DESIGNING FOR HIGH AVAILABILITY: MEASUREMENTS

By Floyd Piedad in conjunction with Harris Kern's Enterprise Computing Institute

Availability: A User Metric

Availability is measured from *the user's point of view*. A system is available if the user can use the application he needs - otherwise it is unavailable. Accordingly, availability must be measured *end-to-end* - all components needed to run the application are available. Many IT organizations mistakenly believe that availability is simply equal to main server or network availability. Some may only measure the availability of critical system components. These are grave mistakes. A user may equally be prevented from using an application because his PC is broken, or his data is unavailable, or his PC is infected with a computer virus. IT organizations that subscribe to a narrow, or undisciplined, availability mindset go through several stages of alienation from their users.

User unhappiness is the first and least severe stage. Users simply express unhappiness with poor system availability. The IT organization may either recognize a problem or deny it, citing their host or network availability statistics as proof. Those who deny the problem's existence bring their organization to the next stage of user alienation.

User distrust is characterized by user disbelief in much of what the IT organization says. Users may begin to view IT's action plans as insufficient, or view the IT organization as incapable of implementing its plans. They gradually lose interest in helping IT with end user surveys and consultations. IT organizations that can deliver on promises and provide better availability *from the user's point of view* can prevent users from moving to the next stage of user alienation.

User opposition is the third stage of alienation. Here, users do not merely ignore IT plans - they begin to actively oppose them, suggesting alternatives that may not align with IT's overall plans. Users start to take matters into their own hands, researching alternatives that might help solve their problems. The challenge for the IT organization is to convince users that the IT plan is superior. The best way to meet this challenge is to conduct a pilot test of the user's suggested alternative, then evaluate the results hand-in-hand with users. In contrast, we have seen some IT organizations react arrogantly, telling users to "do what you want, but don't come crying to us for help." These organizations find themselves facing the final stage of user alienation.

User outsourcing is the final stage of user alienation. Users convince management that the best solution lies outside the IT organization. Outsourcing can take the form of hiring an outside consultant to design their system, going directly to an outside system supplier, or even setting up their own IT organization. At this stage, users have completely broken off from the IT organization, and reduced — if not totally eliminated — the need to fund it. Beyond user alienation, there are other serious side effects:

- **Failure to identify root causes of availability problems** — If only a few components are considered when system availability is evaluated, the root causes of the outages may well lie in components whose availability is not monitored. We have seen several banking IT organizations that have denied the existence of Automated Teller Machine problems by pointing out that their mainframes, switches, and network are always available. They fail to observe that the ATM machines themselves cause most ATM outages.
- **Conflicts between IT divisions** — Many IT organizations usually delegate critical elements of their systems to individual groups within IT. Each then measures the availability of its assigned area, without correlating it with the availability of other areas. This leads to territorial disputes where one group blames others for poor system availability. "Don't blame my group, our network was up 100 percent of the time…"
- **Expensive and ineffective remedial measures** — If you do not know what the root cause of a problem is, you'll probably spend money on the wrong solution. Or, you'll concentrate on improving only *your* assigned system component, without regard to overall system availability.
- **Inability to determine true system health** — Availability measurements of each component cannot easily be "added up" to reveal true system availability. Ninety-nine percent host availability + 99 percent network availability + 99 percent database availability is not equal to 99 percent system availability. Outages in each area usually occur at different times, and each outage in any component brings the entire system down. In this example, actual system availability can be anywhere from 97 percent to 99 percent.

Why do many IT organizations fall into the trap of measuring only a few system components and not actual end-to-end availability? There are two reasons.

First, it is easier to measure a few system components. Few tools are available for analyzing and monitoring end-to-end system availability. Many tools measure network or host availability, but few actually check for application outages from the perspective of the user. Second, it is easier to achieve higher availability on a per component basis since outages rarely occur repeatedly on the same component. Outages for different components usually occur at different times but may all affect the availability of the system to the user, resulting in far worse availability statistics.

Measuring End-To-End Availability

To accurately estimate end-to-end application availability as experienced by end users, you must first thoroughly understand the system's configuration; all the components and resources used by the application, both local and remote; and the hardware and software components required to access those resources. Here is an example:

Sales Personnel Call Management System

Local resources	Sales personnel data
	Call reports
Remote resources	Contact management data at each sales reps' computer
Hardware components	Personal computer, LAN adapter, LAN cabling, network switch, print server, network printer
Software components	Windows 98, MS Access, contact management software, call management application

The next step is to monitor all these components for outages. If outages are detected on multiple components at the same time, treat the outage duration as just one instance. To calculate end-to-end availability, add all the outages of each component. Then, apply the formula presented earlier in this chapter.

Sounds easy in principle, but taxing in practice? Definitely. That's why you need to automate measurement as much as possible. The simplest way is to use a tool that monitors availability of local and remote resources from a user's PC. This tool regularly attempts to get a response from the resources in question, and records times that critical resources are unavailable. More advanced tools can query an application for problems or execute certain tasks on the application. If the application fails, an outage is recorded. This approach does not identify the source of the problem, but the error condition may help support staffers identify the cause.

There is a great demand for automated end user system availability monitoring tools — utilities that can be installed in user workstations that would periodically test the applications for availability. In the absence of such tools, you would have to resort to random sampling of users' availability experiences.

You won't get precise measurements of *every* user's availability experience - that's unrealistic. Do, however, recognize that users have an availability requirement you must pay attention to. Don't get too dependent on technical measurements for rating your performance - ultimately, what matters most is that users are happy with the service that the IT organization provides.