

Quick Problem Detection and Triage with Keynote Red Alert

White Paper



2855 Campus Drive
San Mateo, CA 94403

www.keynote.com
1-800-KEYNOTE
(1-800-539-6683)

Table of Contents

What is Triage?	3
Triage Using Keynote Red Alert	4
Red Alert for Web Server System Triage	5
Red Alert for Internet-Server Triage	7
Summary	8

All rights reserved. No portion of this document may be reproduced without prior written consent. Keynote Systems, Inc. shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance, or use of this material.

With the current state of technology, it's impossible to guarantee that your Internet and Web presence will never go down -- but it is possible to decrease that downtime by quick detection and repair of problems. And quick repair starts with *triage*.

What is Triage?

Triage is the process of performing just enough diagnosis to select the appropriate support team and convince them that the problem is truly theirs to solve. An effective triage system eliminates finger-pointing among the support groups and the resulting loss of valuable time. That time savings, repeated for each failure, can easily amount to many hours per month -- time that can be worth tens of thousands of dollars or more, in lost sales, performance penalties, and customer dissatisfaction. Triage works well if it can be performed quickly and reliably, so that its results are available almost immediately and so that the support teams trust those results. If it takes too long, or involves complex, difficult-to-use tools, or delivers questionable results, it will not be usable in a production environment. Support teams won't use it, or they won't trust the results, and the classic finger-pointing exercises will reappear.

It's important to avoid thinking that triage is the same as detailed diagnosis. The goal of triage is to get the right team working on the problem as quickly as possible, not to diagnose the problem to its root cause. Diagnosis is properly performed by the appropriate team, as they'll have the specialized tools and access to the data that is necessary for accurate analysis. Attempting a diagnosis with general-purpose tools, although possible in some cases, almost always increases the total time spent repairing the problem. It's far better to get the right team involved quickly, instead of wasting large amounts of time trying to do a complete diagnosis without the best tools.

For example, the general-purpose traceroute tool can be used in attempts to diagnose both Internet peering problems and server-room problems, but using traceroute alone, without specialized tools and data, can take a long time and lead to inconclusive results. For Internet peering, the Internet Service Providers have better tools than traceroute to diagnose and fix the root cause of their peering problems, and they'd never let customers access their peering data anyway, even if customers had those tools available. For server-room diagnosis, straightforward use of standard SNMP-based tools such as HP Openview or manager-of-managers tools such as CA/Unicenter are faster and more detailed than traceroute. The key is to get the correct group -- Internet Service Provider or server room support staff -- involved in the problem quickly, then let them do their diagnosis with the appropriate specialized tools and with their access to the appropriate data. And that's the role of triage.

Triage Using Keynote Red Alert

Each server-room component of your Web and Internet presence should be monitored by the Red Alert system to ensure that difficulties are detected quickly and that triage is performed quickly and accurately.

- Red Alert has built-in, automated triage tests for the server room's Internet connectivity and the local server-room network, for the Web servers, and for additional Internet-related servers such as the DNS, FTP, and email servers.
- Red Alert has a quick cycle time for rapid detection of problems.
- Red Alert has alarm paging (including email, pager, and mobile telephone messaging with TAP/IXO) and sophisticated alarm escalation facilities to ensure rapid, effective response.

We'll look at the use of Red Alert for both Web servers and for your other Internet-related servers, with Web servers first.

Red Alert for Web Server System Triage

Red Alert should monitor the primary IP address of your server farm, such as your load-distribution device, and also the major individual Web servers within the server farm, such as your primary image servers and search servers.

- Red Alert has the flexibility needed to access your servers; it can handle secure pages, redirection, CGI queries, POSTs, and authentication.
- Red Alert has the power to examine the retrieved data to test for delivery of the proper response, with the proper content and the proper length.
- Red Alert decides where to send each alarm depending on the error type and the time of day; it automatically performs triage on the situation and alerts the appropriate team.
- Red Alert will be believed and acted upon by your team to repair server situations quickly; its sophisticated triple-test system means that Red Alert responses are reliable, with very few false or unnecessary alarms.
- Red Alert is an inexpensive, completely outsourced service. It is therefore very easy to configure, and a small investment in Red Alert is often repaid by the first fast triage.

The recommended setup for a server system that includes an entry (boundary, or gateway) router, load distribution, Web servers, image servers, and search servers is outlined below.

- **Use Red Alert to monitor the entry point of your server farm to detect failure of the load distribution device, failure of the local server-room network, or failure of the access link or entry router to your server farm.** This can be done by targeting the Web page of the load distribution device itself (such Web pages are used for device configuration), or by inhibiting Red Alert's ability to follow redirects issued by the load distributor. Red Alert's "gateway ping" facility should be used to monitor the entry router's IP address, and Red Alert's "sibling ping" facility should be used to monitor another device on the same local network segment as the load distribution device. That will allow Red Alert to perform automated triage on the router and on the network. Red Alert's sophisticated triage testing will then be able to detect failure of the server farm's Internet access, and *Red Alert will automatically perform triage on the situation to differentiate among catastrophic failure of Internet connectivity, failure of the access router, failure of the local server-room network, and failure of the load distribution device.*¹ The Red Alert triple-test will help avoid sending an alarm when there is a partial failure due to peering problems somewhere in the Internet; it will alarm only when there has been a catastrophic failure of the server farm's Internet connectivity. (Geographically-based partial failures and other, more-subtle failures in the Internet or the server system can be detected by Keynote's Perspective alarms. For example, Web Site Perspective alarms can be set to detect peering problems in limited geographic areas, and Transaction Perspective alarms can detect problems in an individual page deep within a complex transaction.)

- **Use Red Alert to monitor through the load distribution device and also to monitor each Web server directly; this will detect load distribution device or Web server failure.** This is done by targeting a Web server URL (for example, of your home page) and allowing redirection or other connection through the load distribution device, and also by targeting the Web servers directly, if possible, using their individual IP addresses or URLs. Because Red Alert can handle secure pages, authentication, and redirection, it can penetrate the load distribution device and most security mechanisms to retrieve a page from the Web servers. Red Alert should also be configured to evaluate the retrieved page, checking for the correct content and length. This will detect failure of the load distribution device, the authentication process (if any), and the Web servers. If each Web server is measured directly, Red Alert will also instantly detect and report the failure of a particular server.
- **Use Red Alert to monitor each major Web page that's critical to customer satisfaction and that may fail independently of the page measured above.** Some server systems, especially those involving dynamic page generation, may be unable to generate certain pages even if the primary page, measured in the step above, can be delivered by the server. In such situations, you will probably want to take the precaution of measuring a representative set of those pages directly with Red Alert. It's very inexpensive insurance against embarrassing failure. Red Alert's ability to request a URL with an embedded Query String will allow you to handle many types of dynamically-generated pages.
- **Use Red Alert to monitor each major image server.** Failure of an image server may create the appearance of complete Web site failure, because most browsers receive only a few images in parallel. If, as a result of bad luck, all of those images are on the same failed or unusually-slow server, then the user may see a complete browser freeze while all of the browser's parallel access attempts time-out, waiting for the server. Avoid this by using Red Alert to retrieve an image from each image server, testing to be sure it's of the correct length.
- **Use Red Alert to monitor the search engine.** Because Red Alert can generate a Query String or a POST, it can be used to test the response to a search request. That will find search engine failures promptly.

This is the Red Alert setup for a complex Web server farm situation; simpler situations are simpler to set up. For example, if you have a single Web server instead of a load balancer and multiple servers, you'd combine the first two bullet points into a single Red Alert monitoring of a Web server URL.

Note that in all cases, geographic issues, such as peering difficulties, problems with Content Distribution Networks, or difficulties with Ad Servers are not handled by Red Alert; that's a job for Keynote Perspective alarm services.

Red Alert for Internet-Server Triage

There's more to e-business than just Web sites; e-business also relies on email, chat, file transfer (FTP), and Authoritative Domain Name System (DNS) services. Red Alert can play a key role in the quick problem detection and triage of these services because of its ability to open a TCP connection to *any* server system, not just Web servers. Red Alert's ability to test file length is also of great benefit in evaluating the correct performance of any system that generates a log file as a part of its operation; sequential Red Alert tests showing that the log file has or hasn't changed in size can indicate a failed system.

A TCP connection, as used by Red Alert, is a far more reliable indicator of success than the classic ping test. A ping reaches only the outermost boundary of a server's Internet protocol software, which may be handled on the I/O card. It does *not* test the availability of an application. Some server systems may respond correctly to a ping if there's power to their I/O card, even if their CPU is powered down. In contrast, a TCP connect succeeds only if an application has notified the Internet protocol software that it's running and willing to accept incoming connections. TCP connect is therefore a much more reliable indicator of application availability than is ping. That's the reason the Red Alert uses TCP Connect, not ping, as a means of testing non-Web servers for availability.

- **Red Alert should monitor the appropriate TCP application ports of your Internet servers (such as your email, chat, FTP, and DNS servers) to ensure that their applications are available and willing to accept incoming connections.** Some systems may Red Alert to monitor a number of different ports to test for the availability of different services. For example, Red Alert might monitor both TCP Port 23 for Telnet and TCP Port 21 for FTP. Your Authoritative DNS server uses both UDP Port 53 and TCP Port 53; it's the latter that should be monitored by Red Alert to see if outside users are able to reach it.
- **Red Alert should test the length of log files.** Red Alert can be set to generate an alarm if the size of a file either changes or doesn't change between sequential file retrievals. The failure of a log file to grow indicates system failure in many Internet systems; in other cases, the increase in size of an error file may indicate a failure.

Summary

Keynote Red Alert is a highly cost-effective way to provide quick problem detection and triage within a server farm and with the server farm's Internet connections. It should be used as an adjunct to Keynote Perspective services to provide very fast, accurate triage of both Web servers and other Internet servers (email, chat, FTP, DNS) within the Web farm. *The small cost of Red Alert will probably be more than paid for with the first speeded-up triage!*

¹ This gives an excellent example of how Red Alert operates. When a failure occurs, Red Alert will automatically perform the testing needed to choose among the following triage messages, which it will send to the appropriate pager/email, depending on the message and the time of day:

- Catastrophic Internet failure or gateway router failure; gateway router cannot be reached from any of three separate Tier 1 (major) ISPs.
- There's a local network problem: the gateway router is accessible, but neither the load distribution device nor another device on the same local network will respond to anything, even a ping.
- Gateway router is accessible, and load distribution device will respond to a ping, but the load distribution device won't respond to a TCP connection attempt.
- Gateway router is accessible, and load distribution device will respond to a TCP connection attempt, but it returns an HTTP error when a Web page download is attempted.
- Gateway router is accessible, and load distribution device will accept an HTTP request for a Web page, but no data is received after 30 seconds or the TCP connection is interrupted.
- Gateway router is accessible, and load distribution device will accept an HTTP request for a Web page, but the proper keyword cannot be found in the delivered data or the data delivery is interrupted for more than 30 seconds after it starts.
- The size of the delivered Web page is incorrect.

If the problem persists past a user-set interval, Red Alert will automatically escalate the situation and page additional people according to your requirements.