

# Score High on Business Scorecards. Monitor What Matters.

*Leverage existing resource/application  
monitoring tools to enhance  
the end-user experience*



## Table of Contents

1	Business scorecards/dashboards—key IT performance indicators
2	Current resource/application monitoring solutions—an architectural shift
3	IT as the e-business factory—it's all about the business throughput
3	Performance monitoring from the end-user perspective—Keynote Service Level Delivery
4	Streamlined performance management workflow—Keynote Service Level Delivery with existing resource monitoring tools
5	The application is slow!
5	Separate ISP issues from application issues using Private Agents
6	Outside-the-firewall peering/DNS/third-party content issues
7	Peering problems with a remote ISP
7	Failure of third-party content, ad, or content delivery network servers
8	DNS failure leading to site unavailability
8	Behind-the-firewall application issues—page and metric-level drill-down for issue triage
9	Static content/Web server issues
10	Java 2, Enterprise Edition or Microsoft .NET application issues—root cause analysis
12	Keynote Service Level Delivery supports the business top line

As the popular adage goes, “The only thing constant is change.” Several enterprises today have focused their strategic initiatives on aligning constantly changing critical business processes with their underlying IT infrastructures. In an effort to make IT proactive and allow it to contribute to the business top line rather than react to customer issues, business executives are very interested in tracking key IT performance indicators as components of business scorecards.

To score high on such business scorecards, specifically with respect to Web application management, IT executives need to streamline application deployment and performance management workflow. Today, it takes Fortune 1000 enterprises approximately 25 hours to resolve Web application performance issues with an average loss of \$18,500 in revenue per hour of downtime, according to a recent study by The Newport Group. To enhance the top line rather than affect the bottom line, application support and operations managers need a repeatable, streamlined process to detect performance issues from the end-user perspective, triage the problem, determine the root cause, and validate end-user impact—all while leveraging existing investments in resource monitoring and systems management frameworks.

For several years, Keynote Systems has provided the industry’s most comprehensive, accurate, and credible real-time measurement of end-user performance and associated outside-the-firewall diagnostics. Now, with Keynote Service Level Delivery, Keynote combines those end-user measurements with a new, integrated set of behind-the-firewall diagnostic measurements that complement existing systems management tools for quick, targeted diagnoses of Web application problems.

This paper shows how Keynote Service Level Delivery complements popular resource monitoring, application monitoring, and systems management solutions to provide an end-to-end performance management platform. After a brief summary of the existing processes surrounding systems management tools and their potential shortcomings, we will explore how Keynote Service Level Delivery transactional monitoring combined with existing tools can address specific performance issues experienced in network operations centers (NOCs) today.

## Business scorecards/dashboards—key IT performance indicators

Today, business executives are interested in monitoring the success of IT in supporting highly dynamic critical business processes alongside other key business performance indicators. Specifically, with respect to Web applications that support the most critical customers and internal users, business and IT executives are very interested in tracking the business throughput of the IT infrastructure by measuring several factors on a regular basis:

- *Revenue-generating transactions processed*
- *Transaction volume breakdown by customer type*
- *Transactions compliant to availability and response time service-level agreements (SLAs)*

Business executives realize that perfect control is unattainable, so they also want to track business process exceptions handled by IT by measuring the following:

- *Transactions that exceed SLAs*
- *Business process/application downtime (absolute and percentage basis)*
- *Customer-reported performance and availability issues*
- *Mean time to respond to customer or application availability issues*
- *Mean time to resolve incidents/problems*

All of these measurements are made within the context of IT investments made to date, so that further process enhancements or controlled investments can be made in the future.

Ultimately, business executives would like to see 24x7, always-on IT performance for all systems on their scorecards. How can IT executives consistently deliver on the business expectations reflected in these scorecards? How can they exceed business expectations while applications grow more complex, deployment times shrink, and budgets stay flat or increase minimally?

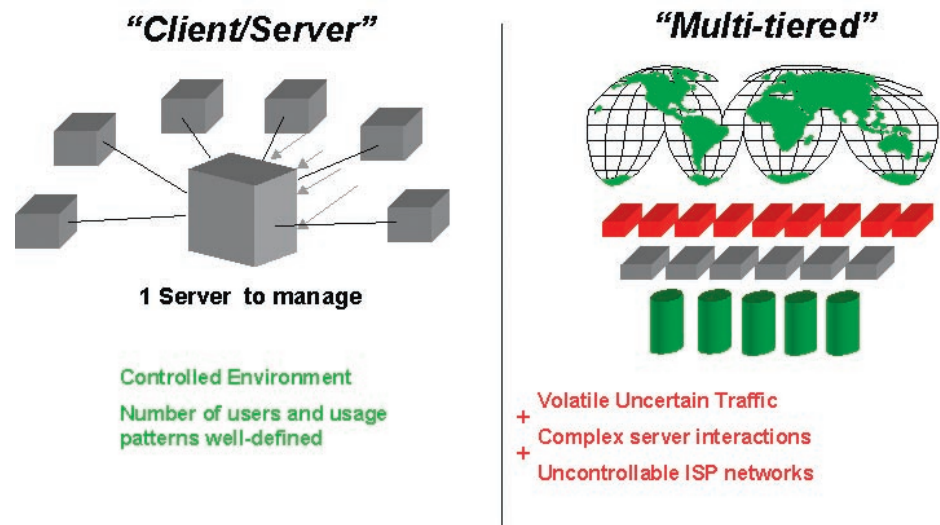
By monitoring what matters with Keynote Service Level Delivery—and driving performance management workflows based on the impact on the end-user experience. We will examine current resource monitoring tools deployed in IT environments and how they can be used in conjunction with Keynote Service Level Delivery to deliver what matters—an enhanced user experience for mission-critical apps.

### Current resource/application monitoring solutions—an architectural shift

To deliver on business expectations of application performance, operations teams have traditionally spent large sums of their budgets and extensive amounts of time to implement resource and application monitoring tools. Application monitoring refers to the software packages (Web servers such as Microsoft Internet Information Services, Sun-Netscape Alliance iPlanet, and Apache; application servers such as BEA WebLogic, Apache Tomcat, and IBM WebSphere; databases such as Oracle and SQL) that provide the integrated development environment (IDE) for mission-critical e-business applications.

With these tools, operations managers are able to gain a good understanding of their site health, receive quick notification of component level problems, and keep resource availability high. This approach works well for controlled mainframe and client/server environments in which the resource health provides a clear indication of the performance delivered to the end users (see Figure 1).

Figure 1  
Controlled client/server vs.  
uncontrolled Web application  
environment



However, in today’s complex multi-tiered Web environments, unpredictable traffic volume and user mix as well as frequent application level changes and upgrades make it impossible to directly correlate the underlying resource health to the end-user performance delivered.

Consequently, significant investments made by IT departments in best-of-breed resource and application monitoring tools have not yielded the desired results—results critical to business since the end-user satisfaction affecting the business top line is at stake.

## IT as the e-business factory—it's all about the business throughput

Monitoring/management solutions are a necessary requirement for quick notification of component-level health. For this discussion, we will focus on resource and application monitoring. These tools monitor infrastructure components (servers, networks, applications, and databases) for abnormal conditions relative to thresholds based on historical usage. Most major corporations in the world use standard enterprise management solutions that include Computer Associates Unicenter TNG, IBM Tivoli Cross-Site, BMC PATROL, Candle OMEGAMON, and HP OpenView software.

However, this approach provides numerous datasets, including CPU utilization and Oracle extent creation errors. The bottom-up approach adopted by these tools generates several false alarms that may not correlate to the actual end-user performance delivered. This anomaly can be compared to an automobile manufacturing plant that manufactures X cars per week with a cycle time of Y. Knowing that each assembly machine in working condition is operating at 90 percent is good information for maintaining the health of the machines. But does the plant manager really care if a machine is at 90 percent capacity unless it affects the cycle time and throughput of the automobiles manufactured daily? On the other hand, even if certain machines are at 40 percent capacity, can the plant manager afford to ignore the resulting throughput and cycle time degradation?

Similarly, just because the CPU on an application server is running at 85 percent capacity, which may exceed a threshold set at 80 percent, it may not be a business-critical issue if the performance for all users meets the SLA. On the other hand, an increase in memory consumption on a database instance from 20 percent to 40 percent may not generate any resource level alerts, but if it increases the response time for an investing transaction from four seconds to nine seconds—thereby exceeding the 8-second SLA—then the risks to your business could be dramatic.

In a nutshell, resource and application monitoring can help operations personnel in several ways:

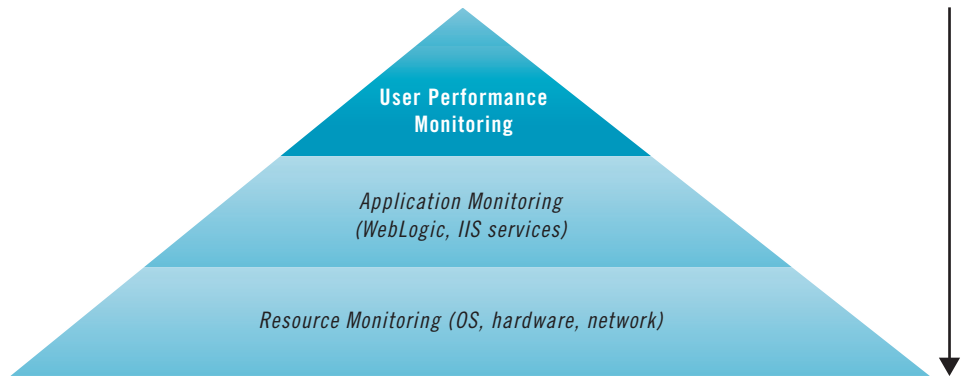
- *Identify a resource problem*
- *Help determine the root cause of the problem through the use of cohesive correlation rules*
- *Help identify who to call to deal with the problem*

Usually, operations personnel are only able to determine that they have a problem and that more investigation is required to determine the exact cause and appropriate resolution. The actual impact on the user community is much more abstract and may never be known, much less improved. Moreover, the process of correlating large datasets from various inconsistent sources without first understanding the end-user impact is highly cumbersome and could exponentially increase the operational expenditure without yielding any performance enhancements or top-line business benefits.

## Performance monitoring from the end-user perspective— Keynote Service Level Delivery

With Keynote Service Level Delivery, enterprises can now begin with end-user performance thresholds relative to business SLAs, and then use resulting violations to prioritize and drive the exceptions generated through resource and application monitoring. By adopting this approach, operations teams can now focus on events that affect the end-user experience, and by resolving these issues, they can directly impact customer satisfaction ratios vital to the business top line (see Figure 2). At the same time, the seamless linkage between Service Level Delivery and leading resource monitoring and management frameworks including NetIQ AppManager, OpenView, and Tivoli Enterprise Console streamlines the performance management process—optimizing operational expenditures while maximizing return on existing investments.

**Figure 2**  
Top-down end-user performance  
monitoring process



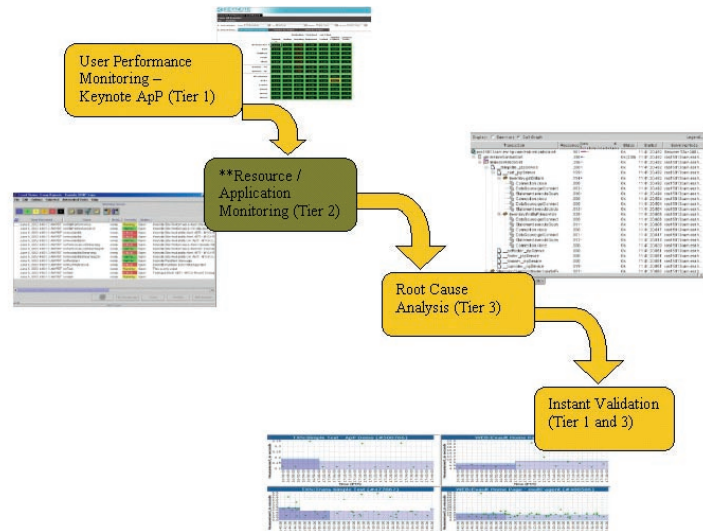
### Streamlined performance management workflow—Keynote Service Level Delivery with existing resource monitoring tools

Keynote Service Level Delivery combined with existing resource and application monitoring tools provides a streamlined performance management workflow that directly maps and enhances the existing operations process (see Figure 3).

Typical IT organizations have three or more tiers of operations. Variations occur based on company size and philosophy, but common tier structures feature these responsibilities:

- **Tier 1** is the network operations center, where personnel are trained to watch a central console that receives events from numerous monitors spread throughout the environment (including performance monitors outside the data center). Their main responsibility is to detect problems and notify the appropriate Tier 2 personnel who have the skill sets necessary to correct the problem.
- **Tier 2** personnel are the system/database/application administrators focused on incident management. They have vertical skill sets in their respective areas of expertise, such as UNIX, Microsoft Windows NT, Oracle, and SAP. Tier 2 personnel can perform triage, apply corrective actions and workarounds, and provide detailed problem reports to Tier 3 operations personnel.
- **Tier 3** operations personnel are the application support developers who drill down to the root causes of recurring issues and create permanent fixes for production applications.

**Figure 3**  
Typical application performance  
management workflow  
*\*\*Note: All components except for the green shaded component are part of the Service Level Delivery platform, which integrates with the existing customer resource and application monitoring solution*



## The application is slow!

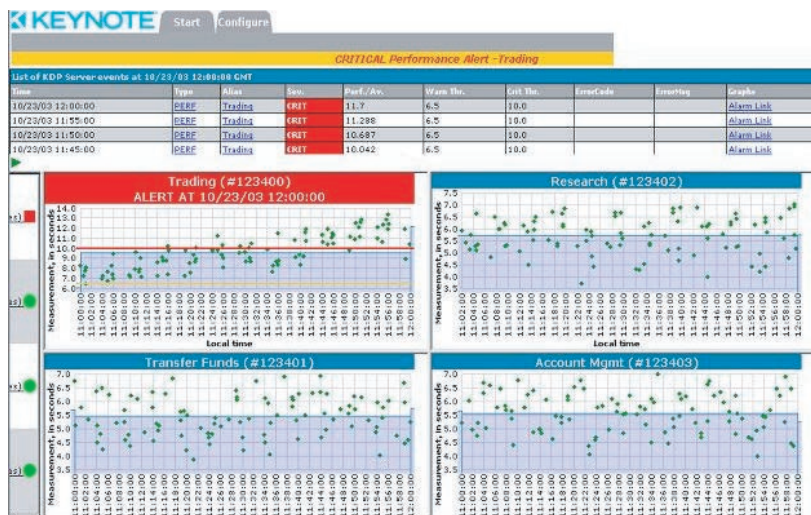
In line with the hierarchy above, Tier 1 staff detects a performance issue as it relates to application end users in the Service Level Delivery Console (also called MyKeynote™ Inside). Customers can count on the credibility and coverage of the Keynote end-user measurement system (see Figure 4).

Keynote's statistical distribution of over 1,700 measurement agents across the world's primary ISP backbones and the world's major metropolitan areas ensures that Keynote subscribers quickly become aware of any problems encountered by their end users. With Keynote, you are quickly aware of any problems reaching the major distribution nodes of the major internet service providers, and you are not distracted by false reports of problems from poorly controlled measurement agents with congested local access links.

Furthermore, events and exceptions shown in the Keynote Service Level Delivery Console are also bridged to your systems management console.

Whether you use an OpenView or Micromuse Netcool console as your event management system or have a separate screen in your NOC for the Service Level Delivery Console, all issues affecting end users can be visualized to help trigger resolution processes.

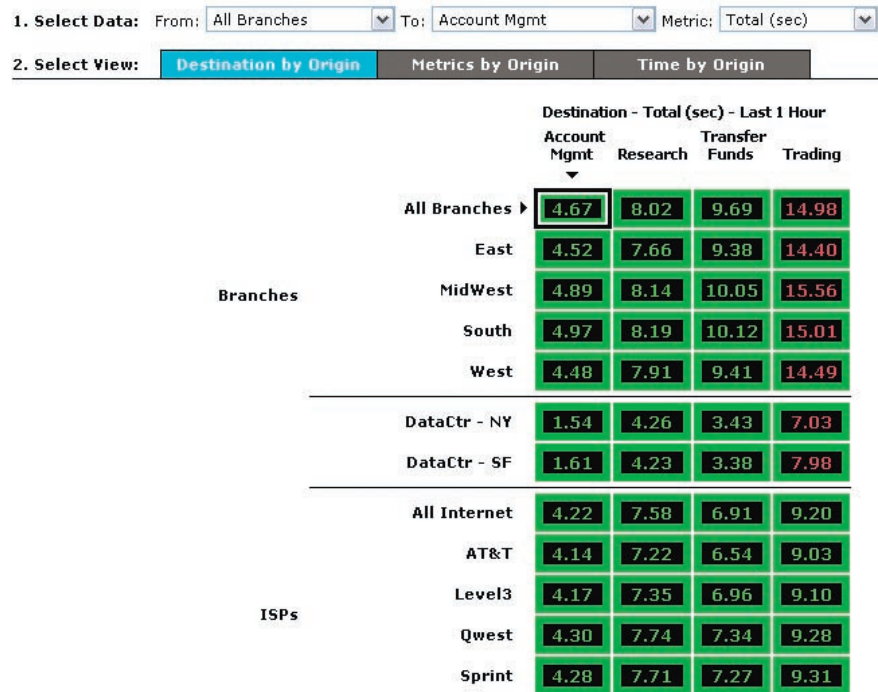
Figure 4  
Real-time Service Level Delivery Console (MyKeynote Inside)



## Separate ISP issues from application issues using Private Agents

Keynote Service Level Delivery is based on Application Perspective™ or Transaction Perspective agents that measure from global locations as well as from private diagnostic agents on the customer's premises to provide an apples-to-apples comparison. Keynote's Application Perspective private agent is a multi-threaded, high-frequency agent capable of generating up to 3,000 pages per hour. By comparing the external public and internal private agent transactional monitoring data, Tier 1 personnel can quickly determine if the issue at hand is related to an external ISP or whether it is a data center application or content issue.

Figure 5  
Keynote Performance Scoreboard  
Console



For instance, the trading transaction that generated an alert in Figure 4 is now examined in the Keynote Service Level Delivery Console (see Figure 5). The private agent measurements from data center NY and data center SF (load-balanced system), which host the application, are compared to the external end-user locations across the continental United States.

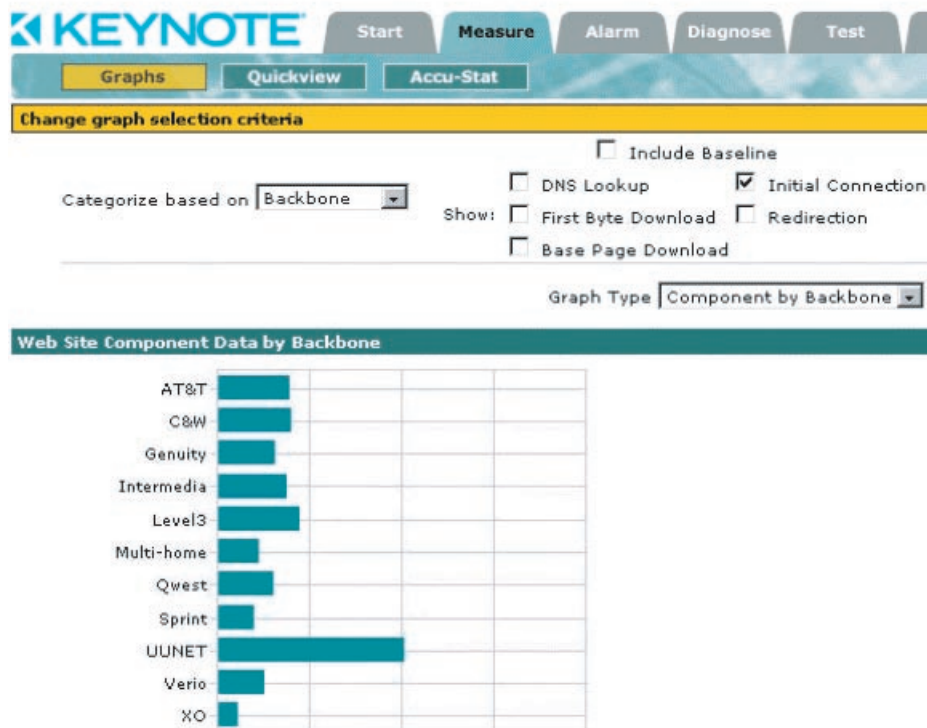
Based on dynamically computed baselines for Internet latency, Service Level Delivery Console can automatically flag the internal measurement to indicate an application issue. In Figure 5, for instance, the average difference between external measurements from branch offices (14 seconds) and internal data center measurements (7.5 seconds) for the five-page trading transaction is below the baseline. In other words, 6.5 seconds (the difference between external and internal measurements) is the expected Internet latency—the increased length of the internal measurement of 7.5 seconds indicates an application performance issue behind the firewall.

Furthermore, the diagnostic private agent can be used to trigger troubleshooting transactions upon receipt of specific events, or can even monitor representative content that indicates the health of individual servers. In a nutshell, the diagnostic private agent provides a critical building block for the implementation of performance management best practices using highly credible Keynote monitoring technology.

### Outside-the-firewall peering/DNS/third-party content issues

If the issue at hand is outside the firewall, customers can rely on Keynote’s diagnostics module, which has long been used by the industry. Time is not wasted on finger-pointing when Keynote measurements are involved; you will never hear an ISP say, “You’re measuring your own congestion,” or “We’re rejecting your trouble report because we believe that your measurement is faulty.” Keynote uses dedicated measurement agents connected to key Internet backbone nodes using uncongested links that are tested every second for proper performance and monitored by a 24x7 team. Keynote measurements have a six-year record of complete credibility with ISPs and other service providers. A few examples should illustrate these classical uses of Keynote measurements for diagnoses of outside-the-firewall problems.

Figure 6  
*TCP connect time (round-trip network latency) by end-user backbone*



### Peering problems with a remote ISP

In most cases, your end users are not using the same ISP that you are. They may be using Sprint while you use Cable and Wireless; they may be using AT&T while you use Level 3. In all cases, however, it is critical that you quickly learn about any problems they may have in reaching your site and that you get the data you will need to have the concerned ISPs to fix the problem fast.

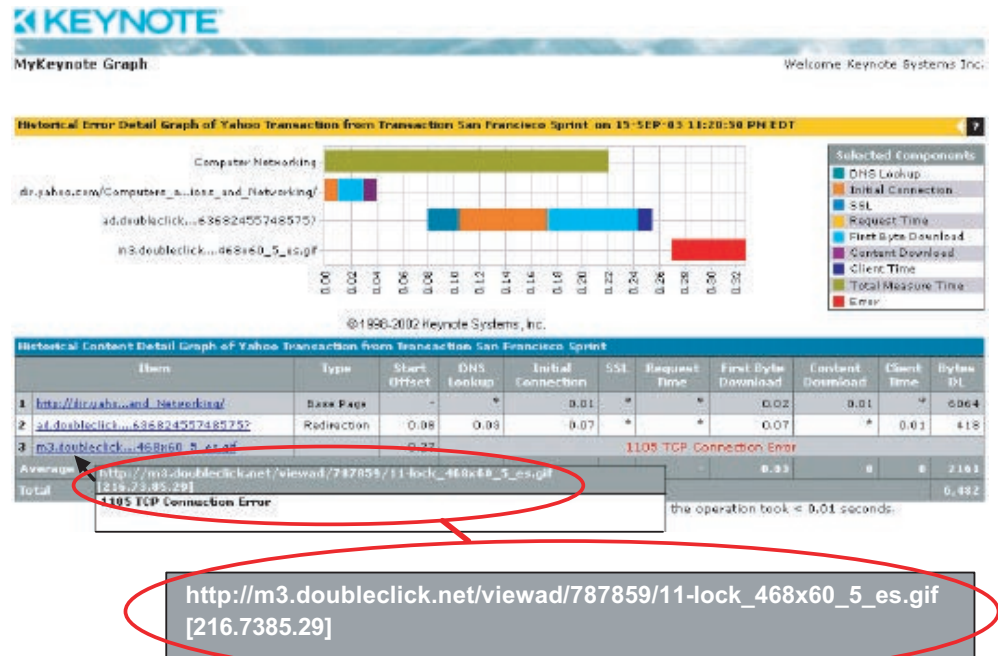
A couple clicks select the Keynote chart that shows TCP connect time (the best indicator of pure network latency) classified by measurement agent for the problem period (see Figure 6). The peering problem to UUnet becomes apparent and is contrasted with the performance to other ISPs. There is no need to spend time running the traceroute utility—the information contained in the Keynote chart is more conclusive. One click prepares a Keynote chart to be e-mailed to your ISP, which will then manage the situation with its peering partners. You do not need to spend time trying to convince your ISP that the problem exists, because all major ISPs are familiar with Keynote, understand Keynote charts and data, and understand Keynote’s credibility.

### Failure of third-party content, ad, or content delivery network servers

Many Web pages contain content that is served from outside the firewall. For example, many financial sites use third-party providers of stock charts; many commercial sites use third-party advertising providers. Sites that need quick delivery of page content use content distribution networks (CDNs). These external server systems usually use geographic distribution to enhance their performance, but the result is that the end-user location affects the behavior of the Web site. End-user measurement is needed to detect regional failures and ensure that service-level objectives are being met.

It is easy to select the Keynote chart (see figure 7) that shows the precise details of the failed page element, which provides all the evidence you need (IP address, precise URL, and exact time) to get a quick response from the external third-party content supplier at hand.

Figure 7  
Historical content error detail—  
showing all page elements for  
failing page



## DNS failure leading to site unavailability

Failure of the Domain Name Server (DNS) usually makes it impossible for a remote site to convert your hostname into an IP address. Without end-user measurement, DNS failure is not detectable. Fortunately, it takes just two clicks in the Keynote user interface to produce the two charts shown in Figure 8. Those charts can be used to locate geographic regions and backbones that are seeing the DNS failures as distinct from other errors that may be occurring at the same time. Clearly, in this example, the DNS problems are local to a single ISP in Dallas (AT&T) while TCP connection timeouts are widespread and are probably caused by difficulties in the server room.

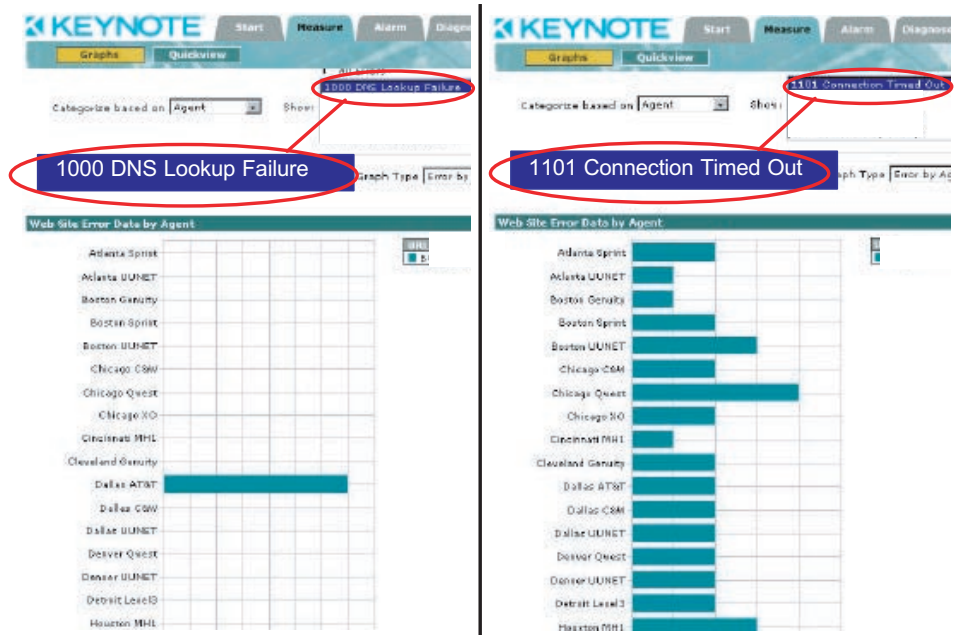
It is charts such as these, backed by Keynote's reputation for accurate, credible measurement, that result in instant action from ISPs. You and your ISP or other external service provider can be looking at the same chart in just a few seconds. For six years, this accuracy and reliability is why Keynote has been the leader in diagnoses of outside-the-firewall problems.

## Behind-the-firewall application issues—page and metric-level drill-down for issue triage

Keynote is now taking a new step toward fast resolution of behind-the-firewall problems by leveraging existing systems management tools in the NOC. As mentioned earlier, Keynote Application Perspective integrates application-level measurements from private agents inside the firewall with Keynote's existing public agents to form an end-to-end measurement solution.

The excellent accuracy, representation, and comparative benchmarking abilities of the Keynote public measurement agent network are combined with the outstanding diagnostic abilities of the application-level measurement system to provide an end-to-end application performance management platform. A few examples will illustrate the end-to-end system in action.

**Figure 8**  
*DNS failure and connection timeout failure for individual Keynote agents*



### Static content/Web server issues

Based on a comparison of the public and private agent measurements, Tier 1 staff may determine that a problem at hand is an application or content issue behind the firewall. The Tier 1 administrator can drill down into any transactional measurement (collected up to once every minute) to get a page- and metric-level breakdown of the transaction, as shown in Figure 9.

As seen in Figure 9, Content Download on Page 1 is high, but the TCP Connect (Initial Connect), First Byte, and Base Page Download are relatively low. In this case, the Tier 1 administrator can open a trouble ticket for the Tier 2 Web administrator to examine the issue further.

**Figure 9**  
*Transaction page and metric-level breakdown*

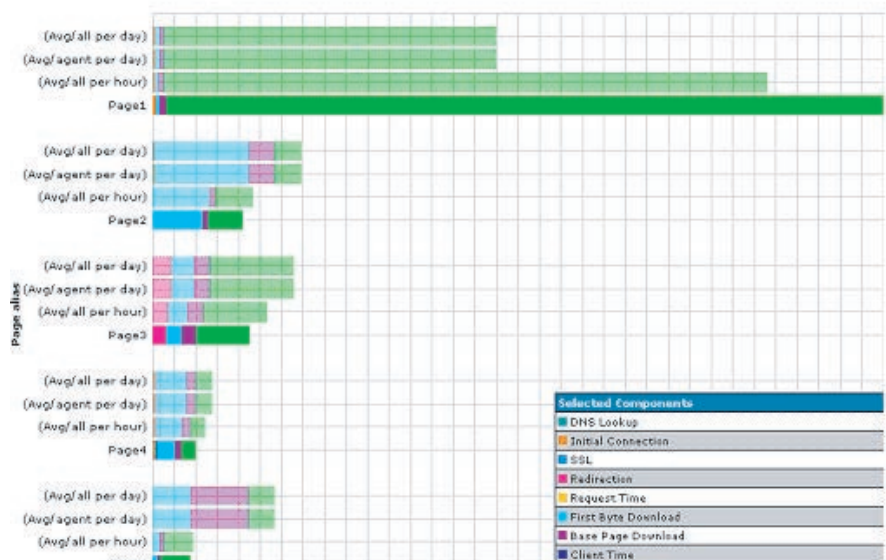
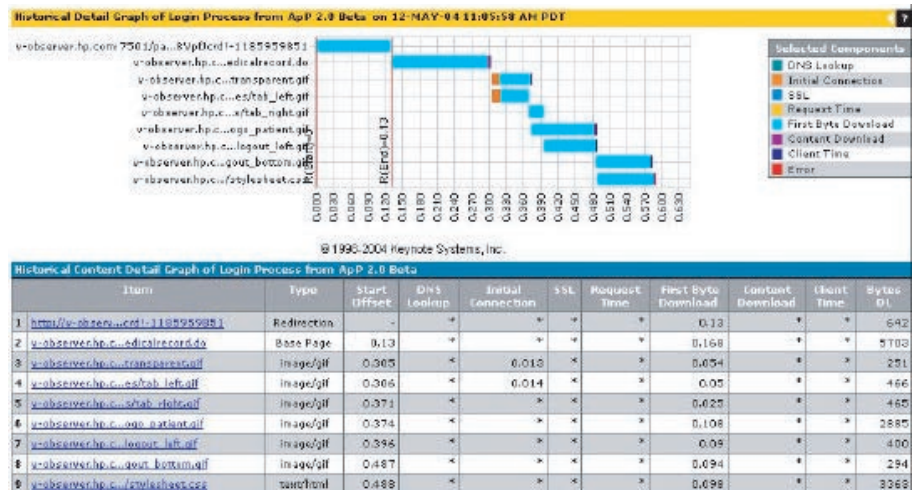


Figure 10  
Object-level breakdown for static content issues



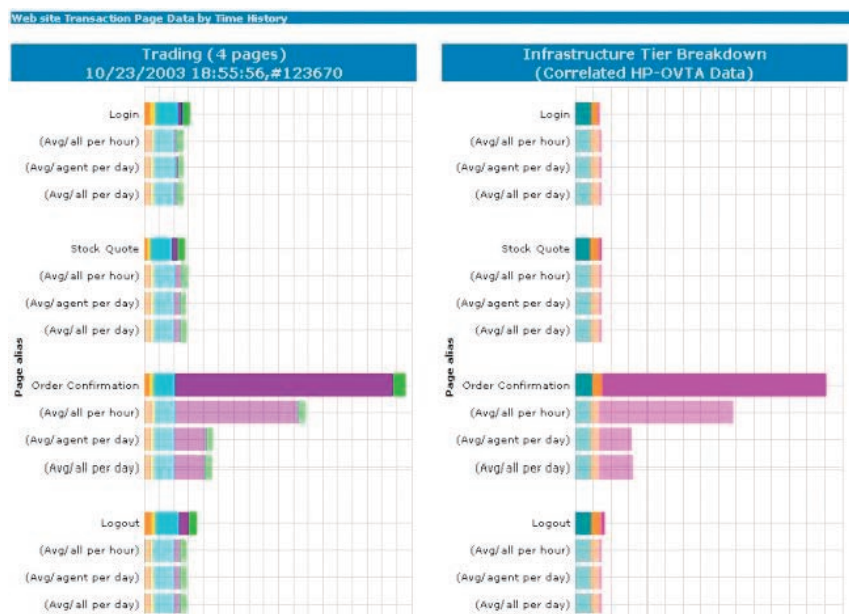
The Tier 2 administrator can look further into the Page 1 content (Figure 10) that indicates the exact object on the Login page that consumes most of the page download time. In this case, the first object results in two redirects and Object 6 (patient.gif) consumes a major portion of the page response time. In this case, the content owner can optimize the image to produce a faster response time.

On the other hand, if First Byte or Base Page Download time in Figure 9 is high, the Tier 2 administrator can look up the resource map for the transaction under investigation and tune the server at fault to manage the incident.

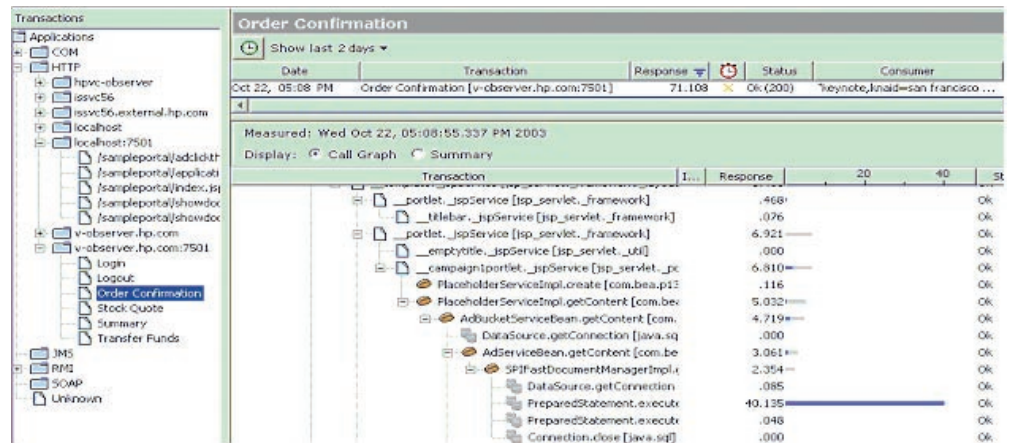
## Java 2, Enterprise Edition or Microsoft .NET application issues—root cause analysis

As shown in Figure 11, if the Tier 1 administrator detects that the order confirmation page for the trading transaction discovered earlier has a high Base Page Download time and most of that time is being consumed in the database tier, the administrator can allocate the issue to a Tier 2 database administrator or Tier 3 application developer.

Figure 11  
Application logic or database query issues



**Figure 12** ❖  
*Root cause analysis at the individual query level*



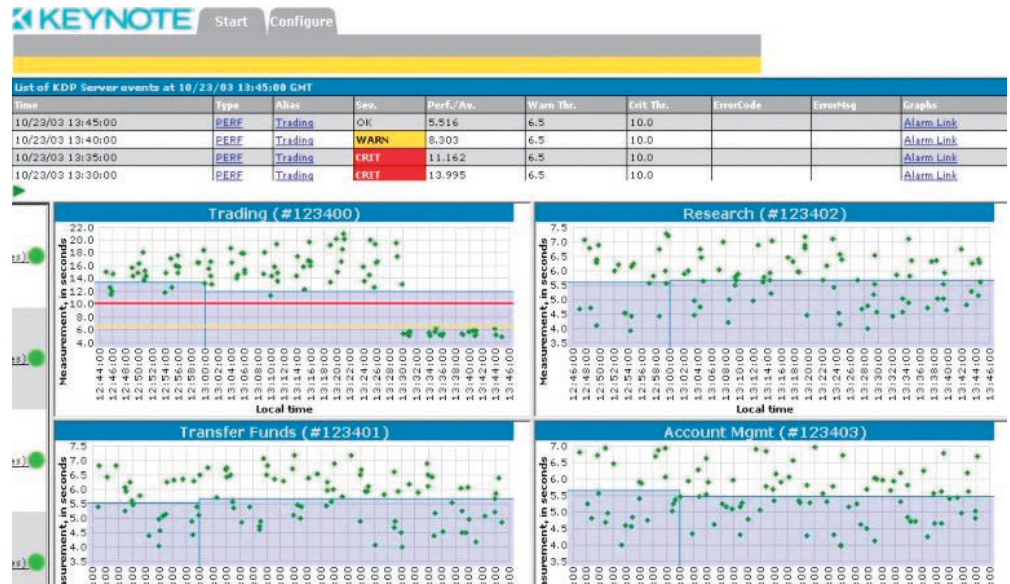
Tier 3 application developers can right-click on the Order Confirmation database section (pink) to conduct root cause analysis, which takes them to the screen represented in Figure 12. The Order Confirmation involves a portlet, which makes a Java database connectivity call to the back-end database resulting in the decrease in performance. Armed with this intelligence, Tier 3 application support personnel can conduct code review and optimization and then deploy the fix in the quality assurance, staging, or production environments.

Following the problem resolution, Tier 1 and 3 administrators can refer again to the Keynote Service Level Delivery Console to conduct instant validation prior to closing the trouble ticket (see figure 13).

Best of all, Service Level Delivery does not require any changes to user applications. It can be used with both custom and packaged software—no modification of the existing Web server or application software is required. Instead, it uses well-proven, standard techniques for inserting software “shims” or “hooks” between the operating system and the applications. Whenever an application makes a call to the operating system for a service, such as performing a database access or sending an inter-process message, Service Level Delivery can be informed of the call and the timings. Overhead for this process is quite small, on the order of only a few percent, and can be made even smaller by appropriate configurations and statistical techniques.

Ultimate diagnostic capabilities have now, for the first time, been completely integrated with Keynote’s industry-leading end-user experience monitoring. Keynote’s new combination of technologies means that your operations team can now use the same tool to perform an instant diagnosis of any problem, outside or inside the firewall, and can collaborate with your application support team to resolve problems at unprecedented speed.

Figure 13  
Instant validation of trading  
performance fix



### Keynote Service Level Delivery supports the business top line

To align IT infrastructures with critical business processes, business executives are very interested in tracking key IT performance metrics on business scorecards. Specifically, with regard to Web application management, IT is measured on the number of revenue-generating transactions processed with SLA compliance (business throughput) and the agility in handling SLA exceptions. To deliver upon business expectations on such scorecards, IT and operations teams have traditionally adopted systems management tools that worked well in closed-loop client/server environments. However, unpredictable Web traffic patterns and the ever increasing complexity of applications present a new set of performance management challenges.

Using traditional systems management tools in isolation, enterprises face the cumbersome task of prioritizing amidst several low-level resource alerts and, more importantly, run the risk of missing business SLA violations that do not manifest themselves as resource issues.

Keynote Service Level Delivery adopts a top-down approach by monitoring what matters—exceptions to business throughput as measured by SLA violations from the end-user perspective. Armed with this intelligence, Service Level Delivery then leverages Keynote’s public measurements and behind-the-firewall diagnostic private agent with existing resource and application monitoring tools to provide a comprehensive performance management platform—all delivered as hassle-free services.

By using this approach, operations teams and application support staff can collaborate and streamline incident and problem management workflow, leverage existing investments to their fullest potential, and actively contribute to the business top line by supporting dynamic business processes.



©Copyright 2005 Keynote Systems, Inc. All rights reserved.

Keynote, MyKeynote, Perspective, and the Internet Performance Authority are trademarks or registered trademarks of Keynote Systems, Inc. in the United States, other countries, or both.