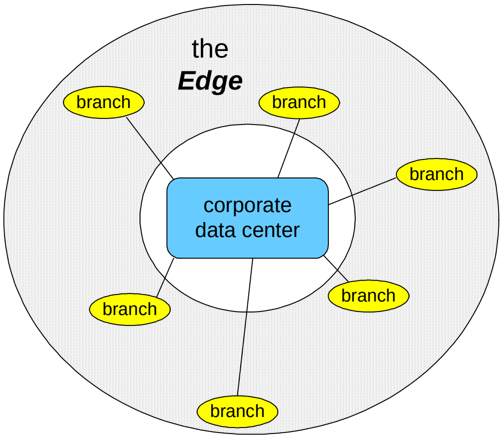
the Availability Digest

Stratus' Avance Brings Availability to the Edge

February 2009

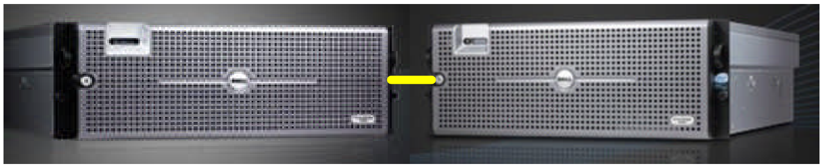
Business continuity has not yet been extended to the Edge. What is the Edge? It is everything outside of the corporate data center upon which the IT services of a company rely. These are the branch offices of an enterprise – the bank branches, the retail stores, the sales offices.



IT services in a branch (or in any small to medium-sized business, for that matter) are typically provided by one or more local servers hidden in a closet. If one of these servers goes down, the operations of the branch or small business can be seriously hampered until the failed server is restored to service. This can take hours or days, and such an outage certainly represents more than an inconvenience to the company and to its customers. However, the costs associated with such downtime often do not warrant the expense of providing fault tolerance at these locations.

Avance™, from Stratus Technologies (www.stratus.com), seeks to overcome this problem. Avance[1] brings high availability to the Edge. It also brings an added capability – virtualization. Not only can the servers sitting in that closet be highly available, but the various applications can also run on virtual machines hosted by a single, highly-available server, perhaps reducing branch IT costs significantly.

Announced on June 10th, 2008, Avance provides an out-of-the-box, fault-tolerant virtualization solution with no high-availability or virtualization skills required of the operations staff. No special hardware is required – it runs on a pair of standard x86 servers interconnected by an Ethernet link. One server acts as the primary node and the other as its backup node. From a deployment and management perspective, Avance creates a single-system image so that the operations staff sees only a single server.



Using the open-source Xen hypervisor to provide a virtualized environment, Avance supports both Windows and Linux virtual machines (VMs). Avance continually monitors the health of the

[1] "Avance" is French for "advance." In addition, the "AV" signifies availability and virtualization.

physical servers and the virtual machines running on them. It takes immediate and automatic corrective action if it detects a fault. The backup server seamlessly takes over the functions of the faulty VM or processor for most faults. In the worst case, a catastrophic server crash can take up to two minutes to recover. Avance provides over four 9s of availability (less than an hour of downtime per year).

Planned downtime is eliminated as upgrades can be rolled through the system one node at a time.

Avance was awarded the 2009 Technology of the Year Award by InfoWorld in the category of Platforms and Virtualization.

Where Does Avance Fit?

Avance is applicable to any environment in which the availability of data-processing services is critical but in which the cost of downtime does not warrant a fully fault-tolerant solution. These environments include corporate branches, ISV products, and small to medium businesses (SMBs).

In addition to high availability, Avance allows environments that host multiple servers to reduce their number of servers by consolidating them as virtual machines onto a virtualized host server.

Edge Applications

Edge applications are the extensions to a data center. Running in a remote location, they exchange information with the data center. Such locations include:

- bank branches
- retail stores
- manufacturing facilities
- distributed warehouses
- departments
- sales offices
- distribution hubs
- hospital clinical systems

Vertical Markets

ISVs (independent software vendors) provide a plethora of products to various business areas. Products include those that support legal offices, medical practices, health care, and public safety (911, fire, police, and emergency medical services).

Small to Medium Businesses (SMBs)

Any SMB that has a number of general-purpose servers, each running its own application, can benefit from the high availability and virtualization of Avance.

Data Centers

Avance even has application to data-center services that are not mission-critical to the enterprise. In effect, Avance reduces the severe pain of downtime to mild discomfort.

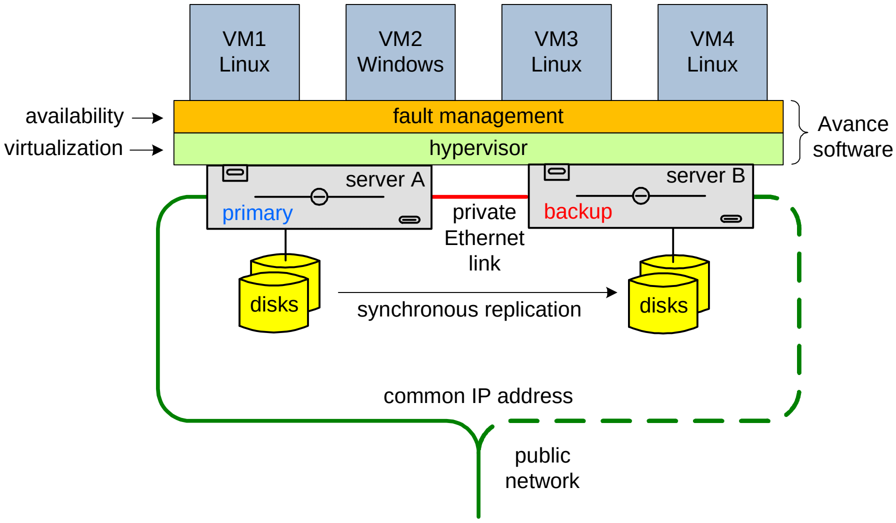
The Avance Architecture

Hardware

Avance is a software product that requires two industry-standard x86 servers interconnected by a dedicated Ethernet cable. That's it. No storage area network (SAN) is required – Avance uses the local direct-attached disks of each server.

The servers must be certified by Stratus but in principle can be any server supported by the Xen hypervisor. Currently, most Dell and HP servers are certified for use by Avance.

Avance creates a single-system image of the server pair. One server (arbitrarily chosen initially by Avance) is the primary server, and the other is its backup. Once Avance is installed, all operation and management of the server pair is seen to the operations staff and the users as that of a single server.



Avance Server

The Ethernet connection is used for two purposes. One is to exchange health information between the two servers. The other is to synchronously replicate changes being made to the primary server's disks to the backup server's disks. The separation between the two servers can be up to 0.5 kilometers.

Up to five public networks can be connected to the Avance server. A public network is connected via appropriate network devices to both servers. Both servers share the same IP address for a public network. However, a public network is driven only by the primary server.

Software

Avance comprises an embedded Xen bare-metal hypervisor[2] and an embedded fault-management layer. It is installed directly on the two servers without any intervening operating system.

The Xen hypervisor supports multiple-core processors. Xen has been hardened by Stratus and has been modified to support Avance's fault-management services. In addition, all certified device drivers have been hardened by Stratus.

[2] Fault Tolerance for Virtual Environments – Parts 1, 2, and 3, *Availability Digest*; March, April, and June, 2008.

Avance maintains synchronization between the primary and backup disks using synchronous replication. Replication is at the cache block level. No change can be made to the source cache until that same change has been replicated to the backup cache. In this way, it is guaranteed that the two disk systems are always identical.

Virtual Machines

Avance can support up to eight virtual machines (VMs). Each VM can support a different guest operating system. Supported guest operating systems include Windows Server 2003, Red Hat Enterprise Linux, and CentOS open-source Linux.

Fault Management

Fault Monitoring

Using predictive fault detection, Avance continually monitors the health of the primary and backup servers and takes immediate action if it detects any anomaly. If it does detect a fault or a potential fault in the primary server, it will initiate a failover to the backup server if the backup server is in better health. Its goal is to keep the logical server alive regardless of faults within it.

Avance monitors the health of the two servers via heartbeats sent over the private Ethernet link. Faults are rated by severity and affect the health score for the server. Such faults might include the failure of a disk unit, the failure of a fan, the failure of a power supply in a server with dual power supplies, or the failure of a link to a public network.

Backup server faults that leave that server still operational are simply reported. If the fault is in the primary server, and if the backup server is in better health, Avance will fail over to the backup server, as described later.

A particularly onerous fault is the failure of the private Ethernet link. Server-health information can still be exchanged via one of the public networks, but the backup disk cannot be synchronized. In this case, an attempt to fix the problem is to restart the backup server, as many faults are transient software faults and can be corrected by rebooting. A restart can take on the order of five minutes, but it is transparent to the users since the other server is still running the applications.

If the restart is successful, the backup database is resynchronized; and high availability is restored. If the restart is unsuccessful, the backup is declared down.

A crash of the primary server is a catastrophic fault. In this case, all work in progress is lost; and the backup server must take over. It takes twenty to thirty seconds for the backup server to take over, at which point the VMs are rebooted. As soon as the VM reboot is complete, the system is returned to normal operation. Except for the unlikely simultaneous failure of the two servers, this is the worst case for recovery times. All other failovers impact the users for a minimal amount of time, typically unnoticed by the user, with Linux performing better than Windows in the current release.

Fault Reporting

Stratus provides 24x7 monitoring of its systems in the field. This is accomplished via the *Call Home* feature built into every system.

Call Home uses one of the public networks to communicate warnings, faults, and recovery actions initiated by the system. If the problem requires a parts replacement, Stratus will immediately ship that part to the customer. In many cases, a replacement part arrives at the customer's site before the customer is even aware of the problem.

For critical software problems, Call Home will send logs, traces, and memory dumps. Stratus personnel can access the system remotely to try to determine what is going wrong and to propose a fix.

Some companies are reluctant to give a third party direct access to their systems. In these cases, the debugging data can be copied to a memory stick and transmitted to Stratus for analysis.

Failover

There are several failover scenarios to describe. In all cases, failover to the backup server will occur only if the backup server is in better health than the primary server. For instance, if a RAID disk and one of the dual power supplies on the backup server is down, a fan failure on the primary server will not cause a failover.

Backup Server Fault

If a fault is detected in the backup server, there is no impact to the users. A fault such as the failure of a disk in a RAID array or the failure of a power supply (if the server has dual power supplies) simply results in a Call Home for a replacement part. The backup stays in service.

Should the backup server experience a VM failure, the VM is restarted.

A more serious fault such as the failure of a network adapter leading to the loss of connectivity to a public network or a server crash will cause the backup server to reboot. There is a chance that such a fault is transient and will be cured by a reboot. However, if the calculated mean-time-before-failure of the failing component drops below a specified threshold, the backup server will be declared down.

Noncritical Primary Failure

If the primary server should experience a noncritical failure, such as a power supply failure in a server with dual power supplies, a Call Home report will be sent. The failover to the backup server will begin, as described below, so long as the backup server is in better health.

Primary Public Network Failure

Should the primary server lose its connection with a public network, communication will continue by using the private link to make use of the backup's network connection. A failover to the backup server will be initiated.

Primary Disk Failure

If a disk on the primary server should fail, Avance will use the synchronized copy of the disk on the backup server by accessing it via the private link. Avance will initiate the failover procedure to the backup server, as described later.

Private-Link Failure

Should the private link fail, the system is running in split-brain mode. Both the primary and backup servers are still operational, but they cannot talk to each other. An attempt is made to correct the problem by rebooting the backup server just in case this is a transient software problem. After repeated attempts, if the calculated mean-time-before-failure exceeds a specified threshold, the backup server is declared down.

Primary VM Failure

If a VM should fail on the primary server, it is simply restarted.

Primary Server Crash

The most catastrophic fault is the crash of the primary server. All work in progress is, of course, lost. The backup server takes over processing. However, though all of the software is loaded and ready to go on the backup server, the VMs must be booted and the applications started. This can typically take on the order of thirty seconds to two minutes.

This is downtime that is observable to the users. In addition, during a primary-crash recovery, Avance does not preserve user connections. However, many of the ISV products are designed to automatically reconnect when the backup server comes into production.

The Failover Process

Avance will decide to fail over gracefully to the backup server if the primary server's health should become less than that of the backup server. Failover is a very controlled process and is substantially transparent to the user. Failover proceeds as follows.

First, the software state for each VM is migrated from the primary server to the backup server by copying its memory over the private link. Avance keeps track of which memory pages have changed and copies only those pages. Memory copying is done in the background, and users continue to be serviced by the original primary server. Software state migration occurs simultaneously and independently for each VM. Some VMs will complete before others.

When most of the memory for a VM has been copied, the VM is paused while the final pages are copied (to prevent continued changes to pages). The VM on the backup server is then started; and the application continues on the backup server, which has now become the primary server for that VM. The user connections are preserved during the failover process.

This process typically takes about ten seconds per gigabyte of memory used by the VM. However, it is transparent to the users. The only user impact is at the end of the copy for a VM when processing is paused while the new VM is started. This time is typically fast enough to generally be transparent to users.

Eliminating Planned Downtime

The Avance failover process can be used to eliminate planned downtime by rolling upgrades through the system. To upgrade the servers to new software versions or to upgrade the hardware, all that needs to be done is to first take down the backup server, upgrade it, and return it to service. The failover process is then used to switch the primary and backup roles, and the other server is then upgraded.

Except for the brief switchover pause, this upgrade process is totally transparent to the users.

Recovery

When a failed server is to be returned to service, the Avance software must be reloaded and the database of the failed server resynchronized with the primary server. Avance will automatically migrate the VMs and the applications from the primary server to the server being restored. It can then be put into operation as the backup server.

Depending upon the size of the database, database resynchronization can take a few hours, but the users are running off of the primary server during this time.

Monitoring

Operator errors are the predominant cause of system downtime in redundant systems. Therefore, Avance provides a system monitor that is aimed at reducing operational steps. In fact, once Avance is installed and running, there is no need for any operator action that would not otherwise be needed for a single server.

Avance's monitor provides a single-system image of the Avance system. So far as the operator is concerned, he is dealing with a single, industry-standard server.

The monitor shows the current status of the hardware components, the VMs, and the networks. It displays fault information down to the component level. Generally, there is no operator action that is required except for hardware-component replacement, as Avance handles all faults automatically. It will automatically alert operations personnel by email if there is a problem that requires attention.

The monitor also maintains a log of all automatic actions undertaken by Avance and of all user actions.

The monitor is browser-based and is intended to be used remotely. Therefore, there need be no IT staff present at an Avance installation. Only authorized personnel can use the monitor.

Installation

Avance is designed to simplify and minimize the installation process. The first step is to install Avance on each server. Avance installation requires that the user answer only one optional question – whether or not to change the default IP address of the server, if desired. Avance installs in ten to fifteen minutes.

The VMs must then be set up. It takes about two minutes to create each VM. The guest operating systems must then be installed. This takes about five to ten minutes for Linux and about twenty to thirty minutes for Windows. If multiple VMs use the same guest operating system, that operating system need only be installed once.

As each VM is created, the user specifies the resources that it will use. They include the number of cores, the amount of memory, the disk capacity, and the number of disk spindles. The servers used to configure Avance do not have to be identical. However, the smallest server must be sufficient to host all of the running VMs. Note that with the capability to roll upgrades through the system without affecting users, the Avance servers can be upgraded at any time.

Purchasing Avance

Stratus sells Avance directly to large customers. In addition, value-added ISVs are integrating Avance into their own products; and system integrators are incorporating Avance into their service offerings.

An Avance system can be licensed directly from Dell. An Avance license costs $2,500 per server. In addition, Windows and/or Red Hat Linux licenses are required for each server. CentOS, being open source, carries no license fee.

Availability Choices – Good, Better, Best

Avance complements Stratus' fault-tolerant offering, ftServer, which recently has exhibited over six 9s availability in field measurements (an average downtime of 30 seconds per year).[3] With Avance's four 9s of availability, companies have a choice of a range of availabilities and costs. A simple example will serve to illustrate this.

First, though, we would like to comment on Stratus' claim of an Avance availability of greater than four 9s. An availability of four 9s translates to 48 minutes per year of downtime. Given that Avance soft failovers impose barely perceptible downtime., and hard failovers take the system down for about two minutes, it is difficult to see how an Avance system could have that much downtime. Stratus says that it is simply being conservative until it has enough field experience with Avance to give a more accurate value. We at the Availability Digest suspect that the actual Avance availability will approach five 9s, or five minutes per year of downtime. In the following example, we will use an estimate of Avance downtime of ten minutes per year.

Let us assume that a company is running an industry-standard x86 server with a sales price of $3,500. It would like to move to a high-availability environment by buying a backup server. The company estimates that the primary server will fail an average of twice per year and that the failover to the backup server will take two hours. The company now has three high-availability options:

Value	Configuration	Downtime	Cost
Good	Active/Backup – two servers	4 hours per year	$7,000
Better	Avance – two servers plus two Avance licenses	10 minutes per year	$12,000
Best	ftServer	30 seconds per year	$25,000

The proper availability option depends upon the cost to the company of downtime. In this example, if downtime cost is insignificant, it should probably stick with an active/backup configuration. If the cost of downtime is $1,000 per hour, Avance will pay for itself in a little over a year. If the cost of downtime is $40,000 per hour, ftServer will pay for itself after the first server failure in an active/backup configuration. Compared to an Avance system, ftServer will pay for itself in this case in two years.

Since many Edge applications have a downtime cost measured in thousands of dollars per hour, Avance is well-positioned to pay for itself rapidly in these environments.

About Stratus

Stratus is the leading provider of high-availability and fault-tolerant products for industry-standard servers. Located in Maynard, Massachusetts, in the U.S., Stratus has been delivering fault-tolerant systems since 1980. Starting with its classic Continuum servers, which provide five 9s of availability, Stratus' current offerings of its four-9s Avance and its six-9s ftServer give it a broad portfolio of high-availability and fault-tolerant products.

Stratus also provides high-availability professional, management, and support services through its CALM (Continuous Availability Lifecycle Management) portfolio.

Summary

Avance brings high availability to the Edge – the branches and small businesses that to date have been reluctant to invest in high-availability environments. Avance is a software product that sits on a pair of standard x86 servers interconnected by a dedicated Ethernet link. It uses

[3] This includes Stratus hardware and software incidents. See Stratus' Availability Meter on www.stratus.com.

predictive fault detection and can protect itself from most faults with little if any impact on the users of the system. In addition, it provides a virtualized environment that can support up to eight virtual machines running Windows or Linux as guest operating systems.

If the cost of downtime of an Edge application is as little as $1,000 per hour, Avance can pay for itself very quickly, perhaps in a year or so.