

Darryl La Gace

Sophisticated Fault-Tolerant Architecture Keeps Applications Available

By Darryl La Gace, Director, Information Systems, Lemon Grove School District and Wayne Enseki, Technical Consultant, ECS

It's 7:55 a.m. and school is about to begin. Students are milling about the quad with wireless tablets. Two students are putting the finishing touches on a science presentation so they can send it in before the bell rings. As class begins, students take their seats and login to begin the day's lessons. The teacher takes attendance on-line as a call from the office rings on the IP telephone. It's a message for little Johnny letting him know his mother went on-line and deposited the money for his lunch account. It's not even 8:15 a.m. and your network and its resources have probably made a zillion transactions.

The morning scenario described above is becoming ever more present in our own school district. With a 1:2 computer/student ratio in every class and an instructional delivery system that relies heavily on technology, it has to work. In today's demanding environment for information anytime anywhere, it's more critical than ever that the networks we design are built for high availability and fault tolerance. So, last year we took on the task of building just such an IT environment. Not only were we looking for fault tolerant hardware, we wanted to take it to the next level and build fault tolerant data centers. We had already suffered a disaster and although the recovery plan worked effectively, it prompted us to rethink our entire design. After realizing the potential disastrous issues that could have occurred had we lost our network, we weren't content with just two node clusters in the same building anymore.

Lemon Grove School District worked with ECS and its industry-leading technology partners to build a highly available architecture that enables mission-critical applications to

be continuously accessed by the more than 10,000 students, teachers, parents, administrators and other community end-users. The focus of this article is on the backend Windows and Storage Environment as it is probably the most complex and important piece of the high availability puzzle, but we will also provide a high-level overview of the rest of the architecture for reference.

Network Design

The design process began by clearly defining our user requirements. We realized that in order to achieve the goals of our highly available architecture, the means to access data and services needed to be redundant and fault tolerant. (We address in detail how redundancy was achieved within the data centers in the next section). However, for clients to access fully fault tolerant data centers, we needed to establish redundancy in our communication infrastructure.

Our remote sites connect to the data centers via a tertiary location. Each of the remote sites has two network paths to the access point with two specific site connection types. First is a triangle configuration where a site has a primary link direct to the access point and an alternate path to the second site. That second site also has a primary link direct to the access point. Our second site connection type has two paths direct to the access point. These sites can also act as intermediaries for downstream sites in which case our downstream sites have two connections to their host site.

In order to maintain the connection state and ensure fault tolerance, we utilized a distance vector routing protocol that dynamically identifies the best path to the access point. Since connectivity needs to be seamless to the user community and function

within the latency specifications of our applications, we established metrics during the design process to use as a guide during deployment. Based on our design, the detection of link state, link selection and transition time became critically important. We selected EIGRP to provide this function. Some modification to the protocol settings was needed to meet the specific requirements of our environment.

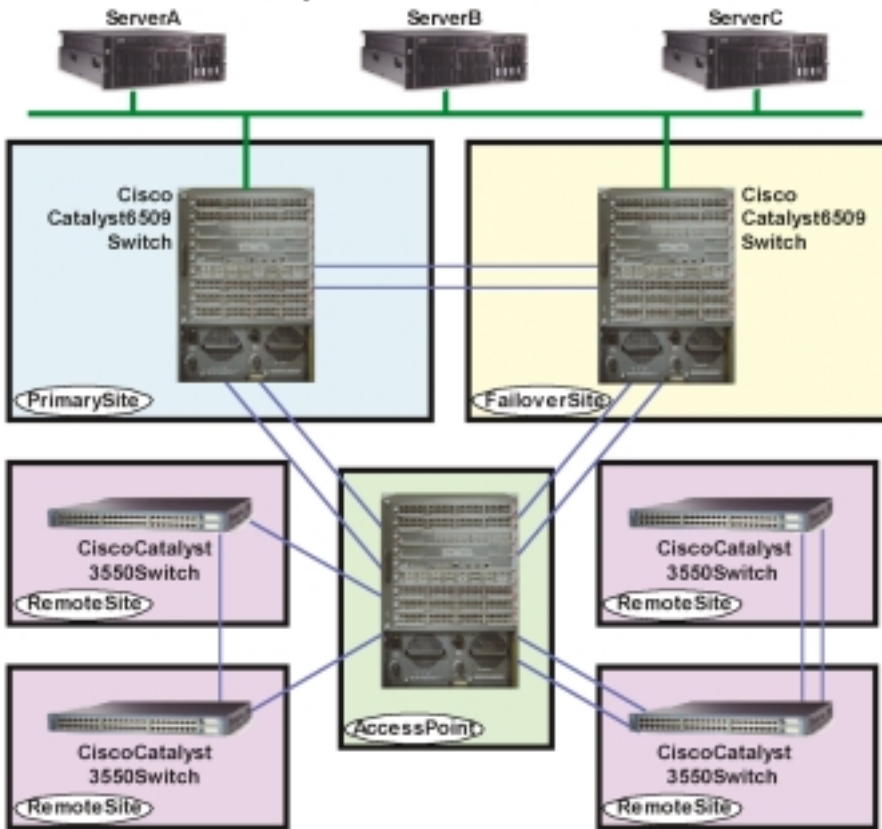
Internet access was established with selected service providers attached to our network on separate external routers. These connections are then filtered through two clustered firewalls (4), and passed to the access point. We determined that having these connections available through the access point would alleviate unnecessary traffic within the data centers and provide better Internet response time. By selecting two ISPs with access through different telephone company offices, we were able to extend our design requirements right out to the Internet.

The following picture on facing page illustrates the network design:

Building a Better Architecture

The new Lemon Grove School District high-availability architecture can be best described as a geographically distributed cluster with real-time, synchronous data mirroring between sites. The noteworthy feature of this solution is that it can survive multiple system failures (up to an entire site) while continuing to provide high-levels of application availability and do so automatically with little or no intervention by IT operations staff. This architecture is unique in its ability to survive a continuum of faults from a component failure to full disaster recovery with unprecedented ease and transparency.

Campus Network Architecture



The architecture utilizes building blocks of two-node clusters and distributed SAN storage, where each application server and its associated storage are located in separate data centers. Under normal circumstances, the system operates as a typical cluster that is simply 'stretched' across two data centers; one located at the district office and the other located at the middle school a few blocks away. However, utilizing a combination of data-mirroring technology and a distributed SAN and LAN, each site is fully capable of operating independently of each other. Should a disaster strike either site, the surviving site will be able to provide services with little or no interruption and no loss of data.

The following picture at right illustrates the basic architecture:

The following picture on page 16 is an example of the capability of the architecture to survive multiple simultaneous failures while still providing near-continuous application availability:

Even in this extreme example, if everything in the diagram were to fail at the same time, Exchange would only experience a short outage while

the cluster restarted the services on the second node. This recovery would also occur automatically and it would be completely transparent to Exchange that all of the other hardware failures occurred.

This sophisticated solution was able to achieve the following functionality:

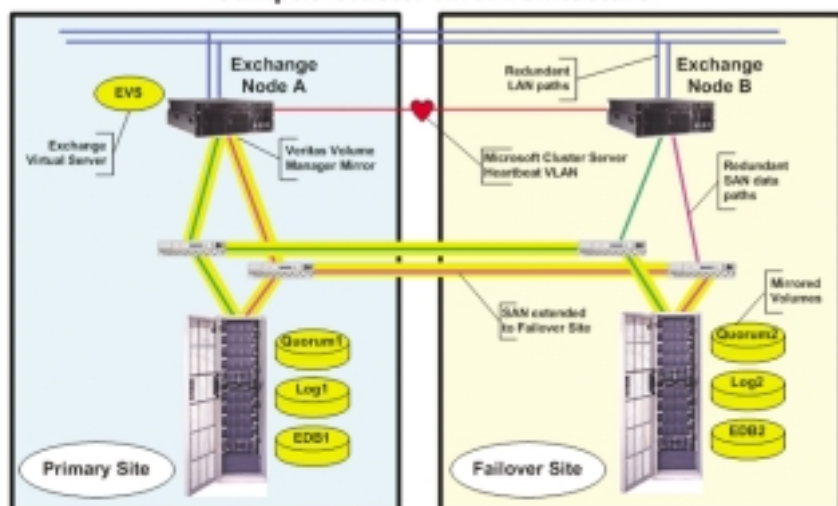
- Failure of any single disk subsystem, LAN switch, SAN switch, or any individual hardware component in the LAN, SAN, storage, or servers would not result in an interruption in service. Failures at this level would be completely transparent to the end-users, who would experience continuous availability mission-critical applications.
- Failure of any single-server or multiple-core components within a site would result in an automatic failover and only a brief interruption in service. If, for example, the active Exchange server failed or if both the SAN and the LAN failed simultaneously, the cluster would automatically detect and move the application resources to the other data center. End-users would experience a brief delay while the application resources restarted on the failover node.
- Failure of an entire site would constitute a disaster and recovery procedures would be implemented. However, through the use of physically distributed cluster nodes and the complete mirroring of all business critical data between sites, the remaining site could be brought back online in a matter of minutes using automated recovery procedures.

Implementation Strategy

The following is a more detailed description of the technical architecture from the bottom-up:

Continued on page 16

Campus Cluster H/A Architecture



Storage Subsystem

The storage consisted of four Compaq/HP StorageWorks EMA12000 storage subsystems with 24 Fibre Channel controllers and 504 disk drives for a total of approximately 28 terabytes of raw disk capacity. The storage subsystems were configured for both maximum availability and performance and utilized redundant controllers and RAID1+0 hardware-based disk mirroring. The failure of a storage controller or disk drive would be handled transparently at a hardware level without any interruption of disk I/O.

Storage Area Network

The SAN utilized four Brocade 3800 Fibre Channel switches operating at 2Gbps for maximum performance. To achieve no single points of failure, these switches were implemented as two independent SAN fabrics and each fabric connected redundant host bus adapters in the servers to the redundant storage controllers. To support the site-to-site failover and disaster recovery, the SAN extended between the two data centers over multiple, redundant single-mode, long-wave fiber cable runs. This extended, high-performance SAN architecture was the foundation for achieving the resulting failover solution and without it, neither the extended cluster servers nor the site-to-site disk mirroring would have been possible.

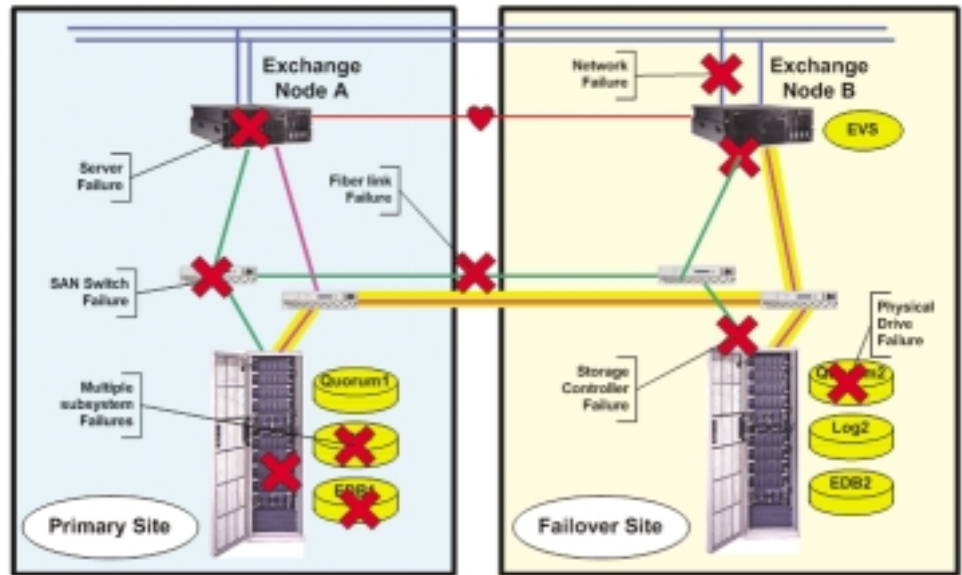
Microsoft Cluster Servers

A total of 10 HP/Compaq DL580, DL380, and DL360 servers were used for the Exchange, SQL, File, Print, and Profile servers. Each server was configured for maximum component redundancy with dual hard disks, power supply, network connections, and SAN host bus adapters. The servers run Windows 2000 Advanced Server and are configured in pairs as five two-node cluster servers. Leveraging the SAN fabric that extends between the two data centers, it became possible to implement geographically dispersed clusters – also known as a ‘campus’ or ‘stretch’ clusters – where each cluster node is in a physically different location. Through this architecture, it became possible to achieve automatic site-to-site failover of services running on the cluster.

Veritas Volume Manager

Veritas Volume Manager is a software upgrade to the Windows 2000 Disk Management capabilities. What this enhanced logical volume manager pro-

Multiple Fault Resilience



vided was the ability to use dynamic disks and software-based volume mirroring in a Microsoft Cluster. When used in combination with the extended SAN, it became possible to implement RAID1+0 mirrors that transparently mirrored NTFS volumes in one site with other volumes located in the second site. This software mirroring solution provided continuous data synchronization across geographically separated disk subsystems along with the capability to fail these volumes between cluster nodes in both sites. The extremely high levels of transparent and automatic fault-tolerance were achieved through the use of this technology.

Systems Management Software

Critical to a high-availability architecture is the capability to detect, notify, and act on events and failures that occur in the data center. The systems management solution was built around Microsoft Operations Manager (MOM) that serves as the central event correlation manager and integrates into the individual platform management capabilities of the various hardware, software, and operation system components.

Local Area Network

The data-center backbone consists of three Cisco Systems 6509 switches. One provides access from the remote sites (located in a tertiary location), while the remaining switches are located in each data-center. Each switch was configured with redundant hardware including supervisor modules, Ethernet port capacity and power supplies. All of the switches are connected using multiple redundant single mode,

long-wave fiber optic cables which have been attached to Gigabit interfaces and “trunked” together. Similarly, each cluster node also uses two separate LAN connections to its local switch. By leveraging redundant paths, the physical connectivity mitigates issues related to the failure of devices, modules or physical links to ensure uptime and availability.

Routing Protocol Selection

Enhanced Internet Gateway Routing Protocol (EIGRP) was selected to provide route information to the remote sites about the nature of the network. The data-centers are configured as a series of single IP subnets. A server or cluster node located in Data-Center One will be on the same IP subnet as its counter part located in Data-Center Two. Since each data-center is connected to the network access point via Gigabit trunks, the routing protocol advertises the data-center connections as equal cost routes, allowing traffic to be load balanced between each site. In the event of a link failure between the network access point and a data-center, traffic will continue to flow down the alternate path without interruption to the end user. This distance vector protocol provides the same seamless route selection for remote site connectivity to the network access point. This allows traffic to use the fastest available connection without a disruption of service.

Continued on page 32

Advertiser's INDEX

Sophisticated Fault-Tolerant Continued from page 16

Considerations in Implementing Fault-Tolerant Solutions

In order to successfully implement a fault-tolerant, high availability infrastructure solution, it is important to consider the following before starting:

- Understand the distinctions between various levels of high-availability as they pertain to the end-user**
There are many, sometimes subtle, distinctions between vendor terms such as 'high-availability', 'continuous-availability', 'no-single-point-of failure', 'transparent fail-over', etc. Rather than get caught up in technical details, what ultimately is most important is what the end-user experiences when a failure occurs. Have the solution defined in terms of the impact on the end-user or software application when a failure occurs, since this is ultimately how IT SLA's are measured.
- Clearly define the pros and cons of the various technology options**
Ultimately, the systems architecture will be determined by the best combination of dozens of trade-offs. Functionality, implementation cost, and performance will vary widely across the various implementation options. As one example, to replicate data from site-to-site, you must choose between I/O driver-based, O/S-based, application-based, SAN-based, or storage controller-based data replication technologies as well as synchronous or asynchronous I/O. Unless you clearly understand the pros and cons, it is difficult to determine which option best meets your business needs.
- Take a holistic approach to the architecture**
High-availability systems architecture is comprised of multiple layers of integrated component architectures that must operate as a whole. To achieve the desired end-state functionality, not only must each technology be understood, but its ability to inter-operate with other technologies while retaining the desired level of availability must be preserved. The system is only as good as the weakest link.
- Document the current physical infrastructure and develop a technology roadmap**
Which specific architectural solutions are technically and financially feasible depends on the current network, server, and storage infrastructures as well as the actual distances between disaster sites. To avoid implementation and integration difficulties, develop an architectural roadmap that starts with the current environment and clearly defines future implementation phases. Always document infrastructure changes to stay current.

Conclusion

Building a fault-tolerant, high-availability architecture was essential to ensure that Lemon Grove School District end-users remain connected and productive. Through careful attention to the numerous issues involved in building a back-up data center, selection of the right technology partners, and thorough implementation planning, the Lemon Grove School District is able to provide protection against catastrophic events and allow 100 percent uptime for its 10,000 community end-users. ■

The author's Web sites can be accessed at www.lgsd.k12.ca.us and www.thinkecs.com.

| | |
|---|-------|
| 365 USA..... | 6 |
| AAL (Admin. Assist. Ltd.) | 34-35 |
| Action Learning Systems Inc. | 31 |
| Allied Telesyn | 9 |
| Arey, Jones Educational Solutions | 33 |
| Chancery Student Management Solutions | 23 |
| Decotech Systems Inc. | 36 |
| Eagle Software | 29 |
| Edupoint Educational Systems | 24-25 |
| Excelsior Software/Pinnacle Plus | 30 |
| Internet Software Sciences | 32 |
| Lightspeed Systems | 18-19 |
| Microsoft | 2 |
| Planware Systems, LLC | 4 |
| QSS | 13 |
| Spectrum | 17 |
| Starnet Data Design | 3 |
| VL Systems | 27 |

Need HELP?

The Simple Solution
Automate your help desk!

Web + Center Web Based Help Desk



- Customer Self-Help/Knowledge Base
- Customer & Case Mgmt Tracking Tool
- Customizable & Extensive Reporting Ability
- Designed for the Education IT Support
- 20% of CA Community Colleges use Web+Center
- Web+Center presented at 2002 CETPA Conference
- FREE (3) Tech Versions Available
- 30% Educational Discount

Internet
Software
Sciences www.inet-sciences.com
13851 Flomont Pines Lane - Los Altos, CA 94022
(650) 949-0942 FAX (650) 917-0913