BEYOND FIVE-NINES

A WHITE PAPER ON DESIGNING A VOIP NETWORK FOR APPROPRIATE AVAILABILITY

BY: ROBERT BELLMAN, PRINCIPAL, BROOKTRAIL RESEARCH AND MARK LANGANKI, CHIEF ARCHITECT, SPANLINK COMMUNICATIONS

SPANLINK

Designing a VoIP Network for Appropriate Availability

"Five-nines" is often cited as the pinnacle of system or network availability. A system rated at five-nines is available 99.999% of the time, or conversely, unavailable only about 5.26 minutes per year. However, the method for calculating the availability of voice systems was defined when circuit switching ruled the roost. Now that packet-switched voice over IP (VoIP) is taking over, the goal of attaining five-nines—or better—is still valid, ***but traditional availability calculations no longer apply.***

VoIP systems are generally more distributed than circuit-switched systems, and they can include redundancy and parallel structures that legacy systems cannot support. This creates more opportunities to build in high availability and even to exceed the availability of legacy systems. It also calls for a new approach to evaluating availability. Since the business productivity of a VoIP solution depends on the entire system and not just a central device, availability calculations must take the entire system into account, including all of its hardware, software and service components.

The system approach to VoIP availability is a powerful business tool. By considering the entire system and all its components, the system approach gives enterprises the information they need to accurately evaluate the availability of their VoIP solutions. Using the system approach, IT managers can set availability goals based on both system performance and life-cycle costs and then determine the most cost-effective way to maximize VoIP availability and boost business productivity.

The Reality of Five-Nines

"Availability" is the likelihood that a device or system or network will be available for use. How likely is it that a system component will be working rather than out of service? How likely is it that a user can make a connection or complete a transaction? Availability is usually expressed as a percentage. A system that is always up boasts 100% availability. A system that is down 10 out of every 100 hours has 90% availability.

Availability is often expressed as some number of nines. A system rated at five-nines is up 99.999% of the time and down 0.001% of the time. In other words, a five-nines system is down no more than 5.26 minutes in a 365-day year. Those 5.26 minutes can be all at once or a few seconds here and there over the course of the year. Table 1 shows availability ratings from one- to five-nines. Traditionally, five-nines is considered a worthy goal—less than an hour of downtime every 10 years. But five-nines is not what it used to be.

Table 1. Availability by the "Nines"

Nines	Availability	Downtime Minutes/Year
Five-nines	99.999%	5.26
Four-nines	99.99%	52.6
Three-nines	99.9%	526
Two-nines	99.0%	5,256
One-nine	90.0%	52,560

The math behind availability ratings is simple. The basic availability calculation takes the following form:

$$A = MTBF / (MTBF + MTTR)$$

where A = availability, MTBF = mean time between failures, and MTTR = mean time to repair. (MTTR includes not only the time to fix the broken device or system but also the time required to bring it back into service—reinstall the hardware, reboot the software or whatever.) For example, if your television goes two years on average between failures and takes a week on average to repair and return to your family room, its MTBF is 730 days and its MTTR is 7 days for an availability rating of 99.05%.

730 / (730 + 7) = 0.9905

But what is the business significance of the availability formula? What does it mean when it's used to rate an enterprise VoIP network, for example? In fact, simple availability calculations don't have much business significance. They offer some insight into the performance of an individual device or network component, but they don't look at the whole system. And business productivity depends on the availability of whole systems. If the VoIP PBX is up but the network between the PBX and workers' phones is down, no one can make phone calls.

Availability measurement has roots in the telecommunications industry. Service Providers and their equipment suppliers rate the availability of all kinds of network devices, everything from central office Class 5 switches to premises-based PBXes. To ensure uniform comparisons, telco availability measurement and calculation methods are based on Telcordia (formerly Bellcore) standards. But the Telcordia standards focus on the core system and draw an "availability boundary" that excludes factors such as power outages, preventative maintenance, network loss and more. So a system based on a five-nines device may really deliver only two- or three-nines availability.

Traditional availability ratings, therefore, have little business value. They are merely box-level operational numbers with little relevance to today's distributed, multi-device VoIP solutions. They do not provide a basis for answering important questions like: Where should I add redundancy to my VoIP system? Is additional redundancy worth the expense? Which availability improvements will be most cost-effective? Clearly, a better method is needed for evaluating the availability of VoIP solutions—a method that takes all of today's technology and business requirements into account.

AVAILABILITY IN THE VOIP WORLD

Until fairly recently, most voice solutions were based on highly centralized designs—a single device that hosted all the intelligence and complexity surrounded by relatively "dumb" peripheral devices. An enterprise voice network, for example, comprised a PBX in a closet or equipment room with twisted-pair cables radiating to desktop phones in each office (Figure 1). One or more DS1/T1 tie lines would link the PBX to the public switched telephone network (PSTN). In these solutions, the availability of the central device was fundamental to system availability. If the PBX crashed, none of the phones worked. If a single desktop phone broke, no one else suffered. So it made sense to focus on the availability of the central device, even though the availability boundary and five-nines calculation excluded "external" components.

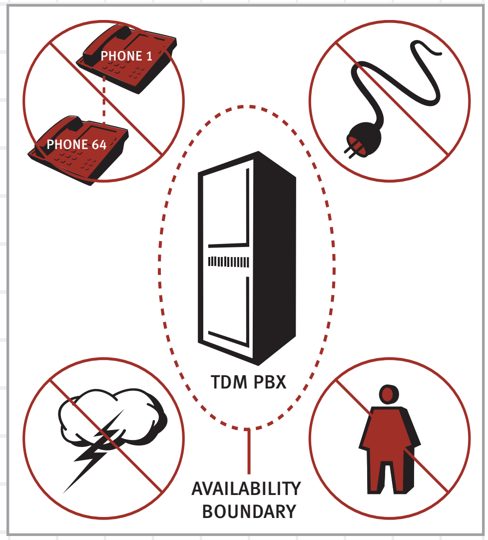
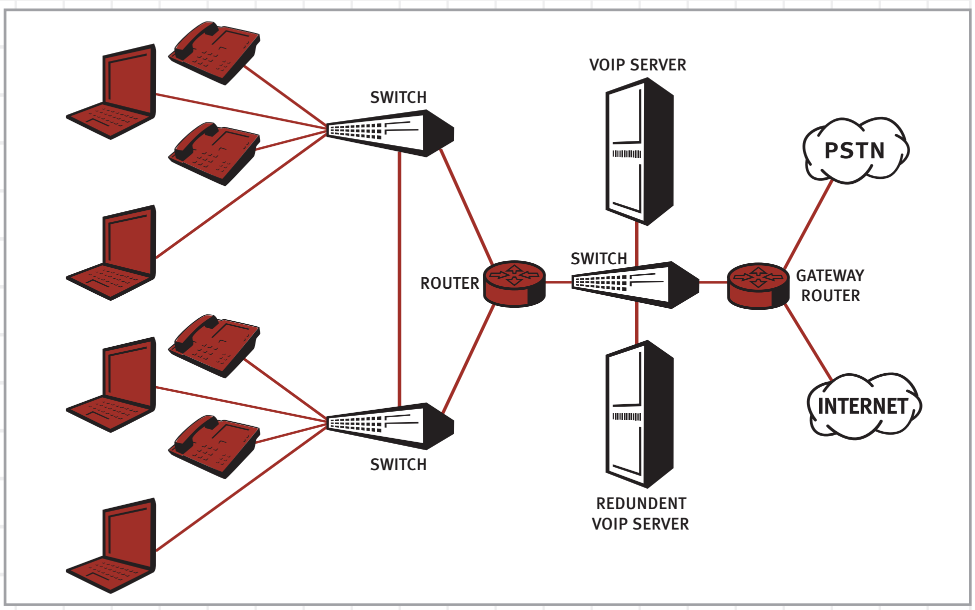


Figure 1. In a Traditional Enterprise Phone System availability is limited to the central device

Figure 2. In an Enterprise VoIP System availability must be measured for the entire system



VoIP solutions are different. Intelligence is distributed to devices throughout the enterprise network: routers, firewalls, media gateways, LAN switches, IP phones, the VoIP PBX, and so on (Figure 2). Each device is itself a complex hardware/software system. Moreover, VoIP systems are inherently more resilient than legacy voice systems. Because VoIP is packet-switched rather than circuit-switched, it's easier to create parallel structures that eliminate single points of failure. If, for example, a cable between LAN switches is cut, the switches will automatically send packets over an alternate path. To be useful, therefore, availability calculations have to consider both the components and the topology of the VoIP solution. With VoIP, business productivity depends on the entire system, not just specific devices.

CALCULATING SYSTEM AVAILABILITY

The first step in calculating the availability of a VoIP system is to determine the availability of individual devices: every router, LAN switch, gateway, server, PBX, IP phone, etc. For most devices, the availability must take into account both hardware and software. This inventory should also include any external services that are part of the overall solution: the PSTN, IP VPNs, and so on.

Next, the availability of individual components must be combined. The combination formula depends on whether the components operate in series or in parallel. Figure 3 shows two light bulbs that operate in a series. If either bulb burns out, the electrical circuit is broken and the combined subsystem fails. Thus the availability of a serial subsystem is the product of the availability of its components:

$$A_s = A_1 * A_2$$

where A_s = availability of the subsystem, A_1 = available of component 1, and A_2 = availability of component 2. The availability of a serial subsystem is always less than the availability of its components. If, for example, A_1 = 99.99% and A_2 = 99.0%, then A_s = 98.99%.

Figure 4 shows two light bulbs that operate in parallel. If either bulb burns out, there is still an intact electrical path through the subsystem and the other bulb stays lit. The formula for parallel availability is as follows:

$$A_p = A_1 + A_2 - A_1 * A_2$$

where Ap = availability of the parallel subsystem, A_1 = available of component 1, and A_2 = availability of component 2. The availability of a parallel subsystem, therefore, is greater than the availability of its components. If A_1 = 99.99% and A_2 = 99.0%, then A_p = 99.9999%.

Using these formulas, the availability of two-component subsystems can be calculated. Then pairs of subsystems can be combined to find the availability of larger subsystems and so on until the availability of the entire solution has been calculated.

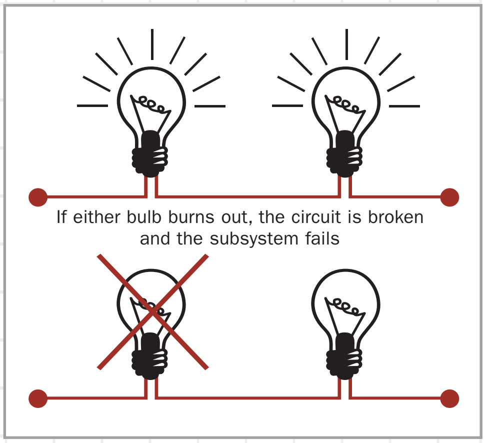


Figure 3. A Serial Subsystem

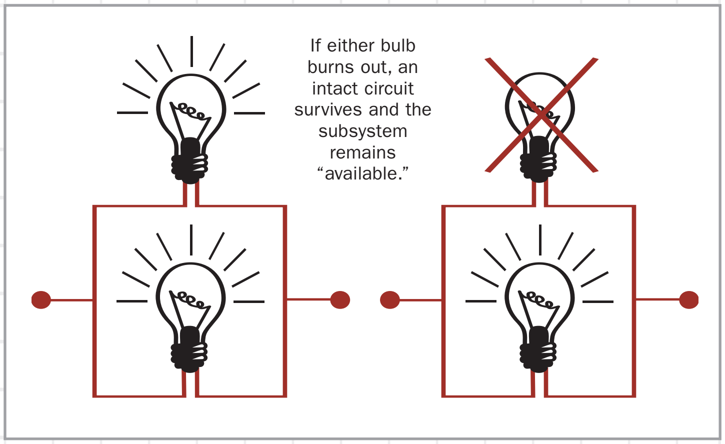


Figure 4. A Parallel Subsystem

For example, Figure 5 shows a simplified client-server network. The client workstation connects to the central router by way of a single LAN switch. The server connects to the router through parallel LAN switches. To calculate the availability of the network, including the server, we first calculate the combined availability of the parallel switches. At 99.999%, the combined availability is better than the 99.997% availability of the individual switches. Since the rest of the network elements are in series, we then multiply their individual availabilities together to get an overall availability of 99.952%, somewhat less than the availability of the least available element (Figure 6).

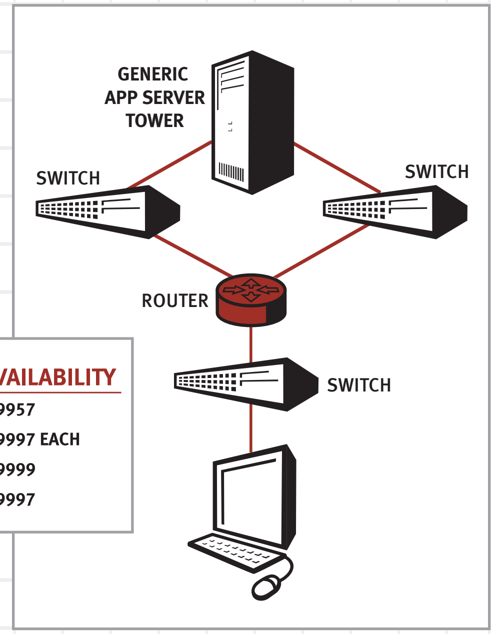
FROM AVAILABILITY TO PRODUCTIVITY

Clearly system availability is more meaningful than device availability, especially as it pertains to complex VoIP solutions. System availability relates directly to business productivity and to how users perceive the system. If the system is up, workers are productive. If the system is down, productivity suffers regardless of which component caused the outage. Furthermore, system availability doesn't make excuses. It doesn't draw an availability boundary that excludes essential components like wide-area network service from the calculation.

Many organizations have quantified the cost of system downtime, usually by quantifying the benefit of system uptime. Consider, for example, an airline call center that operates 24x365 and takes reservations worth $1.9 billion annually. If peak hour revenue is $1.52 million, per-minute revenue ranges from an average of $3,615 to a maximum of $25,333. If system availability is rated at three-nines (99.9%), total annual outages will be around 526 minutes for potential revenue losses of $1.9 million to $13.3 million per year, depending on when the outages occur. Improving system availability to four-nines (99.99%) will reduce average annual outages to 52.6 minutes and cut potential revenue losses by 90%, down to $190,000 to $1.3 million per year.

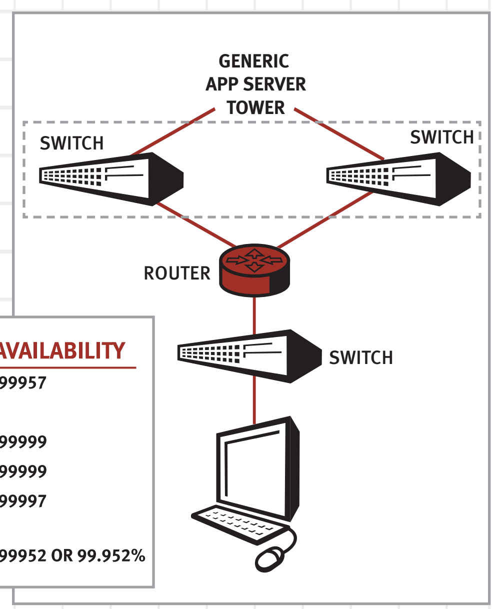
Or consider a small manufacturing company that is open 40 hours per week and takes $1.3 million annually in phone orders. In this case, average revenue—or downtime cost—is just $10 per minute. A system with two-nines (99.0%) predicted availability would have an estimated loss potential of about $52,560 annually. Boosting availability to three-nines (99.9%) would reduce estimated losses to about $5,256.

Increasing the availability of a VoIP solution, therefore, can have direct benefits in terms of improved enterprise productivity. But the enhancement process is a tradeoff. Adding a parallel LAN switch, for example, entails both capital expense (CAPEX) to purchase the device and operating expense (OPEX) to install and manage it. So a number of questions must be asked and answered when designing or upgrading a system for availability: What's the business value of system uptime? What's the business cost of system downtime? How much would it cost to reduce system downtime? Is the added productivity worth the cost?



DEVICE	AVAILABILITY
SERVER	.99957
LAN SWITCHES	.99997 EACH
ROUTER	.99999
LAN SWITCH	.99997

Figure 5. A Simplistic Client-Server Network



DEVICE/SUBSYSTEM	AVAILABILITY
SERVER	.99957
PARALLEL LAN SWITCH SUBSYSTEM	
.99997 + .99997 – (.99997 * .99997) =	.99999
ROUTER	.99999
LAN SWITCH	.99997

SYSTEM AVAILABILITY = .99957 * .99999 * .99999 *.99997 = .99952 OR 99.952%

Figure 6. System Availability Calculations

System availability calculations help enterprises answer such questions and make informed decisions about system design. Enterprises can use these calculations to set practical business goals, to determine what availability level makes the most business sense, to decide what design gives them the most bang for their infrastructure buck.

System availability is not a terribly useful concept if considered in isolation, that is, without tying it to revenue goals, operating costs and other business matters. System downtime has a cost. In an airline reservation center, downtime means missed sales and unhappy customers. At a help desk, downtime can drive customers to the competition. In a hospital, downtime can have the direst of consequences.

For many enterprises, the availability goal is clear. Hospitals, air traffic controllers, and other critical activities need non-stop service. They need a solution that delivers five-nines at a minimum. Other organizations may be satisfied with reaching an optimal cost/benefit balance—the point at which the cost of additional availability improvement outweighs the benefit. In the airline call center example, going from three-nines to four-nines could cut annual losses by as much $12 million, making it relatively easy to cost-justify system enhancement. In the manufacturing example, going from two-nines to three-nines saves around $47 thousand per year—perhaps not worth the cost. Still, organizations of all types want to find the most cost-effective way to reach their availability goals, and system availability techniques provide the tools they need.

THE AVAILABILITY LIFECYCLE

By incorporating system availability techniques into an "availability lifecycle" approach, enterprises can reach the optimal balance between productivity and cost in VoIP solutions (Figure 7). The lifecycle begins with a problem statement and a solution design. The designer weighs CAPEX and OPEX estimates against projected system productivity and sets an availability goal, say three-nines (99.9%). The designer then selects components—hardware, software and network services—and configures them as a system that meets, at least on paper, the specified goal.

Availability calculations take into account all the components and all of the parallel and serial subsystems.

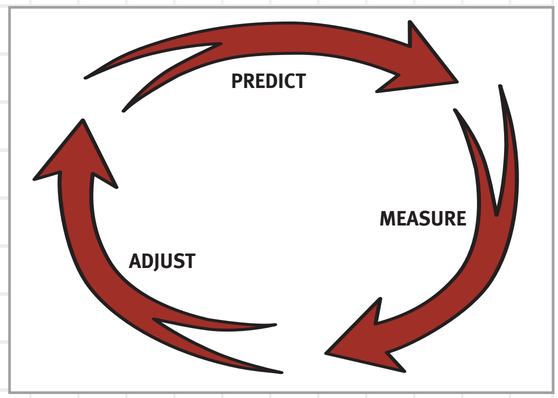


Figure 7. The Availability Lifecycle

Once the system is installed and operating, all business-affecting downtime events are measured and recorded. The events and their elapsed times are accumulated in "buckets" according to cause (see Table 2). Total downtime is used to calculate overall system availability and compare it to the design goal. The example in Table 2 shows total downtime of 1,914 minutes in the month of June for a system availability of 95.569% (assuming 24x7 operation). Since the original goal was three-nines, the system needs adjustment. A review of the June outages provides guidance on how to improve availability. In the example in Table 2, human error accounts for over 73% of the downtime, and power outages account for 21%. So the most promising next steps are to improve human factors with additional training, application of best practices and so on, and to upgrade the power system. Once the human factors and power supply have been "adjusted," the cycle begins again.

Say, for the sake of this example, that downtime measurements in August yield the numbers shown in Table 2. System availability has increased to 99.693%—a worthy improvement over June, but still short of the three-nines goal. Since hardware and network outages are now the dominant failure modes, the solution is adjusted again to add hardware redundancy and parallel network paths. Availability calculations help system designers compare the cost/benefit tradeoffs of potential enhancements and select the most cost-effective options. Downtime measurements in October show that the system, at 99.917% availability, now meets the desired goal. Periodic re-measurement and, if needed, further adjustments will ensure that the solution continues to meet that goal and will help to identify further opportunities to improve business productivity.

TABLE 2. AVAILABILITY LIFECYCLE EXAMPLE

Downtime for Month Type of Outage	June Minutes	August Minutes	October Minutes
Hardware	53	56	0
Software	0	0	0
Software bugs	23	21	20
Network	33	30	2
Environmental	0	10	5
Human error	1,403	20	10
Power	402	0	0
Total	1,914	137	37
System Availability	95.569%	99.693%	99.917%

CONCLUSION

Today's enterprise VoIP solutions offer attractive new productivity features that traditional circuit-switched voice solutions cannot match. But the distributed, decentralized nature of VoIP calls for a new way to evaluate availability and to relate it to business value. The system approach to availability helps enterprises estimate the availability of entire VoIP systems rather than just individual devices. And the availability lifecycle provides a methodology for turning availability estimates into reality. Together, the system approach and the availability lifecycle give enterprises the tools they need to construct VoIP solutions with the most cost-effective balance between availability and productivity.

About Spanlink Communications

Spanlink Communications is a leading provider of REAL customer interaction solutions that leverage VoIP technology.

A Cisco ATP Certified Channel Partner, Spanlink has unmatched experience with Cisco IP communications solutions and is a leading developer for contact center, collaboration and IP communications system management solutions for public institutions and medium and large enterprises.

Spanlink is leading provider of Cisco-based IP Communications solutions. Spanlink provides support services that span converged networks, contact center products and customer self-service applications. Our solutions deliver superior design and efficient deployment process and more practical management models. Spanlink goes beyond the requisite steps to get high-standards for business-focused solution design, post-cut system integrity and system support. Spanlink's expertise allows you to concentrate on your key business and enjoy a faster return on the efficiencies of your new Customer Interaction Solution.

For more information on Spanlink Communications and our IP Communications services and products, visit www.spanlink.com, telephone 800-303-1239 or email Mktg@Spanlink.com.

SPANLINK