Escalando servicios

Servicios de cómputo

- Se requiere que sean escalables y eficientes de acuerdo a la demanda.
- Altos niveles de servicios de cómputo se asocian a tres fases:
  - Cómputo de alto rendimiento (High Performance Computing, HPC).
  - Servicios de alto volumen (High Volume Services, HVS).
  - Servicios de alta disponibilidad (High Availability Services (HAS).

Cómputo de alto rendimiento

- Describe una variedad de problemas que requieren **procesamiento intensivo**.
- Supercomputadoras o *clusters* de computadoras en red, que toman un problema computacional grande y lo dividen en partes pequeñas para resolverlas en paralelo.
- “Computer cluster: A computer cluster is a set of loosely or tightly connected computers that work together so that, in many respects, they can be viewed as a single system.” (Computer cluster. (s.f.). En Wikipedia. Recuperado 16 de febrero de 2018 de https://en.wikipedia.org/wiki/Computer_cluster).
- Requiere de de alta capacidad de cómputo y conexiones de baja latencia.
- Grid computing: is the collection of computer resources from multiple locations to reach a common goal. The grid can be thought of as a distributed system with non-interactive workloads that involve a large number of files. Grid computing is distinguished from conventional high-performance computing systems such as cluster computing in that grid computers have each node set to perform a different task/application. (Computer cluster. (s.f.). En Wikipedia. Recuperado 16 de febrero de 2018 de https://en.wikipedia.org/wiki/Grid_computing).

Servicios de alto volumen

- Se refiere a aplicaciones o servicios que proveen contenidos diseñados para manejar un número elevado de transacciones.
- Se logra implementando estrategias para compartir cargas de trabajo entre host servidores (llamado balanceo de carga).
- Se centra en un alto volumen de peticiones y por lo tanto debe considerar la **escalabilidad**.
- Se requiere un análisis cuidadoso de los cuellos de botella.
- Debe tomarse en cuenta el ajuste del rendimiento de los servicios.
- La capacidad de entregar niveles de servicio confiables depende tanto de los recursos disponibles en el centro de datos como del patrón de demanda impulsado por los usuarios.

Servicios de alta disponibilidad

- Son aplicaciones diseñadas para situaciones de misión crítica, donde es esencial que los servicios estén disponibles para los clientes con una latencia muy baja o un tiempo de respuesta muy corto.
- Se requiere redundancia y detección de fallas en el sistema.
- Es necesario que un servicio esté disponible con una calidad asegurada en el tiempo de respuesta con base en el servicio objetivo.
- Debe decidirse que significa “disponible” en términos de una métrica objetivo.

Recursos

- Administración de recursos.
- Utiliza y comparte entre diferentes tareas, usuarios y procesos:
  - CPU
  - Disco
  - Memoria
  - Capacidad de la red

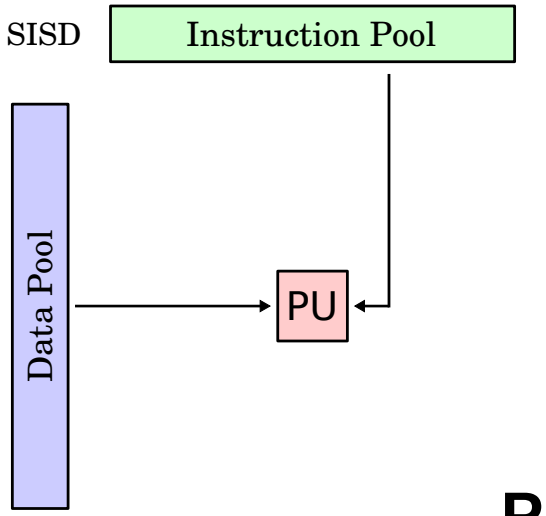
CPU

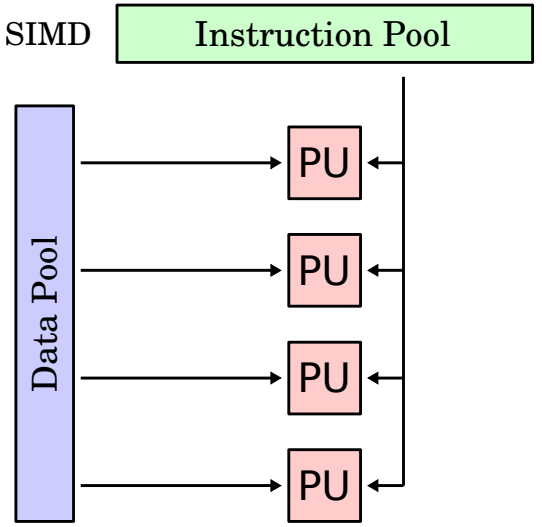
- Para trabajos intensivos en el CPU las herramientas básicas para compartir son:
  - Multitareas e hilos a nivel sistema operativo.
  - Computadoras multiprocesador, vector de máquinas y *cluster beowulf* los cuales permiten paralelismo.
  - Cómputo en mainframe.
  - Sistema de distribución de trabajo con un “Grid Engine” para cómputo en paralelo.
  - Herramientas de programación MPI (*Message Passing Interface*)

Taxonomía de Flynn (1)

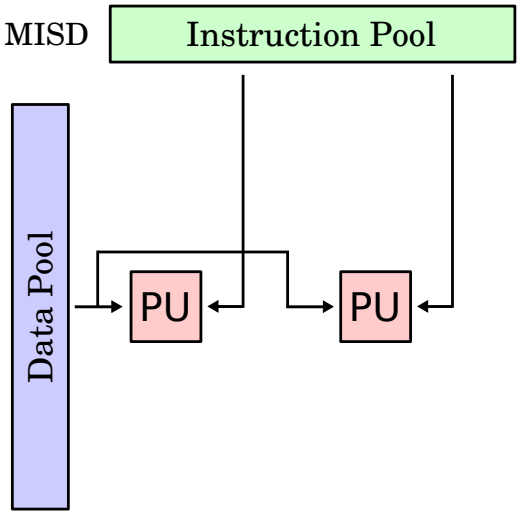
- Es una clasificación de arquitectura de computadoras propuesta por Michael J. Flynn en 1966.
- Single Instruction Single Data (SISD)
- Single Instruction Multiple Data (SIMD)
- Multiple Instruction Single Data (MISD)
- Multiple Instruction Multiple Data (MIMD)
- Single Instruction Multiple Threads (SIMT)

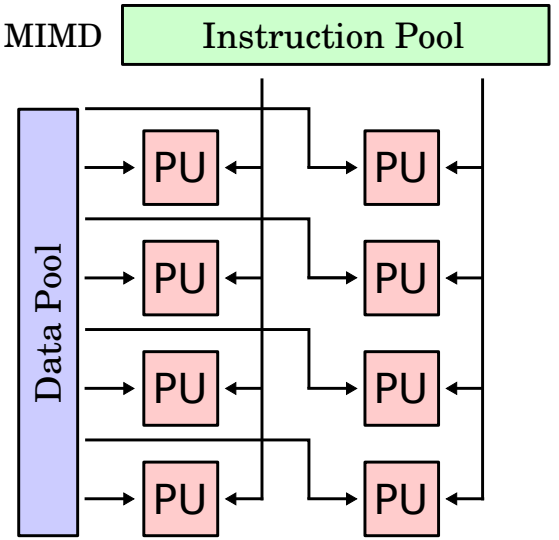
Taxonomía de Flynn (2)





PU=Processing Unit





Cluster de balanceo de carga

- Toda la carga de trabajo proviene de unos o mas *front ends* de balanceo de carga.
- Incluyen características de alta disponibilidad.
- También se le conoce como “**server farm**”.

Disco

- El rápido almacenamiento de altos volúmenes de datos es un componente clave.
- El almacenamiento en disco requiere de movimientos mecánicos de brazos y cabezas de disco para lectura y escritura en disco.
- Servicios de disco que tienen un cache en RAM proveen operaciones rápidas de lectura/escritura.
- Redudant Array of Independent (Inexpensive) Disk (RAID) incluye estrategias para mejorar el rendimiento y confiabilidad.
- RAID son arreglos de disco que tienen controladoras especiales diseñadas para optimizar velocidad y confiabilidad.
- Incluye tolerancia a fallos y paralelismo.
- La tolerancia a fallos incluye características para reducir los riesgos de fallas de disco y pérdida de datos sin pérdida de tiempo. Para esto usa codificación redundante de datos y *mirroring.*

Disco (2)

- La velocidad de la red es mayor a la velocidad de los buses de datos internos, lo que ayuda a tener mejor rendimiento y gestión en servicios de acceso a disco en red.
  - *Storage Area Networks* (SAN)
    - Red independiente de dispositivos de almacenamiento.
    - Utiliza *ISCSI* (SCSI sobre Internet) utilizando una interfaz de red dedicada conectada a un arreglo de discos.
  - *Network Attached Storage* (NAS)
    - Trabaja como un sistema de red tradicional.
    - Corre un protocolo de red tradicional como IP.

Red

- Existen diversas tecnologías para la red.
- El rendimiento de la red puede ser mejorado adaptando o mejorando:
  - Tecnología de transporte (Fibra, Ethernet).
  - Políticas de ruteo.
  - Calidad de servicio.
- La diferentes tecnologías ofrecen optimización, rendimiento y costo.

Data Center

- ¿Cuántas computadoras necesita el Centro de Datos?
- Costo
- Rendimiento
- Aprendizaje de prueba y error
- ¿Qué hardware elegir?
  - Comprar muchas computadoras baratas y utilizar nivel alto de carga compartida para mejorar el rendimiento?
  - ¿Comprar computadoras caras diseñadas para nivel bajo de redundancia y rendimiento?

Servidor blade

Es una computadora para los centros de proceso de datos diseñado para aprovechar el espacio, reducir el consumo y simplificar su explotación.

Bare Metal Server

- Metal desnudo o expuesto.
- Servidor dedicado.
- Los recursos están disponibles única y exclusivamente para un cliente.
- También se les conoce como “Single Tenant Server”.

Cuellos de botella

- ~~¿Dónde está el cuello de botella en la computadora?~~
- ¿Dónde está el cuello de botella en una aplicación?
- Los recursos necesitados por las aplicaciones son diferentes.
- Decidir si un proceso pasa la mayor parte de su tiempo utilizando:
  - ¿CPU?
  - ¿Red?
  - ¿Disco?
  - ¿Memoria?
- Entender qué dependencias están presentes en el sistema, las cuales pueden ser un factor limitante.

Solo cuando conocemos como una aplicación interactúa con el hardware y el sistema operativo podemos mejorar el rendimiento del sistema.

Cuellos de botella (2)

- Distancia vs velocidad
- Escritura vs lectura
- Redundancia
- Aplicaciones que mantienen un estado interno vs sesiones a largo plazo vs transacciones de una sola vez.

Diseño de la arquitectura de la aplicación

- Arquitectura de software
- En un mundo ideal los ingenieros de software tendrían cuidado del centro de datos cuando escriben las aplicaciones.
- El diseño de servicios involucra aspectos desde protocolos de bajo nivel de la red hasta experiencias web de alto nivel sobre Internet.
- Mientras que el diseño de bajo nivel del procesador ayuda a una mejor integración, el diseño dominante en arquitectura de sistemas de software es la separación de intereses. El diseño para servicios de aplicación de red es una arquitectura de tres capas: servidor web, aplicación y base de datos.
- Un diseño basado completamente en el código del programa es un enfoque sin sentido en los ambientes de red modernos.

Arquitectura de software

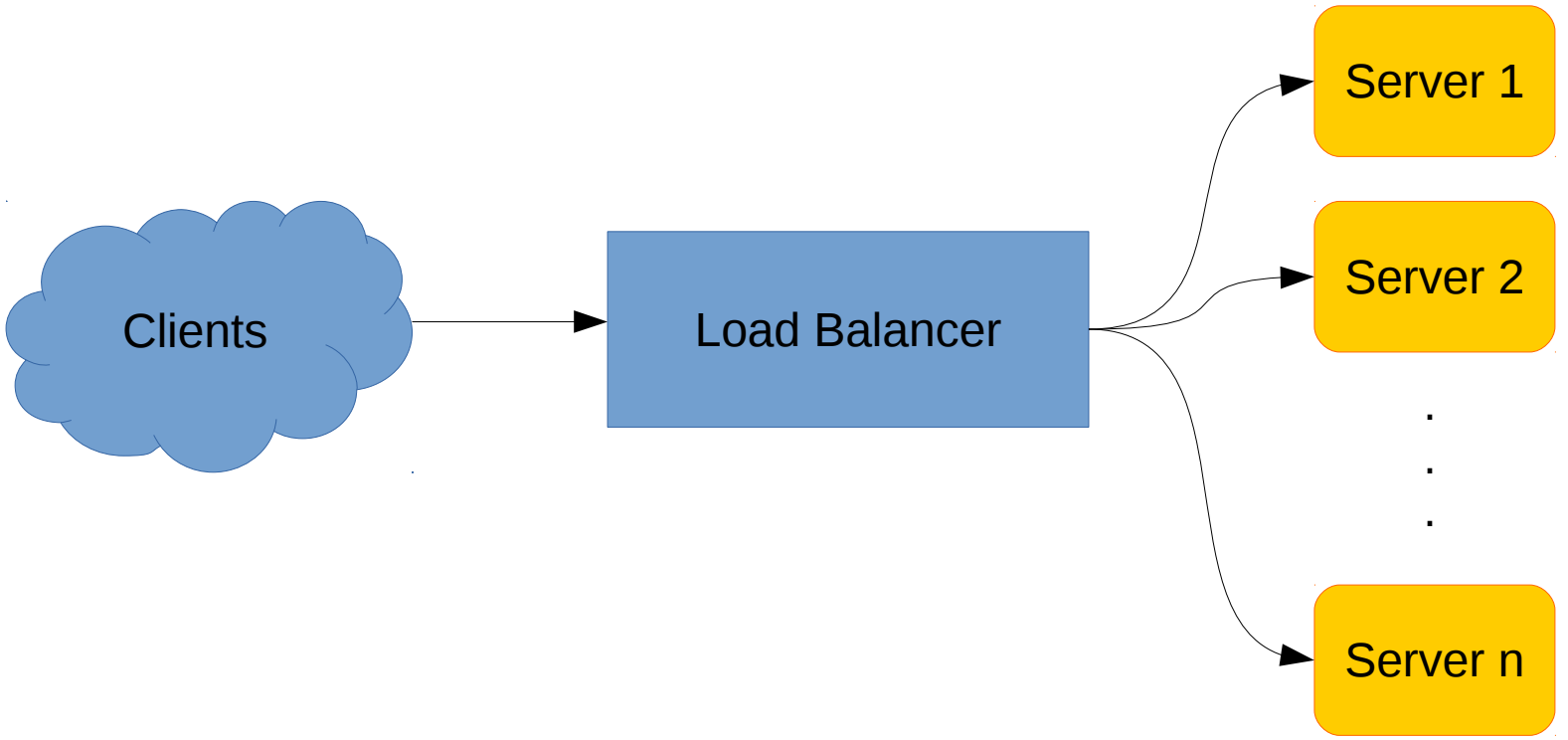
- Se debe tomar en cuenta desde la experiencia del usuario hasta los recursos básicos que que cumplen la promesa de servicio.
- Las consideración en el diseño para servicios incluyen:
  - Exactitud
  - Rendimiento y escalabilidad
  - Tiempo de respuesta (latencia)
  - Confiabilidad
  - Tolerancia a fallas
  - Recuperación de desastres.
- Servicios de alto volumen requieren uso eficiente de recursos
- Alta disponibilidad requiere tanto estrategias de eficiencia y confiabilidad integradas completamente con la estrategia de desarrollo de software.

Escalabilidad

- “Es la propiedad deseable de un sistema, red o proceso, que indica su habilidad para reaccionar y adaptarse sin perder calidad, o bien manejar el crecimiento continuo de trabajo de manera fluida, o bien para estar preparado para hacerse mas grande sin perder calidad en los servicios.” ((Escalabilidad. (s.f.). En Wikipedia. Recuperado 16 de febrero de 2018 de https://es.wikipedia.org/wiki/Escalabilidad).
- Consiste en caracterizar la salida de un sistema como una función de entrada (una tarea).
- La habilidad de un sistema para completar tareas depende de la medida en que la tarea pueda ser dividida en flujos independientes que puedan ser completados en paralelo al igual que la topología del sistema y sus canales de comunicación.
- Podemos pensar en la escalabilidad como la habilidad de un sistema para tratar con grandes cantidades de entrada, considerando como incrementando la entrada afecta la eficiencia con la cual las tareas realizan la salida.

- Pensemos en lluvia que cae (como un proceso aleatorio de peticiones), nos interesa saber la cantidad de lluvia que cae y si podemos drenarla lo suficientemente rápido introduciendo un número suficiente de drenajes (procesadores o servidores). Esta escalabilidad la podemos ver como el rendimiento en función de la carga.
- Algunas veces la entrada es una función de un número de clientes y otras veces la salida es una función del número de servidores.

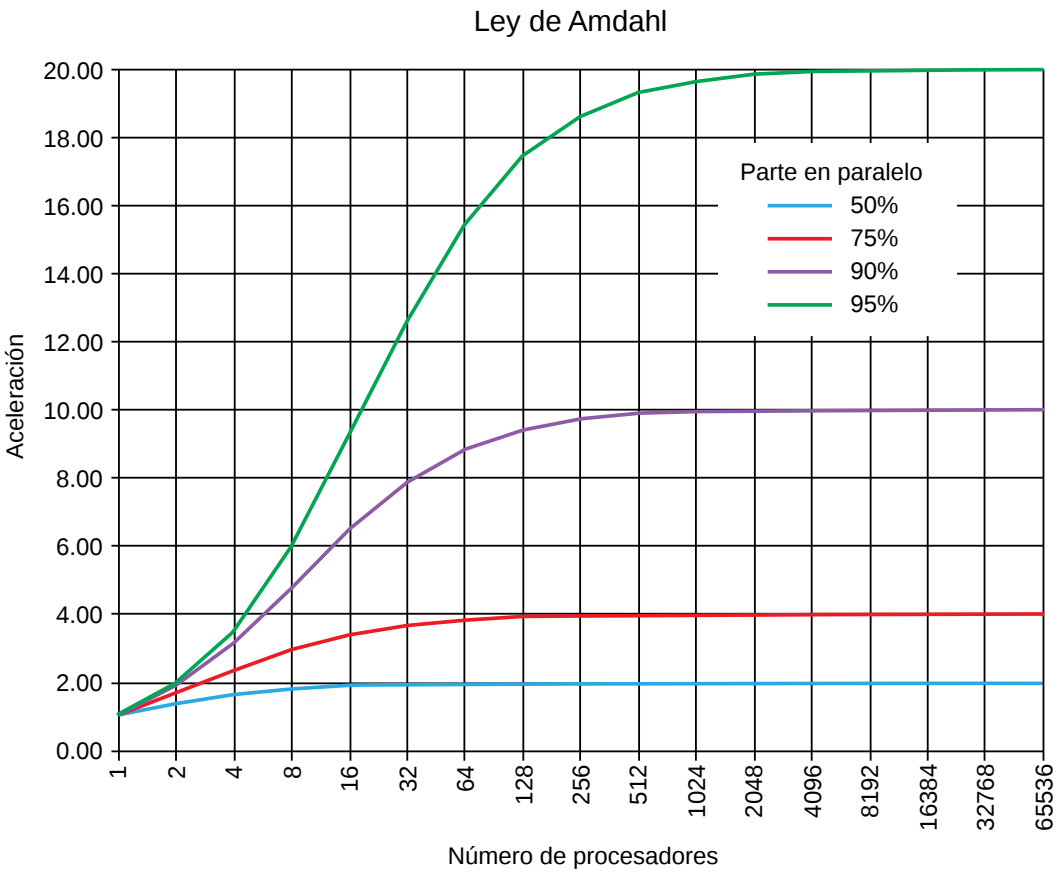
Topología para compartir carga



Ley de Amdahl

- Formulada por Gene Amdahl.
- Se utiliza para conocer la mejora máxima de un sistema de información cuando solo una parte de éste es mejorado.
- Establece que la mejora obtenida en el rendimiento de un sistema debido a la alteración de uno de sus componentes está limitada por la fracción de tiempo que se utiliza dicho componente.
- Calcula la aceleración esperada o incremento fraccional en el rendimiento, como resultado de paralelizar parte del proceso (balance de carga entre N procesadores de la siguiente forma:

$$Tm=Ta.((1-Fm) + Fm/Am)$$


Ley de Amdahl
20.00
18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
Aceleración
Parte en paralelo
50%
75%
90%
95%
1
2
4
8
16
32
64
128
256
512
1024
2048
4096
8192
16384
32768
65536
Número de procesadores

HPC

- Los sistemas de High Performance Computing utilizan el concepto de paralelismo.
- Hardware HPC
  - Multiprocesadores simétricos (SMP, Symmetric multiprocessors).
  - Procesadores de vectores
  - Clusters

SMP

- Multiprocesadores simétricos es un tipo de arquitectura HPC en la cual múltiples procesadores comparte la memoria.

Procesadores de vectores

- La CPU está optimizada para funcionar bien con matrices o vectores.
- Los sistemas de procesadores vectoriales ofrecen un alto rendimiento y fueron la arquitectura HPC dominante en los años ochenta y principios de los noventa, pero los clusters se han vuelto mucho más populares en los últimos años.

Clusters

- Un clúster es un conjunto de procesadores masivamente paralelos (MPP, massively parallel processors).
- Un procesador en un cluster se conoce comúnmente como un nodo y tiene su propia CPU, memoria, sistema operativo y subsistema de E/S y es capaz de comunicarse con otros nodos.

Cluster

- Fail-over cluster
  - En su forma mas simple tiene dos nodos: uno activo y el otro en espera, pero que monitorea al primero. Si el primero falla entra en funcionamiento el segundo.
- Load-balancing cluster
  - Por lo general se utilizan para servidores Web muy ocupados, diversos nodos hospedan el mismo sitio, cada nueva petición se enruta dinámicamente a un nodo con menos carga.
- High-performance cluster
  - Su utilizan para correr programas en paralelo que requieren un uso intensivo de cómputo y que son de interés especial para la comunidad científica.

Grid computing

- Permite utilizar de forma coordinada recursos heterogéneos (cómputo, almacenamiento y aplicaciones) que no están sujetos a un control centralizado.
- Es una forma de computación distribuida en la cual los nodos participantes pueden ser de iguales o distintas arquitectura y cubrir la gama de potencia de cómputo, desde embebidos hasta supercomputadoras.