

Clustering

A computer cluster is a group of computers that work together in unison so that from an outside perspective they form one single computer doing one task. Computer clusters usually contain at least two separate computers, but often can contain dozens, if not hundreds of separate machines. Connecting the computers together into a cluster is an interesting task ~ it can be done in several ways.

The most common way is via fast or gigabit Ethernet over switched networks, but there are other ways of connecting computers such as optical fibre, or "Infinband" switching style fabrics. Regardless of the connectivity medium the end result is the same, and additionally the speed of the cluster may not be affected directly by the speed of its switching fabric.

Clusters are usually deployed to improve performance and/or availability over that provided by a single computer.

Clusters usually fit into the following categories:

1. High Availability (HA) Clusters
2. Load Balancing (LB) Clusters
3. Grid Computing

In the Internet Service Provider arena, the first two categories are the most frequently used, and in reality, Grid Computing is resigned to very high computational tasks and problems such as economic forecasting, seismic analysis, and mathematical problems. It is very rare that an ISP would invoke such expensive hardware on a hosting platform.

High Availability (HA) Cluster

High Availability cluster, which are also known as failover clusters are provisioned primarily for the purpose of increasing the availability of services which would be provided by one or more machines. A High Availability cluster operates by having multiple redundant nodes (nodes are computers attached to the cluster) which are used to provide service in case of failure of one or more nodes.

The most common size for a High Availability cluster is two nodes which is the minimum amount of nodes required to operate in High Availability.

High Availability can be achieved by employing one of two methods, either DNS Round Robin or provisioning a dedicated hardware load balancer. In each case load is split between the nodes who could either be running static content or feeding information back to a back-end database for example.

There are many commercial High Availability software packages available, both on linux/unix or on Microsoft Windows platforms.

Load Balancing (LB) Cluster

Load Balancing clusters are connected physically in the same way as a High Availability cluster, however Load Balancing clusters operate by distributing the received workload evenly between cluster nodes.

Because of the inherent design (and requirement) of a Load Balancing cluster, the typical cluster would be configured with multiple redundant load balancing front ends, with possibly two or more back end nodes running databases.

With load balanced nodes it goes without saying that an identical copy of the content, whether it be static or dynamically generated would need to be placed on the server to provide a seamless experience for the end user.

With both High Availability and Load Balancing clusters, it is possible to distribute the load geographically (for redundancy rather than speed) this is known as Global Load Balancing. This network topology requires a significant amount of resource spent on equipment and may involve DNS changes and web servers in globally distributed locations ~ dependent on the target market.