Published on *Marco de Desarrollo de la Junta de Andalucía* (http://madeja.i-administracion.junta-andalucia.es/servicios/madeja)

Conceptos sobre la escalabilidad

- **Área:** Arquitectura Tecnológica
- **Carácter del recurso:** Recomendado

Código: RECU-0220
Tipo de recurso: Referencia

Descripción

Se entiende por escalabilidad a la capacidad de adaptación y respuesta de un sistema con respecto al rendimiento del mismo a medida que aumentan de forma significativa el número de usuarios del mismo. Aunque parezca un concepto claro, la escalabilidad de un sistema es un aspecto complejo e importante del diseño.

La escalabilidad esta íntimamente ligada al diseño del sistema. Influye en el rendimiento de forma significativa. Si una aplicación esta bien diseñada, la escalabilidad no constituye un problema. Analizando la escalabilidad, se deduce de la implementación y del diseño general del sistema. No es atributo del sistema configurable.

La escalabilidad supone un factor crítico en el crecimiento de un sistema. Si un sistema tiene como objetivo crecer en el numero de usuarios manteniendo su rendimiento actual, tiene que evaluar dos posibles opciones:

- Con un hardware de mayor potencia o
- Con una mejor combinación de hardware y software.

Se pueden distinguir dos tipos de escalabilidad, vertical y horizontal:

- El escalar verticalmente o escalar hacia arriba, significa el añadir más recursos a un solo nodo en particular dentro de un sistema, tal como el añadir memoria o un disco duro más rápido a una computadora.
- La escalabilidad horizontal, significa agregar más nodos a un sistema, tal como añadir una computadora nueva a un programa de aplicación para espejo.

Escalabilidad Vertical

El escalar hacia arriba un sistema viene a significar una migración de todo el sistema a un nuevo hardware que es mas potente y eficaz que el actual. Una vez se ha configurado el sistema futuro, se realizan una serie de validaciones y copias de seguridad y se pone en funcionamiento. Las aplicaciones que estén funcionando bajo la arquitectura hardware antigua no sufren con la migración, el impacto en el código es mínimo.

Este modelo de escalabilidad tiene un aspecto negativo. Al aumentar la potencia en base a ampliaciones de hardware, llegara un momento que existirá algún tipo de limitación hardware. Además a medida que se invierte en hardware de muy altas prestaciones, los costos se disparan tanto de forma temporal (ya que si se ha llegado al umbral máximo , hay componentes hardware que tardan mucho tiempo en ampliar su potencia de forma significativa) como económicos. Sin embargo a nivel estructural no supone ninguna modificación reseñable, lo que la convierte en una buena opción si los costos anteriores son asumibles.

Escalabilidad Horizontal

La escalabilidad horizontal consiste en potenciar el rendimiento del sistema desde un aspecto de mejora global, a diferencia de aumentar la potencia de una única parte del mismo. Este tipo de escalabilidad se basa en la modularidad de su funcionalidad. Por ello suele estar conformado por una agrupación de equipos que dan soporte a la funcionalidad completa. Normalmente, en una escalabilidad horizontal se añaden equipos para dar mas potencia a la red de trabajo.

Con un entorno de este tipo, es lógico pensar que la potencia de procesamiento es directamente proporcional al número de equipos de la red. El total de la potencia de procesamiento es la suma de la velocidad física de cada equipo transferida por la partición de aplicaciones y datos extendida a través de los nodos.

Si se aplica un modelo de escalabilidad basado en la horizontalidad, no existen limitaciones de crecimiento a priori. Como principal e importante defecto, este modelo de escalabilidad supone una gran modificación en el diseño, lo que conlleva a una gran trabajo de diseño y reimplantación. Si la lógica se ha concebido para un único servidor, es probable que se tenga que estructurar el modelo arquitectónico para soportar este modelo de escalabilidad.

El encargado de como realizar el modelo de partición de datos en los diferentes equipos es el desarrollador. Existen dependencias en el acceso a la aplicación. Es conveniente, realizar una análisis de actividad de los usuarios para ir ajustando el

funcionamiento del sistema. Con este modelo de la escalabilidad, se dispone de un sistema al que se pueden agregar recursos de manera casi infinita y adaptable al crecimiento de cargas de trabajo y nuevos usuarios.

La escalabilidad cuenta como factor crítico el crecimiento de usuarios. Es mucho más sencillo diseñar un sistema con un número constante de usuarios (por muy alto que sea este) que diseñar un sistema con un número creciente y variable de usuarios. El crecimiento relativo de los números es mucho más importante que los números absolutos.

Balance de carga

A la hora de diseñar un sistema con compartición de recursos, es necesario considerar como balancear la carga de trabajo. Se entiende este concepto, como la técnica usada para dividir el trabajo a compartir entre varios procesos, ordenadores, u otros recursos. Esta muy relacionada con lo sistemas multiprocesales, que trabajan o pueden trabajar con mas de una unidad para llevar a cabo su funcionalidad . Para evitar los cuellos de botella, el balance de la carga de trabajo se reparte de forma equitativa a través de un algoritmo que estudia las peticiones del sistema y las redirecciona a la mejor opción

Balance de Carga por Hardware

Presenta las siguientes características:

- A partir de un algoritmo (Round Robin, LRU), examina las peticiones HTTP entrantes y selecciona el más apropiado entre los distintos clones del sistema.
- La selección del clon del sistema esta basada en el algoritmo de sustitución y es aleatoria.
- Esto último punto provoca problemas en el diseño, ya que no garantiza que si un usuario realiza varias peticiones sean atendidas por el mismo clon del sistema. Por lo tanto, no hay mantenimiento de la sesión del usuario en servidor y condiciona el diseño.
- La sesión debe de ser mantenida por el desarrollador.
- Al ser un proceso hardware, es muy rápido.

Balance de carga por Sotfware

- Examinan el paquete a nivel del protocolo HTTP para garantizar el mantenimiento de la sesión de usuario.
- Distintas peticiones del mismo usuario son servidas por el mismo clon del servidor.
- Más lentos que los balanceadores Hardware
- Normalmente son soluciones baratas.

Cluster sobre servidores

El concepto de clustering introduce la capacidad de unir varios servidores para que trabajen en un entorno en paralelo. Es decir, trabajar como si fuera un solo servidor el existente. En las etapas primigenias del clustering, los diseños presentaban graves problemas que se han ido subsanando con la evolución de este campo. Actualmente se pueden crear clusters en función de las necesidades

- Unión de Hardware
- Clusters de Software
- Alto rendimiento de bases de datos

En resumen, cluster es un grupo de múltiples ordenadores unidos mediante una red de alta velocidad, de tal forma que el conjunto es visto como un único equipo, más potente . Con ello se pretende mejorar los siguientes parámetros de la arquitectura:

- Alto rendimiento
- Alta disponibilidad
- Equilibrio de carga
- Escalabilidad

El clustering no presenta dependencias a nivel de hardware (no todos los equipos necesitan el mismo hardware) ni a nivel de software (no necesitan el mismo sistema operativo). Este tipo de sistemas dispones de una interfaz que permite dirigir el comportamiento de los clusters. Dicha interfaz es la encargada de la interactuación con usuarios y procesos, realizando la división de la carga entre los diversos servidores que compongan el cluster.

Tipos de Cluster

- **Alta Disponibilidad (HA) y Failover.** Enfocados a garantizar un servicio ininterrumpido, al duplicar toda la infraestructura e introducir sistemas de detección y re-enrutamiento (Servicios Heart-Beat), en caso de fallo. El propósito de este tipo de clusters es garantizar que si un nodo falla, los servicios y aplicaciones que estaban corriendo en ese nodo, sean trasladados de forma automática a un nodo que se encuentra en stand-by. Este tipo de cluster dispone de herramientas con capacidad para monitorizar los servidores o servicios caídos y automáticamente migrarlos a un nodo secundario para garantizar la disponibilidad del servicio. Los datos son replicados de forma periódica, o a ser posible en tiempo real, a los nodos en Stand-by.
- **Cluster Balanceado.** Este tipo de cluster es capaz de repartir el tráfico entrante entre múltiples servidores corriendo las mismas aplicaciones. Todos los nodos del cluster pueden aceptar y responder peticiones. Si un nodo falla, el tráfico se sigue repartiendo entre los nodos restantes.

Enlaces externos

- Pagina de IBM sobre la escalabilidad
- Pagina de Oracle sobre la escalabilidad y rendimiento

Pautas

Área: Arquitectura » Arquitectura Tecnológica

Código	Título	Tipo	Carácter
LIBP-0074	Buenas prácticas en el diseño de la escalabilidad	Directriz	Obligatoria