Lotus Domino Clustering Options for IBM Netfinity Servers

By Stephen Brunner

Jeffrey Conley

Scott Purdy

Gary Sullivan

Abstract

This White Paper presents the advantages of clustering, the clustering options available in a Domino environment and an example of when to use each type of clustering.

Customers using Lotus^R Domino^{TM} as the basis for mission-critical applications need to ensure that these applications are highly available and scalable. There are a number of clustering options customers should consider for accomplishing this task.

Overview of Clustering Benefits

Clustering technology isolates the client from any knowledge of the physical hardware of the cluster and in theory, any number of servers could be providing services. This isolation helps protect the client from changes to the cluster hardware and software. This technique minimizes interruptions to service providing perhaps the most important benefit of clustering: **high availability**. Servers in clusters function as highly available versions of unclustered servers.

What is a cluster?

In simple terms, clustering is the linking or grouping of a set of servers so that they appear as one resource from the client perspective. In this context, the cluster functions as a "single" provider of resources, enabling client requests to be processed in a timely manner regardless of whether any given server is unavailable or too busy at the time the request arrives.

Another huge benefit of clustering is **scalability**. Additional servers can, in most cases, be added into a cluster to support new applications or additional users of existing applications. In contrast to large single systems, good clustering technology performance can scale in a near-linear fashion.

Data backup and archiving are traditional methods of protecting data. Clustering offers another benefit by providing an alternative method of protecting vital data.

Domino Clustering Overview

Two of the most popular clustering approaches are operating system (OS) level and application level, such as Domino Clustering. Both of these approaches provide certain aspects of high availability for Domino applications.

Domino Clustering is a feature of the Domino Enterprise Server that provides high availability and workload balancing for Domino applications. The Domino administrator configures a group of servers into a cluster. Then, if one of these servers becomes unavailable or overloaded, the Notes client includes the capability to locate an equivalent copy of the application on another server in the cluster.

For some customers, Domino Clustering provides all the aspects required whereas in other cases, OS clustering is sufficient. And in some cases, customers need to employ both Domino Clustering and OS clustering to achieve all their high-availability requirements. Thus, Domino Clustering and OS clustering are actually complementary approaches to high availability for Domino applications.

In Domino 5.0, failover and load balancing are also supported for Web clients accessing the Domino Web server via the Internet Cluster Manager (ICM). Domino clustering is easy to install, administer and manage through the use of the Domino Administration client or a browser and provides cross-platform high availability and scalability of the Domino servers in your enterprise.

OS clustering is available for most of the OS platforms on which the Domino server operates. Examples of OS clustering are the Microsoft^R Cluster Service, available for Windows NT^R High Availability/Clustered Multiprocessors (HACMP) for AIX^R, and Sun Clusters for Solaris. While these products have different attributes and capabilities, their basic features and operation are very similar. Netfinity^R Availability Extensions for MSCS (NAE) is a new clustering technology that extends the capabilities of MSCS up to an eight-node solution for IBM Netfinity platforms, thus offering significant enhancements over the base MSCS two-node support.

In OS clustering, all resources needed by the application are "virtualized," and can be moved between servers in a cluster. The OS cluster then monitors the status of each server and application running in the cluster. If a failure occurs, resources can be relocated to another server in the cluster and the application restarted.

Clustering options in a Domino environment

When considering Domino servers, several clustering options are available. The advantages of each are reviewed below.

OS level clustering

OS clustering can make scheduled agents highly available. Scheduled agents are generally designated to run on a specific server. Prior to Domino R5.02,* if that server fails, even when it is a member of a cluster, that server's agents are not run until the server is restarted. Since OS clusters quickly restart a failed server, scheduled agents remain highly available. (*The Domino R5.0.2 server has the capabilities to fail over Scheduled Agents through the Synchronous New Mail Agent Facility not available in previous releases)

For the most part, the strengths of OS clustering stem from fact that the failed server is quickly restarted. All server functions are completely restored. The failover of resources is transparent to the user.

OS clustering can make server tasks highly available. Certain operations of the Domino server are performed by tasks running within the Domino Server. Examples are a fax server, pager gateway, or ccMail MTA. If the server running these tasks fails, their service is no longer available to users. Domino clustering currently provides no means to fail over these tasks to another cluster member. In an OS cluster, these tasks can be automatically restarted along with the Domino server (using Program documents or Notes.ini settings).

OS clustering supports "hot failover" and makes failure transparent. With Domino clustering, if users have an application open when a server goes down, they generally get a Server Not Responding error on the next operation performed. Failure is not triggered within the application -- the user must exit the application, often losing any unfinished/unsaved work, and then reenter the application in order to trigger failover. With OS clustering, failover can occur at any point in the application (as long as work has been saved either at the client or in the database on the server ... and only data stored in memory on the server is lost when the server failed), making failover virtually transparent to the end user.

Domino Clustering

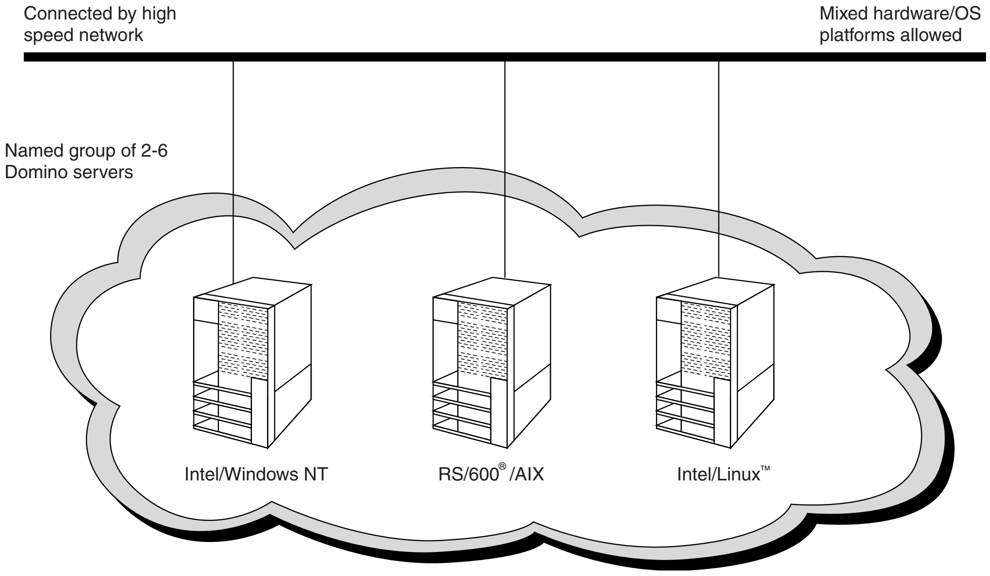
Domino clustering supports six servers in a cluster (six is the current number certified/supported by Lotus -- the clustering software is capable of handling more than six nodes). Some OS clustering products support only two servers per cluster. Having six servers in a cluster offers greater flexibility in distribution of resources in the cluster. It also allows better load redistribution in the case of a failure, since the load of the failed server can be spread across up to five servers, instead of a single server.

Workload balancing ensures that no server in the cluster is overloaded and thus greatly enhances the scalability of Domino applications.

Domino clustering supports workload balancing in addition to failover. Most OS clustering products provide support for failover only. Domino monitors the workload of all servers in the cluster, and if the workload reaches an administrator-defined setting, Domino Clustering redirects new application requests to another server in the cluster.

Domino Clustering supports truly heterogeneous clusters, such as cluster members can run on different hardware and OSs. By its nature, OS clustering requires cluster members to be running the same OS, and often requires similar hardware.

Domino Cluster


Connected by high speed network
Mixed hardware/OS platforms allowed
Named group of 2-6 Domino servers
Intel/Windows NT
RS/600® /AIX
Intel/Linux™

Domino clustering requires no special hardware. Domino Clustering can often be implemented using the customer's existing hardware (assuming sufficient capacity). OS clustering often requires special hardware support for features such as disk sharing and failover.

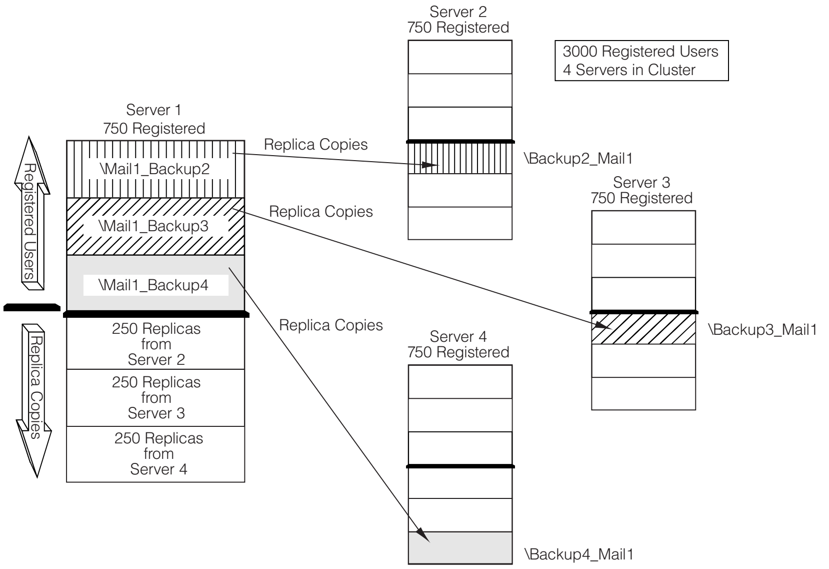
Domino clustering has no distance limit between nodes. This allows not only for high availability and load balancing but basic hot site functionality as well.

Domino clustering supports special Domino-specific availability settings for Domino resources. These include settings to make servers or databases temporarily unavailable to users so that the administrator can perform maintenance activties. OS clustering does not provide these Domino-specific features.

Failover does not occur in the following cases (Domino 5.0.2):

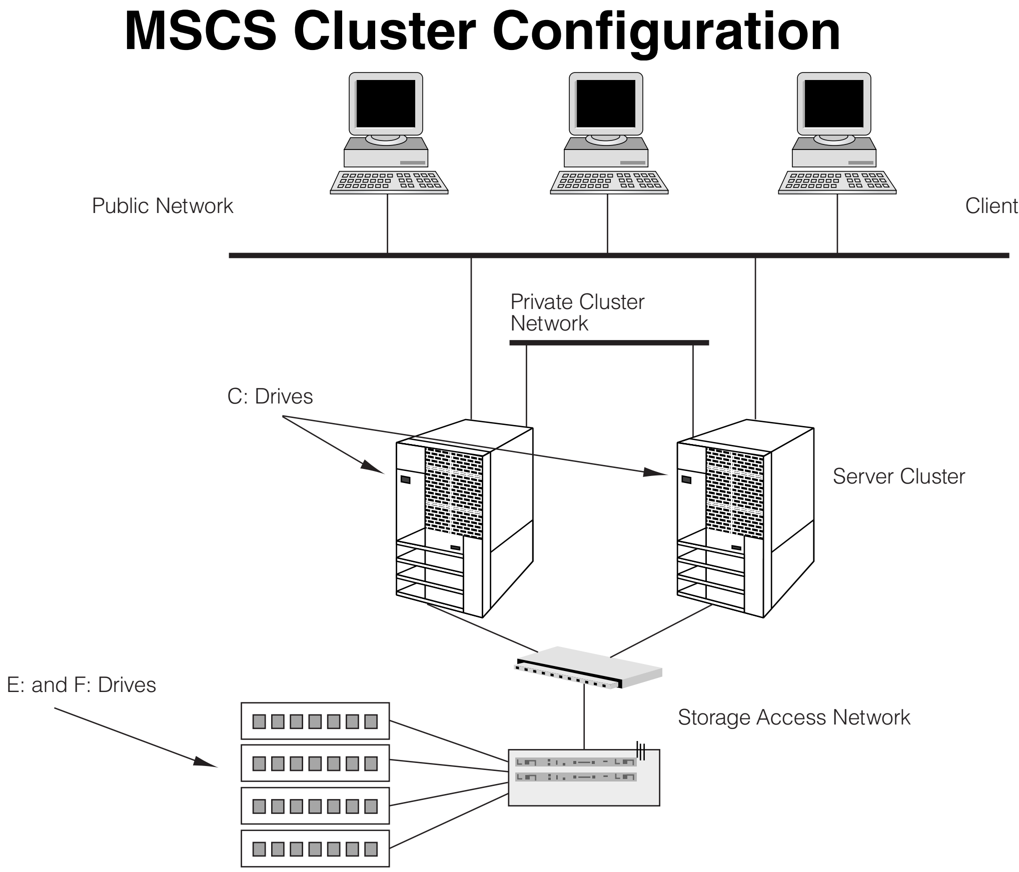
- When a server becomes unavailable while a user has a database open.
- Note: The user can open the database again, which causes failover to a different replica, if one exists in the cluster.
- When a user chooses File→ Database→ Properties or File→ Database→ Open.
- When the router attempts to deliver mail while MailClusterFail Over is set to 0.
- When the template server is unavailable while creating a new database.
- When running agents, other than the mail predelivery agent.
- When replicating with a server that is restricted by the administrator or has reached the maximum number of users or the maximum usage level set by the administrator. Also, when replicating with a database marked "Out of Service." Replication occurs regardless of such restrictions, so there is no need for failover to occur.

Clustering Example



Where are the mail files?

Many Domino administrators can find it a challenge to locate the backup mail files in a cluster environment. Administrators can go to the CLDBDIR.NSF database to find a list of the databases that are in the cluster, but this can be time-consuming. To make this task easier, a naming convention can be used for the mail subdirectories. In the Cluster Replica Chart, we have shown an example of a Domino Cluster with 3000 registered users and four servers. Dividing up the registered mail users equally among all four servers in the cluster leaves 750 registered users per server. When making the replica copies of the mail files, we put a copy of one-third of the registered users from Server 1 on the remaining servers in the cluster. By doing this for each server in the cluster, we have made it possible for any given server to absorb only one-third more workload in the event that a server is brought down for service.



To make it easier for a Domino administrator to know where a user's mail file and replica copy are, we have assigned a naming convention for our mail files. The first one-third of the registered users on Server 1 were placed in a mail file called '\Mail1_Backup2'. This means that the user’s primary mail file on Server 1 and the replica copy are on Server 2 in a directory called '\Backup2_Mail1'. Since the user's primary server and mail subdirectory are listed in the user's person document in Name and Address Book, the administrator can quickly find both copies of a user's mail file. This will assist the administrator in removing the mail files of an employee who has retired, for example.

Microsoft Cluster Service (MSCS)

A specific OS-level clustering option for the Intel platform is MSCS. Since Netfinity Availability Extensions for MSCS (NAE) provides enhancements to MSCS function, a brief review of MSCS is appropriate as background/review prior to exploring NAE.

MSCS provides clustering capability for two servers to provide hardware redundancy. If the OS fails, all applications are restarted on the other server in the cluster. MSCS also provides some capability to restart a failed application on the same server or on the other server (failover) by ongoing monitoring of the application or service.

MSCS utilizes TCP/IP for communication. The two servers in the MSCS cluster usually communicate on an additional network connection, leaving the normal network to accommodate client sessions (two network adapter cards for each node are recommended). Static IP addresses are also recommended (as opposed to Cluster IP addresses from a Dynamic Host Configuration Protocol [DHCP]) to avoid the access problems that are likely to occur if a DHCP lease expires.

MSCS uses the shared-nothing disk topology. Disks are shared at a hardware level, but ownership of the disks is allocated to only one system at a time. A quorum disk is accessed from both nodes since it is used to store information about the cluster.

In MSCS, all the resources on the cluster, such as disks, data files, addresses and applications, are categorized into resource types (12 standard types are supplied by MSCS) and organized into groups. A group serves to identify a set of resources which can reside on one or the other node, but not on both at the same time. A group is the smallest unit that can fail over. Dependency specifications define how resources relate to each other and are used to control the order in which resources are brought on-line or taken off-line.

The MSCS resource monitor obtains state information from the resource DLLs (remember the 12 standard resource types? Each relates to a Dynamic Link Library) and passes this information to the cluster service for restart, failover or failback as specified.

Comparing Domino clustering and MSCS clustering

If you use MSCS, you can set up Domino to run with it even though they use different methods of clustering.

Domino uses "application clustering" to provide high availability, scalability and workload balancing.

Domino monitors the cluster and determines when failover and workload balancing should occur, based on parameters that you set. You also determine how many replicas of a database to create.

MSCS uses "operating system clustering" to provide high availability. MSCS depends on the operating system to monitor the cluster and determine when failover should occur. MSCS supports two servers (nodes) in a cluster; the nodes must share a common disk device. MSCS does not currently support workload balancing.

MSCS failover works differently from Domino failover. If an MSCS node fails when Domino is running on it, the other node takes over. It claims ownership of the disk where the Domino data files reside, uses the same IP address that the Domino server uses, and starts the Domino server. Because the Domino server continues to run on the same IP address with the same data files, users may not notice that failover has occurred.

Netfinity Availability Extensions for MSCS (NAE)

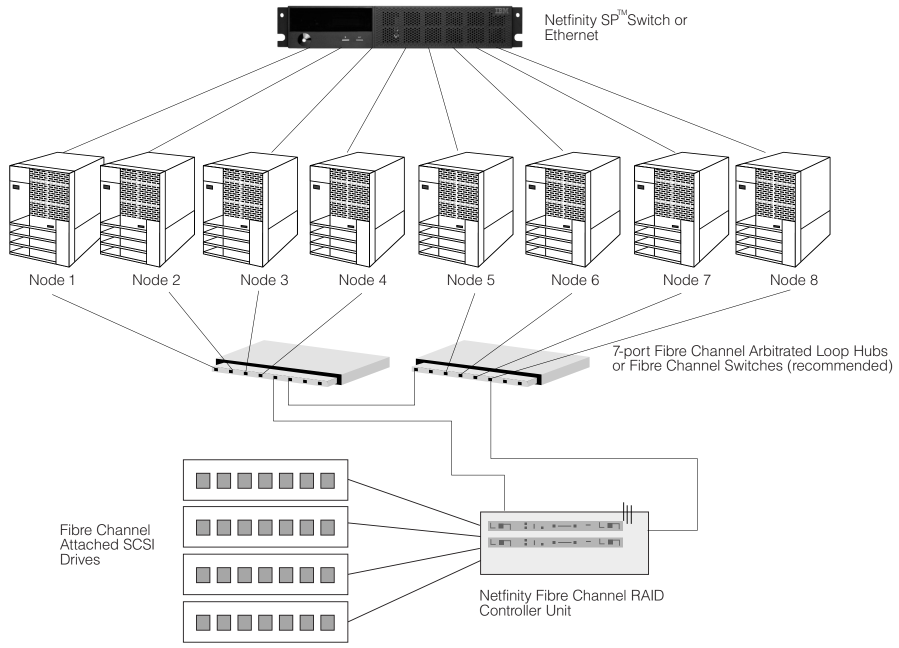
Netfinity Availability Extensions, developed by IBM under the code name *Cornhusker*, is a new clustering technology that greatly augments the capabilities of MSCS. NAE, part of IBM's X-architecture initiative, significantly extends the capabilities of MSCS to a maximum eight-node solution for the IBM Netfinity servers. NAE has roots in High Availability Cluster Multiprocessing (HACMP), the proven high-availability solution for the AIX platform.

NAE manages a collection of MSCS clusters, replacing MSCS function for managing cluster registration, membership, control and failover.

NAE implements a cascading failover policy. That is, in the event of a node failure, a group (for example, an application) on failed node x will be restarted on node n+1 in the best owner list for this group (with a wraparound when the end of the list is reached).

Shattering the 2-node MSCS barrier

IBM Netfinity Availablility Extensions for MSCS


Netfinity SP™ Switch or Ethernet
Node 1
Node 2
Node 3
Node 4
Node 5
Node 6
Node 7
Node 8
7-port Fibre Channel Arbitrated Loop Hubs or Fibre Channel Switches (recommended)
Fibre Channel Attached SCSI Drives
Netfinity Fibre Channel RAID Controller Unit

Expand MSCS up to eight server nodes (*IBM exclusive*)
- Resource failover among all eight nodes
- Support even-numbered and odd-numbered nodes

N+1, N-way, and primary/backup failover
- Better availability and scalability
- Easy configuration and setup
- Reduced hardware cost in cluster environment

Fully MSCS compatible
- Looks and feels like larger MSCS
- No modifications required to MSCS-compliant applications

Cluster Management: Single point of control

Includes:
- Software
- Lower cluster administration costs
- Control up to eight nodes from one screen
- IBM Planning and Installation Services

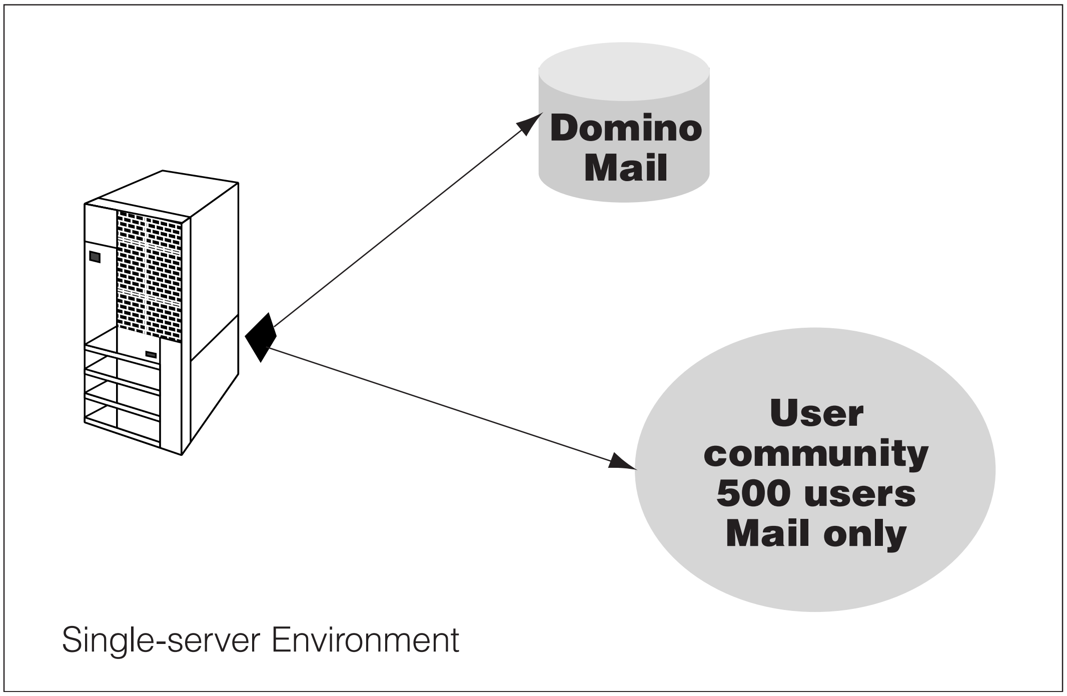
NAE uses a feature called disk pooling. Disk pooling is the ability for all nodes to connect to and share one or more Netfinity Fibre Channel RAID storage subsystems.
NAE reduces hardware and administration costs over MSCS through server consolidation. This is done by reducing the ratio of running servers to standby servers from a 1-to-1 ratio to as great as a 7-to-1 ratio. And by administration of all eight nodes from a single point, instead of administering each node pair separately.

Summary of Features

	Domino	MSCS	NAE
Nodes Supported	up to 6	2	3 to 8
Special Hardware Required	No	Yes	Yes
OS Support	NT, HP UX, AIX, Solaris, OS/400R, OS/390R, Linux, OS/2R	NT only	NT only
Distance Between Nodes	Unlimited	Limit of transmission distance between nodes and shared drives up to 10Km	Limit of transmission distance between nodes and shared drives up to 10Km
Disk Subsystem Fault Tolerance	No shared drive requirement	Must share drive subsystem	Must share drive subsystem
Data Redundancy	No shared drive requirements: thus all data in cluster is replicated on each node	Shared drive requirement allows for a single instance of the data set	Shared drive requirement allows for a single instance of the data set
Dynamic Load Balancing	Yes	No	No
Client Failover	Notes, Web Client	Notes, Web Client	Notes, Web Client
Client Failback	No	Yes	Yes
Server Task Failover Support	No	Yes	Yes
Network Failure Support	Yes	No	No
Single Point of Cluster Management	Yes (6 nodes)	Yes (2 nodes only)	Yes (8 nodes)
N+1 Failover	No	No	Yes

Example of an Evolution to Clustering

Phase I



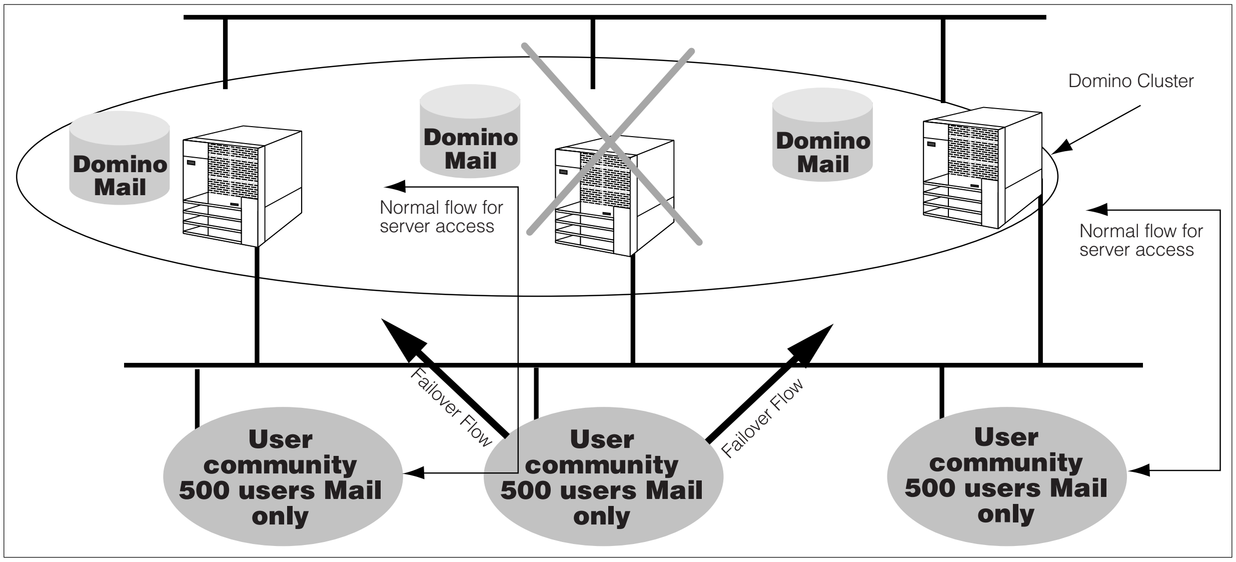
Environment

- Single location
- 500 mail users

Domino Configuration

- Single server
- RAID protection for only drive high availability

Phase II


Domino Cluster
Domino Mail
Domino Mail
Domino Mail
Normal flow for server access
Normal flow for server access
Failover Flow
Failover Flow
User community 500 users Mail only
User community 500 users Mail only
User community 500 users Mail only

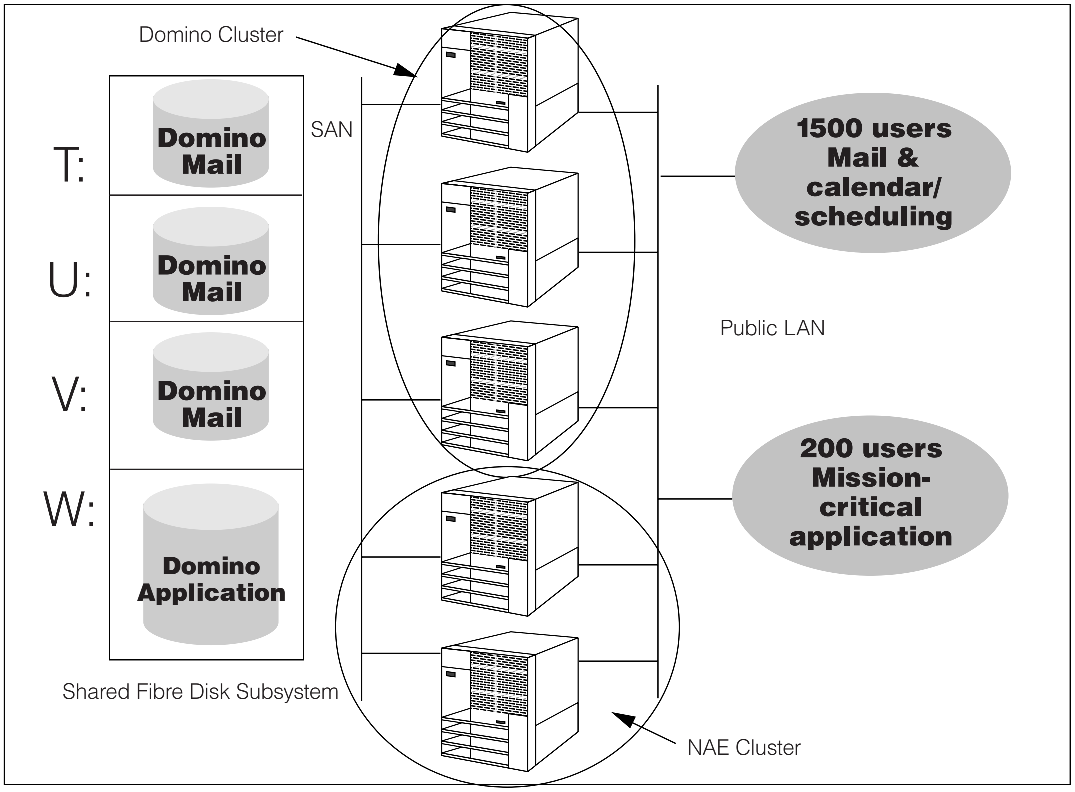
Environment

- Single location
- 1500 Mail-only users (Mail now critical to business)

Domino Configuration

- Three-server configuration
- Users split logically across three servers with failover
- High-speed backbone for cluster communication
- Domino clustering
  - Load balancing and high availability
- RAID protection for drive high availability

Phase III


Domino Cluster
T:
U:
V:
W:
Domino Mail
Domino Mail
Domino Mail
Domino Application
SAN
Shared Fibre Disk Subsystem
1500 users Mail & calendar/ scheduling
Public LAN
200 users Mission-critical application
NAE Cluster

Environment

- Single location
- 1500 Mail calendaring/scheduling users (higher sustained load levels)
- 200 users of mission-critical application
- Dedicated IS staff

Phase III continued

Domino Configuration

- Five-server configuration
  - 3 mail servers
    - Users split logically across three servers with failover
  - 1 Server for Domino mission-critical application
  - 1 Server for hot spare
- Domino clustering
  - Load balancing of e-mail
- Netfinity Availability Extensions for MSCS (NAE)
  - Highly available
    - Mission-critical application
    - Server agents
    - Fax and pager gateways
    - Heavily loaded mail servers
  - Storage Area Network (SAN)
    - Data storage centralization
      - Data Security
      - Manageability
    - Centralized tape backup
  - Server consolidation

Acknowledgments:

Special thanks to Mike Kistler for his insight in clustering and to Roger Hellman for his NAE advice.

Additional Resources:

For additional information, refer to the following sources:

IBM PC Solutions Technical Resource Library — (White Papers: “Lotus Domino Clusters Overview” and “Lotus Domino Clusters Primer”): pc.**ibm.com**/techconnect/tech/resource.html

IBM Redbooks — (IBM Netfinity Cluster Planning Guide SG24-5845-00): redbooks.**ibm.com**/redbooks

Lotus Domino Home Page: lotus.com/domino

Lotus Home Page: lotus.com/

Notes.net Home Page — notes.net/

NotesBench Consortium: notesbench.org/

Netfinity servers: **ibm.com**/netfinity

IBM Personal Computer Company
Department LO6A
3039 Cornwallis Road
Research Triangle Park NC 27709

Printed in the United States of America 05-00

For terms and conditions or copies of IBM's limited warranty, call 1 800 772-2227 in the U.S. Limited warranty includes International Warranty Service in those countries where this product is sold by IBM or IBM Business Partners (registration required).

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. IBM reserves the right to change specifications or other product information without notice.

IBM Netfinity systems are assembled in the U.S., Great Britain, Japan, Australia and Brazil and are comprised of U.S. and non-U.S. Components.

Year 2000 reminder: Visit **ibm.com**/pc/year2000 or call 1 800 426-3395 (and request document number 10020 from our faxback database) for the latest information.

THIS PUBLICATION MAY INCLUDE TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND, INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. SOME JURISDICTIONS DO NOT ALLOW DISCLAIMERS OF IMPLIED WARRANTIES SO THIS DISCLAIMER MAY NOT APPLY TO YOU.