APLICACIÓN DE CLUSTERS (servicios)

Los clusters son agrupaciones de computadores [nodos] (junto con las infraestructuras de comunicación y almacenamiento asociadas) destinados a trabajar de forma conjunta para ofrecer un determinado servicio y/o realizar operaciones de cómputo específicas.

Dependiendo de los objetivos finales de un cluster, tenemos:

1 Clusters de balanceo de carga: conjunto de nodos que se reparten la prestación de un servicio determinado (servir aplicaciones web, dar soporte a un gestor de BD, etc)

Punto clave: ”repartición” del trabajo

Objetivo: atender el máximo número de peticiones al servicio

2 Clusters de alta disponibilidad: conjunto de nodos que garantiza la disponibilidad de un servicio determinado aún en el caso de fallos y/o caídas de algún elemento.

Punto clave: tolerancia y recuperación ante fallos (failover)

Objetivo: garantizar la prestación (y consistencia) del servicio

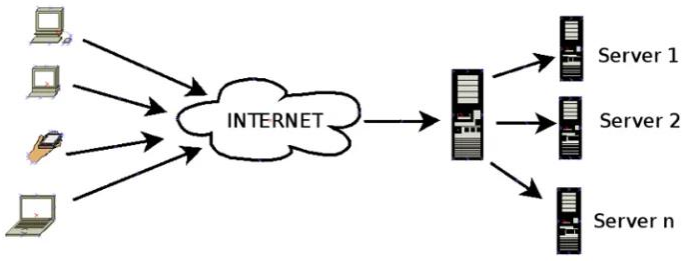
3 Clusters de alto rendimiento: conjunto de nodos que trabajan conjuntamente en tareas de cálculo intensivo (renderizado de gráficos, análisis de datos, predicción) [high performance clusters]

Punto clave: procesamiento distribuido

Objetivo: maximizar el rendimiento y la capacidad de cálculo.

CLUSTER DE BALANCE DE CARGA

Balanceadores de carga: dispositivos hardware y/o software conectados a un conjunto de nodos de procesamiento entre los que reparte las peticiones recibidas por parte de los clientes. Como ejemplo podemos nombrar una granja de servidores web.



Posible solución a problemas de escalabilidad:

- Escalabilidad vertical (scale up): ”mejorar” las máquinas/nodos que prestan un servicio añadiendo más recursos (memoria, capacidad CPU, etc)
- Escalabilidad horizontal (scale out): agregar más máquinas/nodos para prestar el servicio (uso de balanceadores de carga).

Conceptos relacionados:

- Servidor virtual servidores/servicios que se ofrecen a los clientes desde el balanceador (también llamado nodo director)
- Servidor real servidores/servicios que realmente atienden/procesan las peticiones (nodos del cluster)

Alternativas de Implementación:

Balanceo de carga por DNS [balanceo implícito]

Se vincula un servicio a un nombre de dominio DNS. En el servidor DNS se le vinculan distintas direcciones IP (servidores ”reales”) a ese nombre de dominio. Se configura servidor DNS para que reordene la lista de direcciones devuelta cada vez que los clientes resuelvan el nombre de dominio del servicio que se ofrece (balanceo implícito) [DNS Round Robin]

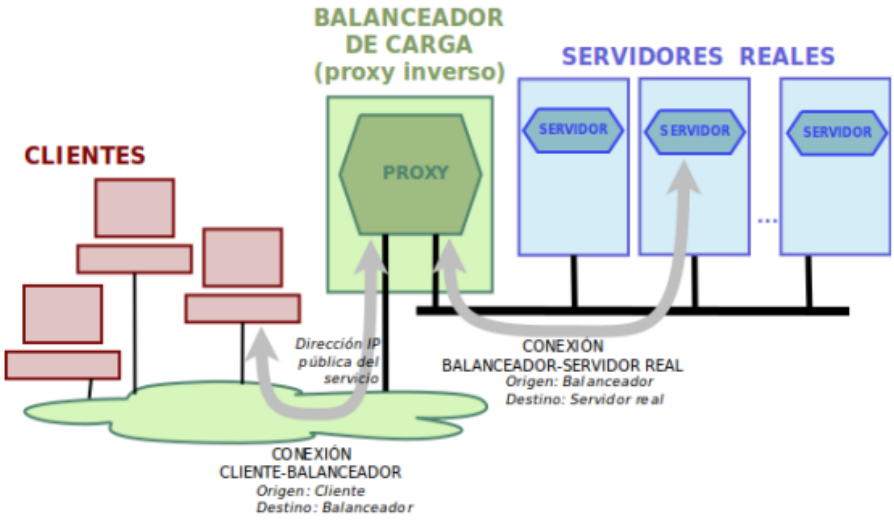
La ventaja es que no requiere contar explícitamente con equipo/s dedicado al balanceo y como Inconveniente se puede se atribuye que el balanceo no equitativo, distorsiones debidas a las caches de nombres.

Balanceo a nivel de aplicación

Mediante proxies ”inversos” Retransmiten a nivel de aplicación peticiones/respuestas a/desde servidores reales. Opcionalmente pueden alterar el contenido de las peticiones/respuestas.

Ejemplos:

- HAProxy (http://www.haproxy.org/)
- mod_proxy y mod_proxy_balancer de Apache



CLUSTER DE ALTA DISPONIBILIDAD

Un cluster de alta disponibilidad (HA: high availability) es un conjunto de dos o más nodos que se garantiza que ante el fallo en uno de ellos no se detendrá el servicio que ofrecen en conjunto. La idea es contar con nodos redundantes que asuman el servicio cuando algún componente falla. El objetivo es garantizar tolerancia a fallos sin provocar inconsistencias de datos.

Tipos de interrupciones/paradas:

- Paradas previstas: mantenimiento, actualización, reparación, ...
- Paradas imprevistas: desastres naturales, fallos hardware ó software, ...

Cluster HA es capaz (sin intervención humana) de:

- detectar los fallos (hardware o software)
- mantener el servicio (retomándolo/iniciándolo en otro nodo)
- garantizar la integridad de los datos

Los Clusters HA suelen usarse para dar soporte a servicios/aplicaciones críticas para una organización que no pueden verse interrumpidas donde se tiene un alto coste de downtime (tiempo que se encuentra parado un servicio).

Aplicable en contextos donde SÍ se requieren disponibilidad y tolerancia a fallos, como ser bases de datos críticas, aplicaciones web de comercio electrónico, sistemas de ficheros compartidos, servidores de correo.

Configuraciones típicas

- Redundancia hardware: replicación de componentes [procesamiento, almacenamiento, comunicaciones]
- Redundancia software: replicación componentes software, ejecución simultánea [replicación] de procesos, logs de sincronización
- Redundancia de datos: réplicas/copias de seguridad, sincronización

Configuración Activo-Pasivo

- Los servicios/aplicaciones se ejecutan sobre un conjunto de nodos activos (al menos uno)
- Otro conjunto de nodos pasivos (al menos uno) actúan como respaldo de los servicios ofrecidos (redundancia)
- Nodos pasivos sólo entran en funcionamiento ante fallo de los nodos activos

Configuración Activo-Activo

- Todos los nodos actúan como servidores activos de los servicios/aplicaciones
- Cualquier nodo puede servir como respaldo ante fallos en los demás nodos

Nota: Se obtiene un mayor aprovechamiento de los recursos cuando es posible combinarlos con componentes de balanceo de carga para repartir la carga entre los nodos.

CLUSTER DE ALTO RENDIMIENTO

Utilizado en aplicaciones con requisitos de cálculo intensivos: Simulación científica, renderización de gráficos, modelos de predicción meteorológica, minería de datos (Big Data).

Están conformados por un conjunto de nodos (habitualmente con un S.O. específico para HPC) junto con una infraestructura de comunicación de alta velocidad, Los nodos del cluster colaboran de forma coordinada en la ejecución de un determinado proceso/procesos concreto con alta demanda de cálculo computacional Habitualmente emplean librerías específicas de programación paralela/distribuida.