Finding and Collecting Availability Measurement Data for the IBM HTTP Server and WebSphere Application Server for OS/390

Mike Bonett
IBM Advanced Technical Support
Gaithersburg. MD
bonett@us.ibm.com
March 2001

Many I/T environments are finding that OS/390 and zOS operation system platforms are prime candidates for implementing e-business applications. These applications rely on the technical foundation provided by the IBM HTTP Server for OS/390 and the WebSphere Application Server for OS/390. These products become part of the end-to-end path for an application, to allow users to submit requests to application logic, and application logic to retrieve the data for presentation back to the user.

The task of monitoring and measuring the availability of the IBM HTTP Server and WebSphere in the OS/390 and zOS environment must be done to properly manage the applications they support. This paper identifies what data is available and how to collect it from two sources of measurement data:

1. ***Event Sources***, locations where products can directly place status information. This includes:
   a. Product logs
   b. SMF
   c. System messages
   d. SNA alerts
   e. TCP/IP traps
2. ***Monitoring Techniques***, which can be used (either directly or as part of a product function) to monitor a product component to determine its status. These include:
   a. Heartbeats
   b. "Ping"
   c. Remote commands
   d. User simulation
   e. Custom monitoring agents

Detailed definitions of these sources can be found in the White Paper ***Finding and Collecting Availability Measurement Data***, which is available from the author.

Each event source or monitoring technique has advantages and disadvantages for both products. The policies within a particular I/T environment will also influence which sources can be accessed or which monitoring techniques can be deployed. Information from the appropriate event sources and monitoring

techniques provides, in various levels of detail, information on IBM HTTP Server and WebSphere Application Server availability.

This white paper is based on the following product levels:
- IBM HTTP Server for OS/390 V5R3
- WebSphere Application Server for OS/390 Standard Edition V3.02

References to OS/390 in the paper also apply to zOS.

Event Sources

Event sources are where information (including status information) is normally recorded about a product or function. Typically these are provided without the need for an additional product (though product customization may be required to activate the event recording). On OS/390 the primary event sources are product logs, SMF, system messages (recorded in the SYSLOG/OPERLOG), SNA alerts, and TCP/IP traps.

Product Logs

IBM HTTP Server

The IBM HTTP Server creates several different logs:

- Server Access Log
- Fast Response Cache Accelerator (FRCA) Access Log
- Proxy Access Log
- Cache Access Log
- Agent Log
- Referrer log
- Server Error log
- CGI Error Log

Of these, the access log can best be used to determine when the HTTP Server was or was not available to serve pages. Each entry in the access log records a request against the server. The entry is time stamped, and shows:
- What was requested
- Who requested it
- The request method
- The file being requested
- A return code indicating the request success or failure

Using this information, different types of availability can be determined. For example:

- When was the server available or unavailable in terms of serving pages? If the HTTP Server normally received one or more requests per minute, gaps in record time stamps that are greater than a minute indicate periods when the HTTP Server was down. In addition, a succession of failure requests for all requests to the server can reveal periods when the HTTP Server was up, but unable to service requests for some reason (a configuration error, missing files, etc.).
- When were specific server resources available or unavailable? Since each request contains the name of requested resource (usually a file, but it can also be the invocation of an executable process such as a CGI program, servlet, or EJB), the log shows when specific resources where available or unavailable to all users, or to particular sets of users.

WebSphere Application Server

WebSphere Application Server also creates additional logs:
- native.log (messages produced by WebSphere C code)
- ncf.log (messages sent to system.out and system.err from WebSphere applications)

The content of these logs will depend upon the applications running in WebSphere Application Server using them to write status and error information. Because of this, these two logs “as-is” may not be the best source of availability information for WebSphere Application Server or its applications. Later in this paper monitoring techniques will be discussed which can be used to log availability information to the ncf.log file.

SMF on OS/390

SMF record type 103 contains configuration and performance statistics for the IBM HTTP Server. The layout and content of these records are documented in the ***IBM HTTP Server Planning, Installing, and Using*** manual. The statistics in these records are very useful for determining HTTP Server performance and required tuning, but not as well suited for availability measurements.

System Messages

IBM HTTP Server

For performance reasons the HTTP Server should be run as a started task under OS/390.
Several key messages are issued to the console and SYSLOG/OPERLOG to indicate the availability status of the HTTP Server:

Message ID	Description
IMW3534I	The HTTP Server task has begin the initialization process.
IMW3536I	The HTTP Server task has completed initialization and can begin processi requests
IMW3535E	The HTTP Server failed to initialize. Errors need to be fixed and the serve restarted before requests can be processed.
IMW3537I	The HTTP Server has been requested to terminate and restart, usually as a result of an operator request.
IMW3538I	The HTTP Server restart was successful, and request processing can resur
IMW3539E	The HTTP Server restart was unsuccessful. Errors need to be fixed and th server restarted before requests can be processed.
IMW3540I	The HTTP Server has stopped processing work and has begun to shut down, usually as a result of an operator request.
IMW354II	The HTTP Server has completed shutdown and the address space is terminating.
IMW3542E	The HTTP Server is ending due to a program check. A dump is taken for problem diagnosis purposes and then it shuts down.

These messages can be captured in real time (using an automation product) or extracted from the SYSLOG or OPERLOG, and used to determine when the HTTP Server was or was not available. For example:

- Messages IMW3536I and IMW3538I indicate availability or “up” status.
- Messages IMW3535E, IMW3537I, and IMW3540I indicate “unavailability” or “down” status.

Calculating the time periods between “up” and “down” messages provides availability and unavailability measurements, in terms of when the HTTP Server was available to process requests. Note that the HTTP Server can be available even though the WebSphere Application Server is unavailable, so these messages by themselves will not give a complete view of availability for WebSphere applications.

WebSphere Application Server

No messages are logged to the console or SYSLOG that report the state of WebSphere Application Server (Messages do appear in the SYSOUT file of the HTTP Server started task, but these are not visible to the console or SYSLOG/OPERLOG). Any messages from WebSphere Application Server to the console have to be generated by applications. Using Tivoli NetView for OS/390 1.2 or later, Java applications (servlets, JSPs, or EJBs) can send messages to the console in the following manner:

1. The application calls the Java **exec** method of the runtime class to send a command to the UNIX System Services (USS) shell.
2. The USS command invokes the **netvcmd** REXX program (supplied with Tivoli NetView), which sends a request to NetView.

3. NetView receives the requests and issues a message using the NetView WTO function.

The ***UpTime*** sample servlet in Appendix A illustrates how this is done.

SNA Alerts

The IBM HTTP Server and WebSphere Application Server do not send out SNA alerts. Alerts can be generated from applications running in this environment (e.g. CGI programs or servlets) by using the USS interface provided by Tivoli NetView for OS/390 V1R2 or later. An alert can be sent by following these steps:

1. The application calls the **netvcmd** REXX program (supplied with Tivoli NetView) and passes to it the NetView GENALERT command with appropriate parameters.
2. NetView receives the command string and executes it.

SNMP Traps

IBM HTTP Server

A SNMP management information base (MIB) and SNMP subagent is supplied with the IBM HTTP Server. This allows any SNMP-capable management system to retrieve status and performance information about the server. Details on implementing the SNMP function are in the *IBM HTTP Server for OS/390 Planning, Installing, and Using* manual.

Currently the HTTP Server does not issue traps. However, the MIB variables can be queried and, in some cases, set. For example, the MIB variable applOperStatus indicates the operational status of the server. An SNMP manager can query this variable at a regular interval and use the query results to determine the HTTP Server availability.

WebSphere Application Server

There is no SNMP support in WebSphere.

Monitoring Techniques

If the event sources do not provide sufficient information for the desired measurements, the following monitoring techniques can be evaluated for use. All require some work to implement, but can provide

greater detail on availability. These techniques all provide real time information on product/function status. This information that can be captured and saved for longer term availability reporting.

Heartbeat

A ***heartbeat*** is a signal at a regular interval indicating that the component is still functioning. When the heartbeat is present -- either logged to a file, or sent as an event to receiving software -- the component is available. It indicates that the component is executing -- but it may not show that it is able to service requests. Absence of a heartbeat indicates the component is unavailable.

IBM HTTP Server

There is no built-in heartbeat function for the IBM HTTP Server. An indirect one can be created from the contents of the server access log. For example, if an automated function requests a URL from the HTTP Server at regular intervals, the access log will record the request whenever the HTTP Server is able to accept them. The server log has to be processed by a program to analyze when heartbeats did or didn't occur, to determine when the server was available. Periods when the automated function was not running would also have to be accounted for.

WebSphere Application Server

As with the IBM HTTP Server, there is no built-in heartbeat function for the WebSphere Application Server. An indirect one can be created in several ways. For example:

- Using the contents of the server access log, as described above for the HTTP Server. The major difference is that the automated function requests a URL that is normally processed by WebSphere. The log information for both the request and the request return code determines if the recorded heartbeat was successful -- a heartbeat recorded but with an unacceptable request code indicates that the HTTP Server was running, but WebSphere Application Server was not.
- From a servlet-created log. Servlets can be autoloaded and started when WebSphere starts. This allows a servlet to write heartbeat time stamp records to a file on the server, or even send them to another platform, while WebSphere is running. As long as the servlet is always running when WebSphere is running, the created heartbeats indicate WebSphere availability.

Ping

A ***ping*** sends an event or query to a target component; the response (or lack thereof) from the target is used to indicate if the component is available. A ping timeout can mean the component is unavailable, or

some other component in path between the ping source and the target is unavailable. This situation must be accounted for when using ping type functions for measuring availability.

The TCP/IP **ping** command is the most well known implementation of this type of monitoring. It conveys that a component's TCP/IP stack is able to communicate and the response time for its "base" communication. It does not report on specific applications or ports. But it is very easy to implement ping-like functions at the port or application level, by knowing what port(s) an application is using and the strings to which it will respond. Sending a string to a port, and measuring the time it takes the response to return, is all the ping function has to do.

IBM HTTP Server

A ping measurement can be implemented for the HTTP Server by sending a string to the server port (normally 80 or, for SSL, 443) and measuring the time it takes to receive a response. A programmed function can, at regular intervals, send an HTTP request to the server port, capture the response, calculate the time it took to receive the response, and use the information to determine if the server is available. The function could send the request using TCP/IP socket programming or, using a language with URL constructs built in such as Java, send a URL to the server port. A timeout waiting for a response, or a response time above a defined threshold, indicates that the server is unavailable.

WebSphere Application Server

WebSphere runs in the same address space as the HTTP Server and receives requests via the same port, so the same technique can be used as described for the HTTP Server. The difference is that the ping request should be a URL the WebSphere normally processes, to accurately determine if WebSphere is running. WebSphere provides the SnoopServlet servlet, which is used to verify that it has been installed properly; this servlet can be used as a "ping", by invoking it from an automation program, determining if a response was received, and measuring the length of time to receive the response.

Remote Command

A ***remote command*** is sent from a monitoring platform to the target component. The results of the command are sent back to the monitoring platform, which looks at the response to determine the component status. The response indicates that the component is running and able to respond to service requests.

IBM HTTP Server

Two types of remote commands can be sent to determine the HTTP Server status:

1. Console commands to determine if the server process is active. On OS/390 the HTTP Server runs best as a started task. The **D A, <name>** command (where <name> is the name of the HTTP Server started task address space) returns whether or not the server is active. The server also supports modify commands, which are documented in the *HTTP Server for OS/390 Planning, Installing, and Using* manual. A command such as **F <name>,APPL=-d stats** (where <name> is the name of the HTTP Server started task address space) returns server statistics; the response (or absence of one) can be used to determine server availability for measurement purposes.

   The console commands can be sent from an automation program or function running on another platform. The ROUTE command can be used if the HTTP Server runs in a sysplex. Tivoli NetView for OS/390 RMTCMD can be used if NetView connectivity exists between the OS/390 monitoring platform and the HTTP Server image. If the monitoring platform is not OS/390, another technique such as TCP/IP socket programming can be used. In all cases, security for sending commands to the target platform must be in place.

2. HTTP commands to the server port to determine if the server process is active. This is the same as sending "pings" to the server port as described earlier, but more information is requested and (hopefully) returned than with a ping function. The HTTP can be URLs normally called via a browser. An automation function or programming language can send the URL request to the server port at a regular interval, capture the response, and analyze it to get detailed server status information.

   The HTTP Server provides a Server Activity Monitor. Details on its use are documented in the *IBM HTTP Server for OS/390 Planning, Installing, and Using* manual. It provides several URLs that can be called to retrieve various server statistics such as request activity, request response times, and network traffic to/from the server. For example, the default URL to access server activity statistics is **http://<servername>/Usage/Initial**. This URL returns an HTML page containing real time statistics, which are updated each time the URL is called.

   Custom URLs can also be written that invoke programs supported by the HTTP Server (such as CGI or GWAPI) to provide information for availability measurements of specific applications or functions within the server.

WebSphere Application Server

There are no OS/390 console commands that interface with WebSphere on OS/390. However, HTTP commands, as described above for the HTTP Server, can also be used with WebSphere. The URLs would invoke WebSphere programs (servlets, JSPs, EJBs, etc.) that return information on WebSphere status. The more detailed the programs, the more information that can be returned about WebSphere.

An example of a command that can be written is the ***UpTime*** Servlet, which is listed is in Appendix A. The servlet performs two functions:

1. When it is first invoked (normally loaded when WebSphere starts), it records the time when it was loaded.
2. When it receives a URL request, it returns the time when it was loaded (which is also the time the WebSphere functions became available), and the elapsed time since it was loaded (which is the length of time WebSphere has been running since it was last started).

Calling the UpTime servlet using a URL such as **http://<servername>/servlet/UpTime** at regular intervals provides a running update on WebSphere availability. A failed URL call (no response returned) is a good indication of WebSphere unavailability.

Additional modifications can be made to UpTime to report WebSphere availability in different ways. Simple changes to UpTime can cause it to take actions such as issuing an OS/390 message or alert at startup and shutdown (using the methods described earlier in this white paper), or monitor and report on the status of back end connections from WebSphere to database or transaction systems.

User Simulation

User simulation attempts to measure availability as an end user would see it. It interacts with the application, either at a keystroke or function level, as a real end user would, and measures the availability and response time of the application. Implementing user simulation is more complicated that any of the other monitoring techniques, but can provide the most accurate measure of the availability experienced by users of applications running on HTTP Server and WebSphere.

For both IBM HTTP Server and WebSphere, implementing user measurements can be done using languages that can easily interact with URLs and the HTTP protocol. Programming a user simulation would include:

- Building the proper HTTP protocol headings and URL
- Invoking form processing by passing the form data a user would normally fill in
- Capturing the HTML response
- Analyzing the response to determine if additional HTTP requests are required
- Measuring the elapsed time from sending a request and receiving a response
- Feeding the appropriate data into a measurement repository (log, relational database, etc.)

Languages such as C, C++, Java, and REXX can be used to implement this function

Another technique, usually as part of a software product or a specialized “script” language, is “screen scraping” or “playback/record”. Actual keystrokes of using the web based application are recorded (or,

if supported, coded into a “script” file), and then played back through a browser. The response back to the browser is captured and analyzed to determine application availability.

The program or script sees only the HTTP protocol or web browser; it does not matter what platform the server runs on. Any language or product with support for these functions and that can connect to an IBM HTTP Server or WebSphere application residing on OS/390 can be used.
Once the user simulation is created, automation products can invoke the program or script at regular intervals to capture the data that will be placed in a measurement repository.

Custom Monitoring Agents

All of the above monitoring techniques can be developed using various programming tools. They are also contained in various software products. These products usually use some sort of agent, running on either the same or different platform as HTTP Server or WebSphere, to invoke one or more of these monitoring techniques and provide information into the data sources to be used for measurements. The products that provide these functions include:

- Event management products, which provide agents to capture events from a variety of components.
- Application management products, which provide APIs to imbed in application code to measure transaction availability and response time.
- Application stress testing products, which provide scripting languages and/or record/playback functions to run against web based applications and produce availability and performance measurements.

Examples of IBM products in these categories are:
- Tivoli Web Services Manager (TWSM), which can measure IBM HTTP Server and WebSphere applications for availability and response time.
- Tivoli Applications Performance Management (TAPM), which provides an API to user for defining transactions in an application and obtaining availability and response time measurements for that transaction
- Tivoli NetView Performance Monitor for IP (NPM/IP), which can monitor the availability and “ping” response time of any IP port, including those used by HTTP Server.

Scalable Server Considerations

The IBM HTTP Server in scalable mode runs in multiple address spaces per scalable system:
- One Queue Manager, which receives the requests
- One or more Queue Servers, which process the requests based on the application environments (work queues) they are defined to service

All of the event sources and techniques described in this paper can be used, with the following considerations:

- In a scalable environment, availability is measured at the HTTP Server subsystem level. So the Queue Manager and (directly or indirectly) each Queue Server is monitored. Availability can be defined as periods when, for a subsystem, the Queue Manager and at least one Queue Server servicing the measured application environment was active and able to process requests.
- Queue Server address spaces do not produce SMF records, or have a console interface, so monitoring or measuring availability using these interfaces is not possible.

Summary

The techniques outlined above are simple actions that provide powerful information. The IBM HTTP Server on OS/390 and WebSphere on OS/390 are becoming critical components for providing end-to-end service for increasingly important business applications. Collecting measurement data from the event sources or monitoring techniques described above, and merging it with existing availability data from the operating system, network, database systems, and transaction systems, will provide a total end-to-end availability picture than can be managed and improved as business requirements dictate.

Appendix A - UpTime Servlet

The following is example code for a servlet that reports WebSphere start time and elapsed time. When called using the URL http://<servername>/servlet/UpTime, it returns the following:

UpTime Servlet loaded at: 2001.02.19 23:32

Uptime Servlet uptime: 1176111016 milliseconds

Uptime Servlet uptime: 326:41:51

The first line shows the date/time the servlet was loaded; since it was defined to load when WebSphere started, it shows when WebSphere began to execute requests.

The second line shows the elapsed time since the servlet was loaded, in milliseconds.

The third line shows the elapsed time since the servlet was loaded, in HH:MM:SS.

This version uses the Tivoli NetView for OS/390 USS command interface to send messages to the console when it is load and when it is unloaded (shutdown).

```
/*UPTIME SERVLET (Author: Mike Bonett)

This servlet performs the following functions:
  - When loaded, writes a message to the System Log
  - When invoked, returns the amount of time since loaded (if loaded at
    WebSphere Application Server (WAS) startup, this is the amount of time WAS
    has been active)
  - When destroyed, writes a message to the System Log. If only destroyed
    when WAS stops, indicated that WAS is down
  - This is an example. No warranties intended, implied, or assumed. Use at
    your own risk. Modify to your hearts' content.


*/

import Java.io.*;
import Java.util.*;
import Java.text.*;
import javax.servlet.*;
import javax.servlet.http.*;

public class UpTime extends HttpServlet
{
 long StartupTime;
 Date startdate;
 DateFormat df;
```



```
 public void init(ServletConfig config) throws ServletException
        {

         //The init method is run once, when the servlet is first loaded.
         //This is where the current date/time is obtained and used as
         //the WAS startup time, and where the startup message is
         //is sent to the SYSLOG via Tivoli NetView for OS/390
         super.init(config);
         df = new SimpleDateFormat("yyyy.MM.dd HH:mm");
         StartupTime=System.currentTimeMillis();
         startdate = new Date(StartupTime);

         Runtime rt = Runtime.getRuntime();
         Process pr=null;
         try
            {
             pr=rt.exec("netvcmd TESTWTO WAS001I WebSphere Application Server
                         Has Started at " + df.format(startdate));
             int rc = pr.waitFor();
            }
         catch(Exception e)
            {
             System.out.println("UpTime init error: " + e.getMessage());
            }
         }

 public void doGet(HttpServletRequest req, HttpServletResponse res) throws
ServletException, IOException
        {

         //The doGet method is run for every URL request.
         //The elapsed time since startup is calculated and
         //Returned in a simple HTML page.

         long Ctime, Dtime, htime, mtime, stime;

         Ctime = System.currentTimeMillis();
         Dtime = Ctime - StartupTime;
         htime = Dtime / 3600000;
         mtime = (Dtime - (htime * 3600000)) / 60000;
         stime = (Dtime - ((htime * 3600000) + (mtime *60000))) / 1000;

         res.setContentType("text/html");
       PrintWriter out = res.getWriter();
         out.println("<HTML><BODY><P>");
         out.println("UpTime Servlet loaded at: " + df.format(startdate));
         out.println("<P>");
         out.println("Uptime Servlet uptime: " + Dtime + " milliseconds");
         out.println("<P>");
         out.println("Uptime Servlet uptime: " + htime +":"+ mtime + ":"
+stime);
         out.println("</BODY></HTML>");
        }
```



```
 public void doPost(HttpServletRequest req, HttpServletResponse res) throws
ServletException, IOException
        {

         //If UpTime is called with the HTTP Post method, invoke doGet
         doGet(req,res);
        }

 public void destroy()
        {

         //This method is executed when the server is destroyed (unloaded).
         //Write out a “shutdown” message to the SYSLOG using
         //Tivoli NetView for OS/390

         Date stopdate = new Date(System.currentTimeMillis());

         Runtime rt = Runtime.getRuntime();
         Process pr=null;
         try
            {
             pr=rt.exec("netvcmd TESTWTO WAS999I WebSphere Application Server
                         Has Been Stopped at " + df.format(stopdate));
             int rc = pr.waitFor();
            }
         catch(Exception e)
            {
             System.out.println("UpTime init error: " + e.getMessage());
            }
         }

}
```

Appendix B. Reference Publications

- HTTP Server for OS/390 - Planning, Installing, and Using (SC31-8690)
- WebSphere Application Server Standard Edition Planning, Installing, and Using (GC34-4806)
- An Introduction to Tivoli NetView for OS/390 V1R2 (SG24-2254)
- Measuring End-to-End Availability: How To Get Started (White Paper available at http://www.ibm.com/support/techdocs)