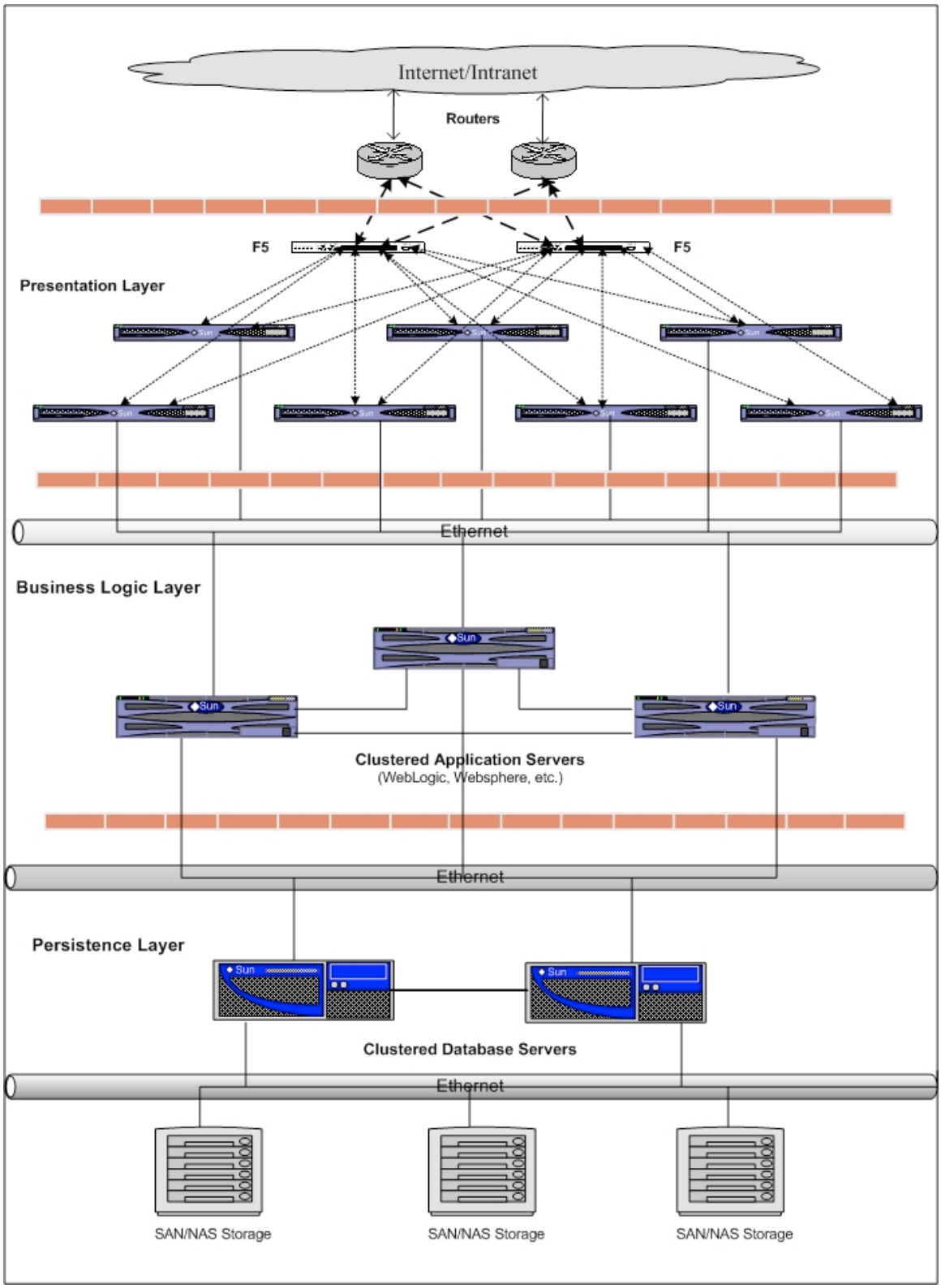
Abstract

Complex multi-tiered e-business applications are prone to performance degradation – whether temporary, caused by short-lived malfunctioning of various system, application or network components, or permanent and long-term, resulting from increased load or imbalances in the overall systems or applications architecture. In order to maintain robust and responsive e-business systems it is necessary to develop a set of quantitative measurements that would unequivocally provide both a performance benchmark and an early warning system. Although each e-business application requires normally a different set of business activity and performance metrics, the framework for building such metrics and correlating them with a set of standard systems metrics for the underlying IT infrastructure are common to most, if not all, applications. Based on the author's many years as IT director at the University of Toronto, and a systems consultant in large banks and brokerage houses, this paper describes such a framework, and provides examples of e-business activity and performance measurements and their correlation with the underlying IT systems performance measurements, with the purpose of developing methods and tools for benchmarking, monitoring, and capacity management, of e-business applications.

INTRODUCTION

Performance of e-business applications matters under most circumstances, and is absolutely critical in some – e.g. on line brokerage, foreign exchange – where under 0.5 sec. response time when requesting a particular operation (e.g. exchange rate update, stock quote, transaction posting) is normally required, and over 1 sec. response time is deemed unacceptable. Overbuilding the underlying infrastructure may not necessarily help, because business performance relies equally on the quantity of available system resources, as on their complexity, and on the quality of underlying applications. Look at typical high-end e-business systems architecture – as in Fig. 1 – to realize that although there may not be a single point of failure in a well-designed system, there may still be many points of potential performance degradation even without taking into account the quality of the code and robustness of APIs between application components.

Figure 1.


Internet/Intranet
Routers
F5
F5
Presentation Layer
Ethernet
Business Logic Layer
Clustered Application Servers
(WebLogic, Websphere, etc.)
Ethernet
Persistence Layer
Clustered Database Servers
Ethernet
SAN/NAS Storage
SAN/NAS Storage
SAN/NAS Storage

Both the science and practice of IT systems performance measurements have been evolving for decades and are well developed, but mostly limited to individual systems (e.g. benchmarks published by vendors and aimed at impressing IT users). Relying on such system performance metrics can be compared to driving a car with impressive performance characteristics, and still getting snarled in daily traffic. Businesses are not primarily interested in e-business topology, technical specifications of deployed hardware, and software characteristics of off-the-shelf or in-house developed applications. They are interested mostly – if not solely – in meeting requirements and expectations of business users. Since problems do occur, enterprise business units insist on developing and collecting measurements that would help in troubleshooting them. And since business volumes are rarely static over time, enterprise architects must have ways of planning for future capacity requirements.

The objective, then, is to measure activity and performance in business context, not in systems context – although the latter is related to, and should provide a solid base for, the former. And because they are related to, we end up focusing on sets of both systems and business metrics in order to help us solve our problems.

WHAT TO MEASURE

Any given e-business application runs within a specific hardware and software environment, sometimes exclusively dedicated to that application, more often shared with others. Even a dedicated IT environment often shares many IT components with others – for example, networks and network devices, security servers and security protocols and procedures, external data feeds, common log servers, service management personnel, etc.

There are generally two broad groups of metrics – systems metrics and business metrics - that can provide us with measurements essential to monitoring the current levels of e-business activity and to planning and executing adjustments in capacity levels of the underlying IT infrastructure, as well as in the applications themselves. Systems metrics have been used for decades by systems administrators, and a vast array of monitoring tools, from simple shell scripts to highly complex external systems costing millions of dollars, are routinely deployed by almost every IT organization. It is less obvious that standard systems metrics, particularly in e-business environments and when looked at without taking the entire e-business context into account, are often misleading and provide poor tools for both business performance measurement and for IT capacity planning. While there is a standard and fairly uniform set of systems metrics collected and analyzed by systems administrators, business metrics depend on the nature of e-business activity, and they can vary widely.

Business metrics can also be divided into two subgroups: **activity measurements** and **performance measurements**. A typical example of an activity measurement is a metric that captures the number of equity trades or foreign exchange trades executed on each business day, or the number of hits that an e-business web site records over a period of time. E-business performance measurements, on the other hand, are reflected by such metrics as the exchange rate propagation time (number of milliseconds elapsed from the posting of a new exchange rate to the moment that rate reaches the trader's workstation), or average, maximum, and minimum web server's response time, in milliseconds, to external requests. Service Level Agreements (SLAs) between IT divisions and business groups list expected activity volumes and their growth paths, but are mostly concerned with performance thresholds acceptable to business. This is understandable, because what appears as a slight variation in performance can often have significant effect on the business in question. For example, just one extra second in the exchange rate propagation time is the cause of significant competitive disadvantage to FX traders when compared to their competitors from other organizations. Average web site response time for active e-business web sites is usually set in SLAs to below one second, with – say – 3 seconds being the maximum. Activity levels, especially for new e-business applications, are more difficult to predict in SLAs – although once they reach certain levels they may have significant impact on performance measurements.

The accepted approach to business and systems metrics (see, for example, various papers by Ron Kaminski and Yipping Ding, listed in the References) is that they should correlate in order to be of interest for performance troubleshooting and/or capacity planning endeavors. On the systems side we have workloads (processes that consume resources) and on the business side we have CBMI – Candidate Business Metrics of Interest. Business metrics can be anything that accurately reflects the way business is expressed and measured – thus, they are difficult to standardize (businesses differ widely), often difficult to measure (there are no standard monitoring tools, like in the case of systems metrics), and there may be considerable disagreement between the players regarding their relative importance. Involvement of enterprise business groups in identifying these metrics is critical, but so is involvement of business consumers, whether internal or external.

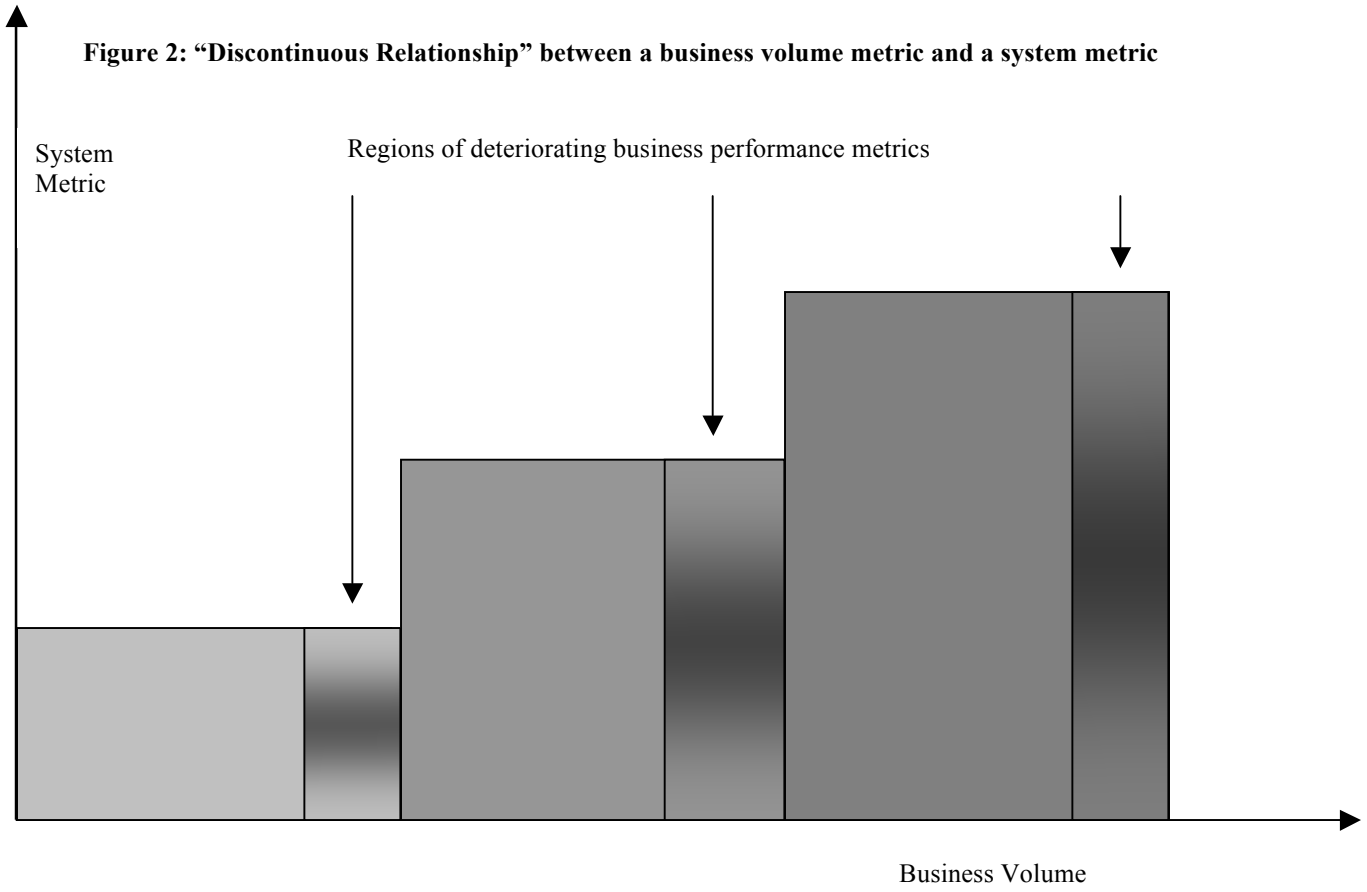
Unfortunately, the corollary of the requirement of statistically significant correlation between business and systems metrics is that those that do not correlate are of little value and interest. In fact, a lot of effort has been directed into developing methods and frameworks that refine the concept of correlated metrics and bring to surface relationships not necessarily obvious at first glance.

Although true in a vast number of cases, the above requirement is misleading and insufficient in a vast number of other cases – with an equally significant impact on business activity and business performance within the enterprise – as we shall attempt to argue in the rest of this paper. Our main premise is that the very concept of correlated metrics, as discussed in literature, is based on the notion of continuity of relationship between those metrics, that is – over the entire spectrum of measurements of A and B there is a statistically significant and measurable relationship. If this is not quite the case, we might exclude, say, 5% of outlying measurements from our calculations and thus strengthen our argument.

What we attempt to point out here is that there exists a class of discontinuous relationships between A and B, where A may be relatively constant over a sub-spectrum of B (signaling lack of correlation and – as *per* existing orthodoxy – lack of interest and importance), and then can become of critical importance once B moves above and beyond the previous sub-spectrum of its values. This differs from "linear subintervals" (Yiping Ding and Chris Thornley "On Correlating Performance Metrics") when correlation coefficients are low over the entire time interval, but fairly high over selected subintervals. Discontinuous relationships are not correlations (in the sense that A moves along with B one way or another), and yet B may critically depend on A (or *vice versa*). Thus, metrics can be highly interdependent and yet show almost no correlation.

The above phenomenon of discontinuous relationships between sets of metrics has become part of the measurements landscape with the advent of resource pre-allocation modes of operation and some modern programming paradigms that let applications run in memory containers. Anyone working for sufficient time with modern middle tier servers (e.g. WebLogic, Websphere) must have spent many hours analyzing and fine-tuning the underlying JVM. Predefined minimum and maximum heap allocation, and predefined internal heap structure within a JVM do have their justifications, but they easily break the correlation (though not dependence) between business workload and resource consumption as captured by standard data collection tools. A typical hypothesis (Kaminski) that the workload consumption should correlate with the volume of business functions may not hold anymore, replaced by a scenario in which resources show little correlation with the actual workload, and then – after reaching a certain level of the volume of business functions – the sky falls. Likewise, consider processes (e.g. relational database systems processes) that grab the entire available RAM, whether very small or very large volume of business work is being done. The dependence of business on such systems resources is unchanged, but the correlation between the volume of business functions and the use of those resources is broken – in other words, a typical case of discontinuous relationship, as illustrated in Figure 2. In that example there is no correlation between business volume metrics and selected systems metrics. However, when volume grows, we reach regions of deteriorating performance, where strong negative correlation between business volume and business performance metrics will be observed.

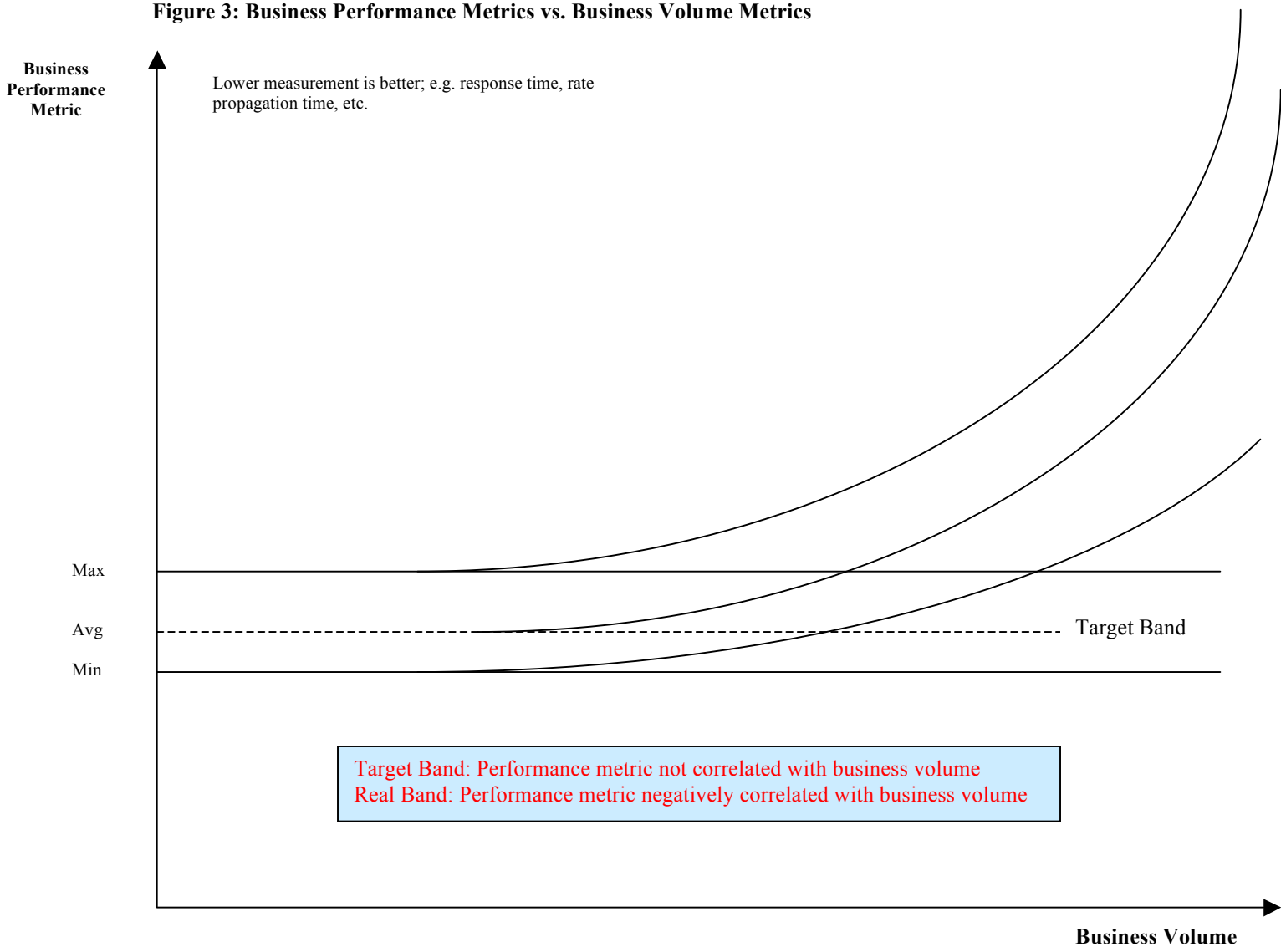
Figure 2: "Discontinuous Relationship" between a business volume metric and a system metric



CORRELATIONS

The question then is: when do correlations still matter? There are many situations where classic correlation between a business metric and a system metric is clearly measurable and statistically significant. This type of analysis is still used widely, for both troubleshooting and capacity planning. Its success depends on good selection of business and system metrics. In short periods of time, however, it is another type of correlations that become prominent – those between volume of business and business performance metrics. If the volume of business fluctuates widely or grows unexpectedly – whether intra-day or intra-week – the underlying IT infrastructure, and thus all IT resources available to business, are usually static in the short term. As indicated above, business functions of an enterprise are measured in terms of volume (e.g. of transactions) and in terms of performance, and in many businesses performance metrics are critical to the success of the business. One would normally assume that when volume of business increases, its performance metrics are prone to suffer – given the static IT infrastructure for any given short period of time. Thus, business performance metrics can be negatively correlated with business volume metrics (without even touching on systems metrics). Business units within the enterprise will insist on maintaining relatively static performance measurements (consistent with SLAs) when the volume is pumped up. Thus, their target can be expressed as a narrow horizontal band of a business performance indicator, with upper and lower lines indicating minimum and maximum measurements for that indicator at every level of business volume (Figure 3).

Figure 3: Business Performance Metrics vs. Business Volume Metrics



In practice, and consistent with my observations at a large capital markets enterprise, the performance indicator band not only does not stay horizontal, but standard deviation of performance measurements significantly increases with the volume of business, as does asymmetry between minimum and maximum measurements (in Figure 3 we assume that higher measurements reflect worsening performance – e.g. response time to a request for information or to transaction posting). It is that latter phenomenon of widely fluctuating performance indicators – even when average readings are perfectly acceptable – that triggers alarm bells, because it reflects many situations where performance thresholds are exceeded to such an extent (or with sufficient frequency) that complains start being escalated through the enterprise hierarchy.

Typical correlation scenarios are as follows:

(a) Case: static IT infrastructure with a fairly high level of resource utilization. Fluctuations in the volume of business normally have significant impact on business performance measurements,

while various IT metrics may reflect the ceiling of available resources. Significant negative correlation between business volume and business performance metrics is usually observed. Correlations between business metrics and various systems metrics may fluctuate widely, with some of them breaking off when resource utilization ceilings are reached for individual IT resources.

(b) Case: static IT infrastructure with high level of resource underutilization and evenly balanced initial overcapacity. Fluctuations in the volume of business do not usually have any major impact on business performance metrics (weak negative correlation between business volume and business performance metrics). Significant positive correlation between business volume and individual systems metrics is often observed. Since balancing of overcapacity is rarely perfect, certain deterioration of business performance metrics may occur, but is not deemed significant.

(c) Case: static IT infrastructure with unevenly balanced overcapacity. This is a typical case of one IT resource becoming a bottleneck at a certain level of business volume. Until that point is reached, the situation resembles the one described above in (b). However, once it is reached, negative correlation between business volume and business performance metrics becomes moderate to significant for individual pairs of metrics. At the same time, systems metrics correlate unevenly (including zero correlation) with the volume of business, with no correlation for subintervals when the IT resource is plentiful, as well as for those when its utilization has reached the ceiling.

(d) Case: dynamic IT infrastructure (various scenarios). Since it is usually a daunting task to measure requirements and build IT capacity in a manner that would guarantee perfect balancing of IT resources in time for increased volume of business (or to counter expected short peaks in business), certain degradation of business performance metrics (and thus moderate to significant negative correlation between business volume and business performance metrics) may take place in discrete time intervals - until another increase in a resource that created a bottleneck enters production. Various systems metrics correlate unevenly (from zero to high levels of correlation) with the volume of business within the same discrete time intervals.

BUSINESS INTELLIGENCE FRAMEWORK

Successful gathering and analysis of business and systems metrics, and their incorporation in business performance troubleshooting and/or in capacity planning, require a sophisticated enterprise level Business Intelligence Framework. That is, measurements must be collected from a vast array of systems and applications (e.g. servers, network devices, business apps, logs, databases) over comparable time intervals, with comparable frequencies, and stored in a data warehouse for further processing and analysis. A typical topology for such a framework is based on:

- grouping of systems (servers and other devices) based on their participation in, and relevance to, various lines of business. Such grouping is relevant, since various sets of hardware and software components may require different sets of metrics and/or different frequencies, and/or different time intervals. Furthermore, post processing of collected information and subsequent analysis of OLAP cubes or other analytical constructs is much easier and uses fewer resources when applications, processes, servers, and other components form logical groups in relation to businesses that they serve.

- deployment of a central monitoring system and of a central data warehouse for metric measurements. The central monitoring system acts as a brain for the entire operation of measurement collection – what data should be collected, where, how often, how should they be stored, with what frequencies various analytical processes should be executed, who should get the results of analyses, etc. For systems metrics, off-the-shelf applications are often available that are capable of automatically collecting, storing, and analyzing information. For business metrics – whether business activity or business performance – identification of metrics and collection of

measurements are mostly done by in-house processes integrated with the afore-mentioned central systems monitoring.

- deployment of data collection agents on all underlying hardware and software infrastructure. Such agents should be as passive as possible – restricted to silent capture of data generated within each system, minimal post processing, and periodic transfer of data to the central monitoring system.

- business measurements built into all relevant applications code. A well-written business application usually already contains switches that allow measurement and logging of various business processes. If not, application developers should be asked to add such monitoring components to the code, preferably in a way that allows turning on and off of various levels of granularity in capturing underlying business activities.

- deployment - within the central monitoring system - of processes that read various application logs, business databases, etc. at regular intervals (say, once a day during off business hours) and convert the extracted measurements into business metrics subsequently stored in the data warehouse and comparable to stored systems metrics (in terms of data collection frequencies and association with particular time periods). These are the business related ETL processes (extraction, transformation, and loading).

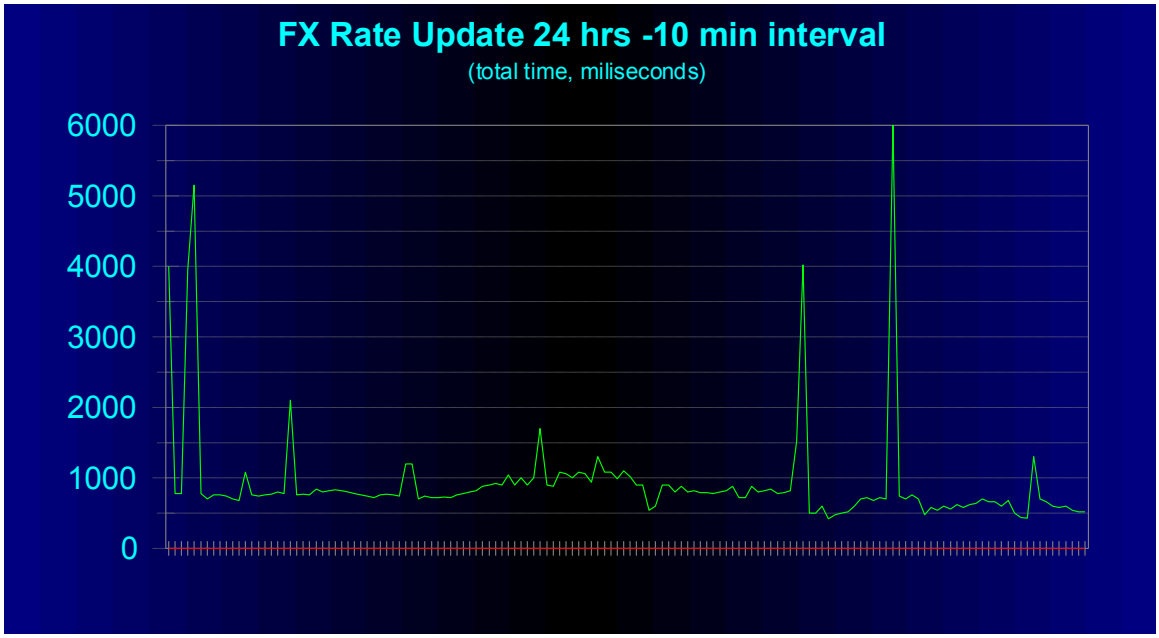
- automatic and manual building of OLAP cubes and other business intelligence constructs for analysis of systems and business metrics behavior with the purpose of troubleshooting reported problem tickets (short term data) or capacity planning (trend analyses).

INTRA-DAY AND INTRA-WEEK MEASUREMENTS (TROUBLESHOOTING)

Somewhat contrary to this section's title, intra-day and intra-week measurements are used for a variety of purposes, not least for measuring short-term overcapacity requirements to weather anticipated peaks and troughs in business activity. Since the underlying infrastructure is static in short term, and business performance metrics are expected, and even required to be relatively static (changing only within a narrow band), the required built-in overcapacity in hardware, software, networks, etc. can be estimated quickly and with a fair accuracy, if not during user acceptance tests and load stress tests, then in the first weeks of production implementation. Nevertheless, such overcapacity is the function of normally anticipated fluctuations in the volume of business, not of the growth in the volume of business. Business analysts and business units are usually attuned to performance metrics for each type of business application, and react to any deterioration in those metrics, even in cases when blips are short-lived and non-repetitive. Thus, the existing business and systems intelligence framework is used to build hourly, daily, or weekly OLAP cubes and other analytical constructs to overlay various systems and business metrics readings in search for clues that could explain various performance issues. A typical case is that of a business analyst calling the Central Intelligence Application analyst with a question: "has there been a problem with the application and/or overall IT systems at and around, say, 9:28 this morning?" If an OLAP cube that spans the time period in question has already been built, the analyst zooms on that hour and drills down across various metrics of interest, looking for strong correlation between the sharply deteriorating performance metric around 9:28 AM and other business metrics ("has the volume of business sharply increased just before that time; has there been unusual activity and increase demand for systems resources from other business applications?"), as well as various systems metrics (e.g. network traffic to and from servers involved in this particular business application; page scans and memory utilization; prolonged CPU spikes that could indicate rogue processes; JVM garbage collection frequency and duration, disk IO and its distribution between IO devices, etc.). Although these are usually *post mortem* (as opposed to real time) analyses, they allow a thorough analysis of large quantities of disparate data in search for explanations and patterns. Real time monitoring systems, particularly for business applications in conjunction with the underlying IT infrastructure, are rarely that comprehensive. They can be used for spotting a build-up to a problem, and thus trigger alarms that will help to solve the problem sooner, but usually do not allow a detailed search for underlying causes.

Figure 4 below shows a typical business performance metric for an FX application. FX traders constantly receive rate updates to their screens and the speed with which those rate updates reach the traders is vital to their business. As we can see, the measurements usually fluctuate between 0.5 sec and 1 sec, the latter being the maximum level of tolerance by the business. Severe performance degradation, when the metric reaches 4, 5, and even 6 seconds, potentially means lost business or lost opportunities, and needs to be investigated. The analyst will zoom on the selected 10-min intervals and overlay all available (within the OLAP cube) systems metrics, as well as other business metrics, looking for strong correlations and other clues that might help explain the observed performance blips.

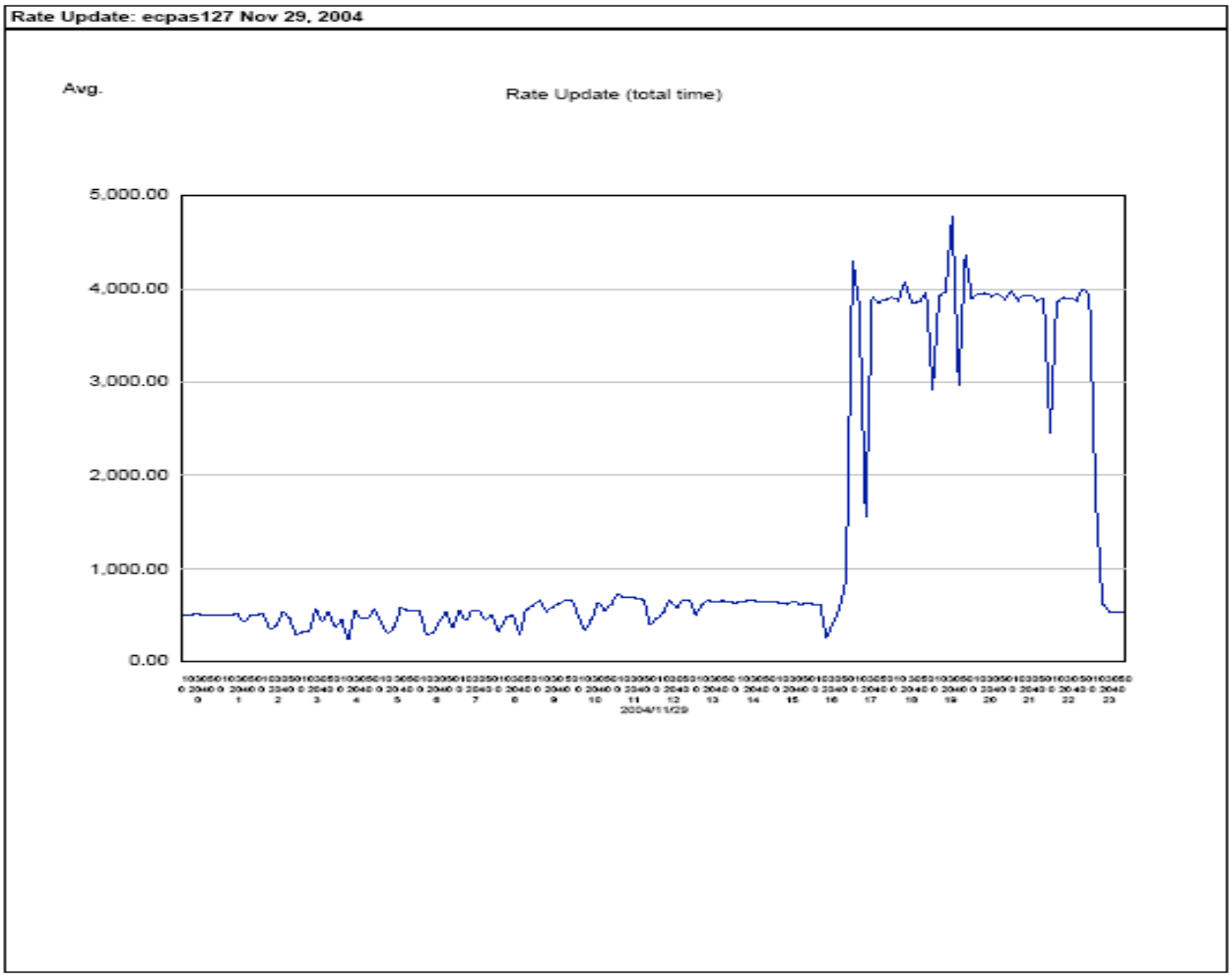
Figure 4: An example of a typical business performance metric – short lived performance blips



Sometimes correlating business metrics with system metrics does not reveal any issues, but splitting the business metric does. For example, in one of our cases a foreign exchange rate update time of n milliseconds was split into a few parts, each one time stamped separately. By doing so and graphing results over time, we could quickly narrow down our search to those sections of the application's code responsible for the most obvious bottlenecks.

Figure 5 shows the same business performance metric on a different day. Again, we can see that the target is approx. 0.5 sec, with fluctuations limited to a narrow horizontal band until approx 4 PM, when a major problem lasting for many hours is clearly visible.

Figure 5: An example of a major problem revealed by a business performance metric



CAPACITY PLANNING THROUGH LONG TERM METRICS MEASUREMENT ANALYSIS

Capacity planning, in the IT context, is about adjusting various components of the IT hardware and software infrastructure to meet demands of the changing volume of business – in a typical case, meeting the projected growth curve of the business. If we can find reliable correlations between the volume of business and other business and systems metrics, we can plan for the controlled system expansion that would facilitate maintaining performance metrics within the narrow horizontal band demanded by business units. In many cases the growth in business volume will fit into some components of the existing IT infrastructure, while requiring expansion of others.

It is clear from the preceding sections that e-business relies on a highly complex multi-layered infrastructure. Components of that infrastructure scale up differently in different layers, and simple correlations between a set of standard system metrics measured within a single server component and

metrics that reflect the e-business application rarely exist. On the contrary, projected changes in the volume of business require uneven, often discrete rather than continuous, expansion of IT components across all layers. As we have seen from the discussion of 'discontinuous relationships' between business and systems metrics, a typical e-business can grow within IT components with pre-allocated resources, where correlations between business and systems metrics – if they exist – are often hidden.

Therefore, the suggested approach is based on thorough analysis of business performance metrics for any particular e-business application. Performance requirements can be expressed as a distribution of measurements with averages, standard deviations, and other statistics describing the "required by business" and real distributions for each performance metric. Further analyses of multiple systems and business metrics, derived from OLAP cubes, and analyses of possible correlations between them, should serve the purpose of finding potential bottlenecks within different IT layers and leading to discrete changes in individual IT components.

CONCLUSIONS

Enterprise Business Units concentrate on business volume and business performance metrics, and not on systems metrics that reflect operation of the underlying IT infrastructure. Static performance targets – with performance measurements distribution within a narrow band – are a 'standard' requirement when business volume is growing or widely fluctuating. In real life business performance metrics usually correlate negatively with business volume measurements.

Metrics correlation is based on the implicit premise of continuity of relationships. In many e-business environments we face a class of 'discontinuous relationships' between business metrics and systems metrics, where lack of correlation does not mean lack of dependence. This is particularly true for e-business applications that follow an IT model of pre-allocated resources. Typical examples include many business logic applications running within JVMs, as well as backend database servers running within pre-allocated memory and/or CPU cycles.

Tracking business performance metrics *via* a Central Intelligence Application and correlating them with a number of systems metrics is useful in troubleshooting instances of performance degradation and finding clues about systems and/or application's bottlenecks.

Long term capacity planning is a complex process of adjusting levels of various IT resources across multiple layers, with the explicit target of maintaining performance measurements within a narrow target band.

REFERENCES

Ron Kaminski and Yiping Ding, "Business Metrics and Capacity Planning" BMC Software Inc. White Paper 2003

Yiping Ding and Chris Thornley, "On Correlating Performance Metrics"

Ron Kaminski, "Automating Workload and Process Pathology Detection"

Frederic J. Riggins and Saby Mitra, "A Framework for Developing E-Business Metrics Through Functionality Interaction"

Metrics for IT Service Management – ITSM Library, Van Haren Publishing 2006