A Survey of Popular Clustering Technologies

Edward Whalen, Performance Tuning Corporation

May 2002

INTRODUCTION

There are a number of clustering products available on the market today, and clustering has become quite popular. However, the term clustering has also become quite popular and there are a number of different products that use the term "clustering" that act and perform differently. In this paper I will attempt to explain the different technologies and terminologies that are used to form clusters.

What is a cluster?

Simply put, a cluster is two or more computer systems that act together to perform a single function. I like to think of a cluster as a black box. When you connect your application into the cluster you see what appears to be one system or database. Whether it is actually two or more systems or not is irrelevant. It appears to the end user to be one system.

Clustering technologies have taken two different approaches in order to serve two different goals. There are *failover clusters* that are designed to provide a high availability or quick recoverability solution and there are *performance clusters* that are designed to provide for more performance. Of course there are also clustering technologies that provide for both of these goals.

Microsoft Clustering

Microsoft has two different clustering products. The first Microsoft Cluster Services (MSCS) is a failover solution that works on NT and supports a number of applications. The second product is the *Federated Server Cluster* and is a product that only works with Microsoft SQL Server.

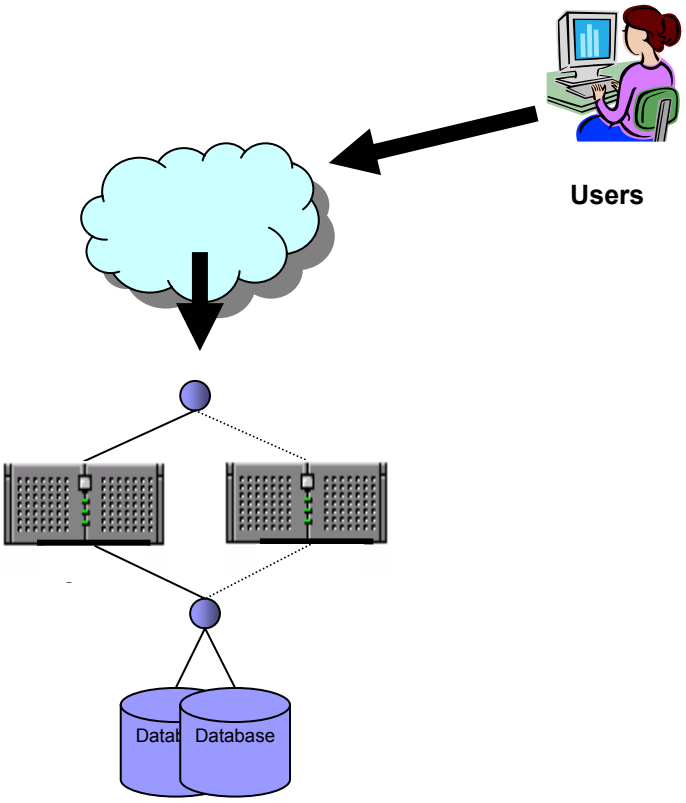
MSCS

Microsoft Cluster Services is a product that is designed for Microsoft Windows NT and Windows 2000 that allows a standby system to take over for the primary system in the event of a failure. There are two major components that make up the MSCS cluster; the shared disk subsystem and the cluster interconnect.

With MSCS each of the systems in the cluster must be able to access the same disk subsystem. Thus it is called a shared disk subsystem. Although all of the systems (or nodes) in the cluster must be able to access the shared disk, only one system has ownership of the shared disk subsystem at any one time. The other node is unable to access the disk subsystem. In the event of a failure, the ownership of the shared disk

subsystem is passed to the standby node and that node will take the disk subsystem and restart applications that were running on the cluster.
The cluster interconnect is a network connection that is used to pass cluster information between the nodes in the cluster. The nodes are constantly passing status information, known as the *heartbeat* back and forth between the nodes. It is when the heartbeat is no longer detected from the primary node that the standby node takes over.

The user that is accessing the application or database on the cluster uses an IP address that is assigned to the cluster, not an individual node. In the event of a failure that causes the standby node to take over the users continue to use the cluster IP address, however now the standby node will respond to that address. An illustration is shown here:



A MSCS Cluster

The MSCS cluster is very popular and works well, however it does not provide any additional performance to the system, since only one node is active at a time.

Active/Passive vs. Active/Active Clusters

With MSCS clusters you can have both Active/Passive Clusters or Active/Active Clusters. An Active/Passive cluster is the cluster that is shown above. One node is active at any one time, the other node is passive, waiting for a failure. Another option that is available is an Active/Active Cluster. An Active/Active Cluster should be thought of as two Active/Passive clusters, each acting as the failover node for the other. Although you can have an Active/Active cluster with both nodes running a database software such as

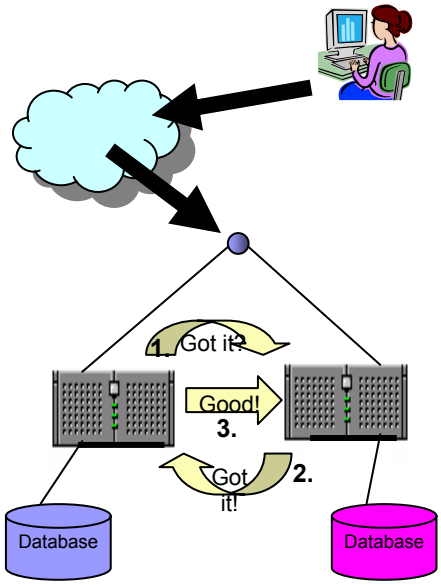
MS SQL Server or Oracle, they cannot be the same database, since each virtual cluster (each Active node) has its own separate shared disk that must be able to fail over independently of the other.

Active/Active Clusters are not as popular as the Active/Passive Cluster, since the Active/Active cluster must be configured such that in the event of a failure it is possible to run two SQL Servers, or two Oracle instances on the same node (both active nodes on the same system). This usually entails tuning down each node so that in the event of a failure one system can handle both loads.

Note: In the event of a failure, the standby node will take over and will proceed to recover the database, just as if it had failed on a single node system. This recovery process could take seconds, minutes or hours depending on the activity on the system prior to failure. The MSCS solution is designed to get the system back up and running as soon as possible, but the time to recover is not guaranteed. Thus, MSCS does not provide continuous uptime.

Federated Servers

The Microsoft SQL Server Federated Server is a distributed "shared nothing" cluster. Each machine in the cluster has a separate database and the data is actually spread across multiple machines. Each machine has a part of the data. The knowledge about which machine has the data is pushed out to the cluster application and on the database servers themselves.



A Federated Server Cluster

Microsoft Federated Servers are very difficult to configure and maintain and is not a very popular solution for clustering. Whereas the Federated Server can provide more performance by distributing the database among multiple systems, there is no failover

capabilities or high availability built into this solution. However, you can create a Federated Server Cluster that is made up of multiple MSCS clusters.

Other Microsoft Solutions

There are a few other high availability/fault tolerant type solutions available within Microsoft SQL Server, however these are not really clustering solutions. Many SQL Server customers use SQL Server replications to create near real time copies of their data in order to offload certain functions, such as reporting queries, etc. This is very popular and works very well, however it is not a "black box", and thus not a cluster.

Many SQL Server customers also use *Log Shipping* in order to create a standby system. With Log Shipping the transaction log backups from SQL Server, which contain the changes made to the database, are shipped to a standby system. This standby system, which was created by restoring a backup of the initial system is constantly being restored with the transaction log backups. In this manner, the standby system is constantly being kept up to date with recent changes to the database.

Log shipping is popular for two reasons. First there is no performance degradation on the primary system. Second, log shipping can be done via tape and can be sent to a system in another town, state or country. Although it is not kept up to date as often, it still provides a good standby solution on a budget. In addition, since it is not automatic, a mistake by a user which destroys or corrupts data in the database is not immediately reflected on the standby system and thus can be avoided.

Oracle Clustering

As with Microsoft, Oracle also provides a number of clustering solutions. On Microsoft Windows NT and Windows 2000 Oracle supports MSCS with a product called *Oracle Fail Safe*. In addition, Oracle has another cluster product called *Oracle Real Application Cluster* (RAC) which is an updated version of *Oracle Parallel Server* (OPS). These clustering products server both fault tolerant and performance goals.

Oracle Fail Safe

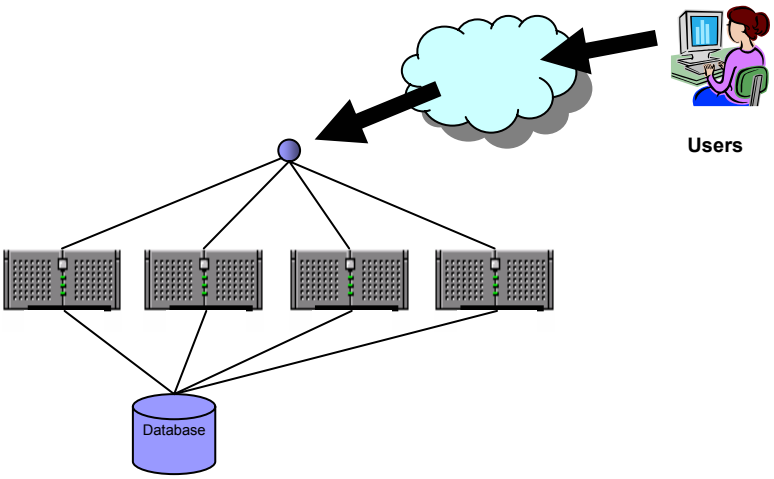
Oracle Fail Safe uses MSCS and works in the same manner. There is a shared disk subsystem and a cluster interconnect that is used to allow for active and standby nodes in the cluster. As with MSCS, Oracle Fail Safe can be configured in an Active/Passive or Active/Active configuration.

Note: In the event of a failure, the standby node will take over and will proceed to recover the database, just as if it had failed on a single node system. This recovery process could take seconds, minutes or hours depending on the activity on the system prior to failure. The Oracle Fail Safe solution is designed to get the system back up and running as soon as possible, but the time to recover is not guaranteed. Thus, Oracle Fail Safe does not provide continuous uptime.

Oracle Real Application Cluster

The Oracle RAC product is different from MSCS, but has some of the same components. In a RAC system two or more cluster nodes share a disk subsystem, however, unlike MSCS, all of the nodes in the cluster access the shared disk subsystem at the same time. The Oracle cluster management software is responsible to make sure that the data is kept in sync controls locking of data between the nodes.

With the Oracle RAC cluster each node in the cluster runs an Oracle instance that is sharing the same database. In this way both fault tolerance and performance is gained. Since all nodes are accessing the same database it is transparent to the user which node they are actually running on, since all the data can be accessed from any node.



Oracle RAC Cluster

The Oracle RAC Cluster is available only with Oracle9i and is just starting to gain in popularity.

Other Oracle Solutions

As with MS SQL Server Oracle also supports replication. However, replication on Oracle is not nearly as popular as replication on Microsoft SQL Server. Oracle also supports *Standby Database*. With Standby Database the Oracle Redo Log backups (archive log files) are shipped to a backup system and restored to the standby database. In the event of a failure, the standby database can take over.

Standby Database is a great solution for two reasons. First there is no performance degradation on the primary system. Second, standby database can be done via tape and can be sent to a system in another town, state or country. Although it is not kept up to date as often, it still provides a good standby solution on a budget. In addition, since it is not automatic, a mistake by a user which destroys or corrupts data in the database is not immediately reflected on the standby system and thus can be avoided. With Oracle9i the standby database functionality is now automated.

Third Party Clustering

There are a number of products available on the market that link into the operating system at a low level and provide mirroring of the OS as well as the disk subsystem. These products can be found on the internet and will not be covered here. Of particular interest to some customers is the ability to mirror disk storage to a remote standby site. This is gaining in popularity and serves a need in the industry and information is available from Storage Area Network vendors.

Summary

This paper was designed to provide a high level overview of the different types of clustering solutions available today from Microsoft and Oracle. Other products are available and their exclusion should not be a reflection of their quality. They are simply not the focus of this paper.

As you have seen, cluster takes on two distinct functions, high availability and performance. Your needs will determine your solution.

About the Author

Edward Whalen is vice president and principle consultant at Performance Tuning Corporation (www.perftuning.com). Performance Tuning Corporation provides database performance tuning, load testing and troubleshooting services on Oracle and MS SQL Server. Edward Whalen was a co-author on for SQL Server books from Microsoft Press;

SQL Server 7 Administrator's Companion,
SQL Server 7 Performance Tuning Technical Reference,
SQL Server 2000 Administrator's Companion
SQL Server 2000 Performance Tuning Technical Reference

Edward Whalen has also authored four Oracle books;

Oracle Performance Tuning and Optimization (Oracle7)
Teach Yourself Oracle8 in 21 Days
Oracle Performance Tuning (Oracle8/9i)
Teach Yourself Oracle9i in 21 Days (2002/2003)

Edward Whalen is considered a leader in database performance tuning.

