System & Service Management
Clustering

Computercluster
Ein Rechnerverbund (engl. Cluster) mit dem Ziel der Erhöhung der Rechenkapazität oder der Verfügbarkeit gegenüber einem einzelnen Computer.
Zu den verschiedenen Arten von Cluster Systemen gehören folgende:

- High-Availability-Cluster: Werden zur Steigerung der Verfügbarkeit bzw. für bessere Ausfallsicherheit eingesetzt. Tritt auf einem Knoten des Clusters ein Fehler auf, werden die auf diesem Cluster laufenden Dienste auf einen anderen Knoten migriert.
- Load-Balancing Cluster: Werden zum Zweck der Lastverteilung auf mehrere Maschinen aufgebaut. Die Lastverteilung erfolgt in der Regel über eine redundant ausgelegte, zentrale Instanz.
- High-Performance-Computing-Cluster: Dienen zur Abarbeitung von Rechenaufgaben. Diese Rechenaufgaben werden auf mehrere Knoten aufgeteilt.

http://de.wikipedia.org/wiki/Computercluster
http://en.wikipedia.org/wiki/Two-node_cluster
http://en.wikipedia.org/wiki/High-availability_cluster
http://en.wikipedia.org/wiki/Load_balancing_(computing)
http://de.wikipedia.org/wiki/Server_Load_Balancing
http://de.wikipedia.org/wiki/Hochleistungsrechnen
http://www.tecchannel.de/server/hardware/458076/it_systeme_ausfallsicherheit_im_kostenvergleich/index5.html

Aktiv/Aktiv Cluster
Ein Aktiv/Aktiv-Cluster ist ein Rechnerverbund, in dem mehrere Rechner (Clusternodes) gleichzeitig aktiv sind. Bei Aktiv/Aktiv-Clustern wird zwischen den Architekturen Shared Nothing und Shared All unterschieden.

Unter einer Aktiv/Aktiv-Konfiguration versteht man in diesem Zusammenhang, dass die so gesicherte Ressource, also zum Beispiel eine Datenbank, auf allen Clusterknoten aktiv ist. Wenn ein Knoten ausfällt, übernehmen die übrigen Knoten die Prozesse des ausgefallenen Knotens, es gibt praktisch keinerlei Ausfallzeiten, eventuell jedoch starke Einbussen in der Performance, da die gleiche Last nun von weniger Systemen übernommen werden muss.

Klarer Vorteil dieses Konzeptes ist, dass die Ressourcen nicht redundant vorhanden sein müssen und der Ausfall eines Knotens nur leistungsabhängige Auswirkungen auf die Verfügbarkeit des Clustersystems als Ganzes hat.

Hauptnachteil eines solchen Setups ist, dass die Ressource, anders als bei einer Aktiv/Passiv-Konfiguration, entweder direkt eine Aktiv/Aktiv-Konfiguration unterstützen oder mit teils hohem Aufwand an eine solche angepasst werden muss. Bei einer Aktiv/Passiv-Konfiguration hingegen können nahezu beliebige Ressourcen verfügbar gemacht werden.

http://de.wikipedia.org/wiki/Aktiv/Aktiv-Cluster

Aktiv/Passiv Cluster
Ein Failover-Cluster oder Aktiv/Passiv-Cluster ist ein Verbund von mindestens zwei Computern, in dem bei einem Ausfall eines Rechners ein zweiter Rechner dessen Aufgaben übernimmt.

http://de.wikipedia.org/wiki/Aktiv/Passiv-Cluster

Shared Nothing Architektur
Die Shared-Nothing-Architektur (SN) beschreibt eine Distributed Computing-Architektur, bei der jeder Knoten unabhängig und eigenständig seine Aufgaben mit seinem eigenen Prozessor und den zugeordneten Speicherkomponenten wie Festplatte und Hauptspeicher erfüllen kann und kein bestimmter, einzelner Knoten für die Verbindung zu einer Datenbank notwendig ist. Die Knoten sind über ein LAN- oder WAN-Netzwerk miteinander verbunden.

Shared Nothing ist auf Grund seiner Skalierbarkeit beliebt für Webanwendungen oder parallele Datenbanksysteme. Wie bei Google gezeigt werden konnte, ist ein Shared Nothing System nahezu unbegrenzt durch Ergänzung zusätzlicher Knoten in Form preiswerter Computer ausbaufähig, weil kein einzelnes Netzwerkelement existiert, dessen begrenzte Leistung die Geschwindigkeit des gesamten Systems vermindert.

http://de.wikipedia.org/wiki/Shared_Nothing

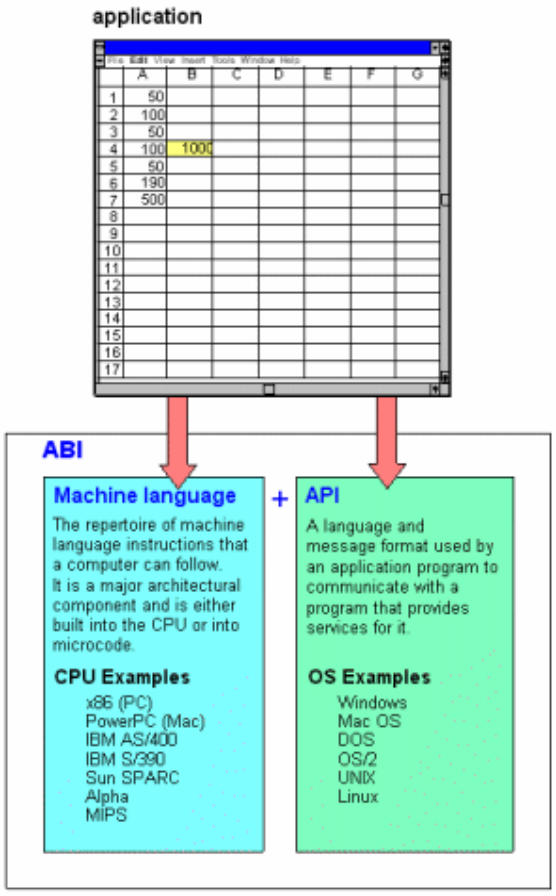
Application Binary Interface
An application binary interface (ABI) describes the low-level interface between an application (or any type of) program and the operating system (or another application).

ABIs cover details such as data type, size, and alignment; the calling convention, which controls how functions' arguments are passed and return values retrieved; the system call numbers and how an application should make system calls to the operating system; and in the case of a complete operating system ABI, the binary format of object files, program libraries and so on. A complete ABI, such as the Intel Binary Compatibility Standard (iBCS), allows a program from one operating system supporting that ABI to run without modifications on any other such system, provided that necessary shared libraries are present, and similar prerequisites are fulfilled.

http://en.wikipedia.org/wiki/Application_binary_interface

A specification for a specific hardware platform combined with the operating system. It is one step beyond the application program interface (API), which defines the calls from the application to the operating system. The ABI defines the API plus the machine language for a particular CPU family. An API does not ensure runtime compatibility, but an ABI does, because it defines the machine language, or runtime, format.

http://www.answers.com/topic/application-binary-interface#ixzz1AFA6WPOA



Eintrittsinvarianz
Eine Routine bzw. Methode wird als eintrittsinvariant (engl. reentrant) oder auch wiedereintrittsfähig bezeichnet, wenn sie so implementiert ist, dass sie von mehreren Prozessen gleichzeitig ausgeführt werden kann. Dabei dürfen sich die gleichzeitig ausgeführten Instanzen nicht in die Quere kommen. Die Ausführung jeder Instanz läuft also gleich ab, egal wie viele andere Instanzen es noch von dieser Methode gibt.

Eintrittsinvariante Programmkonstrukte sind die Basis für viele Multitasking-Systeme. Auch Kernfunktionen von Betriebssystemen müssen reentrant sein.

http://de.wikipedia.org/wiki/Eintrittsinvarianz

Arbitration
Der Arbiter oder die Arbitrationslogik (lat. arbiter, Richter, zu lat. arbitor, beobachten, meinen) ist eine Funktionseinheit in Form einer elektrischen, digitalen Schaltung oder einer Softwareroutine, die Zugriffskonflikte oder Zugriffskollisionen löst oder priorisiert. Dies ist zum Beispiel notwendig im Falle von Bussystemen mit mehreren Busmastern – also Einheiten, die aktiv wie etwa ein DMA-Controller auf den Datenbus zugreifen dürfen – um zu entscheiden, welcher Busmaster Zugriff bekommt.

Im Allgemeinen versteht man unter der Arbitration die möglichst gerechte Zuteilung von Ressourcen auf verschiedene Benutzer (Geräte). Dieses Verfahren kommt auch beim so genannten Token-Verfahren bzw. in der FDDI-Technologie oder bei Token Ring zur Anwendung. Das CSMA/CD-Verfahren ist hingegen ein Beispiel für ein Verfahren, das keine „gerechte“ Zuteilung der Ressource garantiert.

http://de.wikipedia.org/wiki/Arbiter

Quorum
Unter einem Quorum oder einer Voting Disk versteht man eine Komponente des Cluster Managers eines Computerclusters zur Wahrung der Datenintegrität im Fall eines Teilausfalls. Bei Ausfall des Cluster Interconnects (der Verbindung zwischen den Clusterknoten), besteht das Risiko einer Aufspaltung des Gesamtsystems in unerwünschterweise autonom agierende Einheiten, die fast immer die Datenintegrität bedroht (Split-Brain-Problem). Durch wechselweises oder konkurrierendes Schreiben in die logische Struktur der Voting Disk wird im Falle eines unterbrochenen Interconnects entschieden, welcher Teil des Clusters überleben soll. Die Voting Disk liegt auf Shared Storage.

Bekanntes Problem: Die Quorum Disk muss jederzeit verfügbar sein. Bei einem Ausfall würde der ganze Cluster stillstehen. Damit entsteht ein Single Point of Failure. Es gibt viele Möglichkeiten, das Cluster Quorum zu spiegeln. Die Königslösung für sämtliche damit verbundenen Probleme stellt die speicherseitige Replikation (storage-based mirroring) im SAN dar. (Storage-based mirroring ist jedoch sehr teuer!)

http://de.wikipedia.org/wiki/Quorum_(Informatik)

I/O Fencing
There will be some situations where the leftover write operations from failed database instances reach the storage system after the recovery process starts, such as when the cluster function failed on the nodes, but the nodes are still running at OS level. Since these write operations are no longer in the proper serial order, they can damage the consistency of the stored data. Therefore, when a cluster node fails, the failed node needs to be fenced off from all the shared disk devices or disk groups. This methodology is called I/O Fencing, sometimes called Disk Fencing or failure fencing.

The main function of the I/O fencing includes preventing updates by failed instances, and detecting failure and preventing split brain in cluster. Cluster Volume Manager, in association with the shared storage unit, and Cluster File System play a significant role in preventing the failed nodes accessing shared devices.

For example, in Sun Cluster, disk fencing is done through SCSI-2 reservation for dual hosted SCSI devices and for multi-hosted environment through SCSI-3 PR. Veritas Advance Cluster uses the SCSI-3 persistent reservation to perform I/O fencing. In the case of Linux clusters, CFS like Polyserve and Sistina GFS are able to perform I/O fencing by using different methods like fabric fencing that uses SAN access control mechanism.

http://www.dba-oracle.com/real_application_clusters_rac_grid/io_fencing.html

Split Brain Problematik
Split Brain ist in der Informatik ein unerwünschter Zustand eines Computerclusters, bei dem alle Zwischenverbindungen zwischen den Clusterteilen gleichzeitig unterbrochen sind.

Zur Koordination der Transaktionen im Cluster wird in der Regel ein Cluster Interconnect oder ein Quorum verwendet – je nach eingesetzter Technologie. Wird die Verbindung zwischen einem oder mehreren Teilen des Clusters über diesen Weg unterbrochen, kann keines noch unterscheiden ob es sich um einen partiellen Ausfall oder eine Trennung handelt. Alle dieser (nun isolierten) Clusterfragmente arbeiten für sich weiter, um die Bereitstellung des Dienstes aufrechtzuerhalten.

Das Grundproblem von Split Brain ist der Umstand, dass mindestens zwei Teile noch funktionieren, jedoch keine Koordination zwischen ihnen mehr möglich ist. Während dies bei reinen Lesezugriffen noch nicht unmittelbar problematisch scheint, führt ein Schreibzugriff zu massiven Konflikten.

Die Grundlage aller Gegenmaßnahmen ist der gleichzeitige Einsatz von Quorum und Cluster Interconnect. Desweiteren können mehrere Quoren und parallelisierte Interconnects eingesetzt werden.

Im Zusammenspiel zwischen Quoren und Interconnect ist eine zuverlässig automatisierte Entscheidungsfindung notwendig.

http://en.wikipedia.org/wiki/Split_brain
http://de.wikipedia.org/wiki/Split_Brain_(Informatik)

Amnesia Problematik
Amnesia occurs if all the nodes leave the cluster in staggered groups. An example is a two-node cluster with nodes A and B. If node A goes down, the configuration data in the CCR is updated on node B only, and not node A. If node B goes down at a later time, and if node A is rebooted, node A will be running with old contents of the CCR. This state is called amnesia and might lead to running a cluster with stale configuration information.

Oracle Artikel zu Split Brain und Amnesia Problematik: http://docs.sun.com/app/docs/doc/821-0518/concepts-42?l=en&a=view

Weitere Informationen
http://de.wikipedia.org/wiki/Gewichtetes_Votieren
Interessanter Artikel zum Thema MS SQL Mirroring: http://www.dbresource.de/Default.aspx?tabid=199