Tema 2

Alta disponibiliad y escalabilidad

Pedro A. Castillo Valdivieso
Depto Arquitectura y Tecnología de Computadores
Universidad de Granada
pacv@ugr.es

Índice

3

Introducción

disponibilidad **escalabilidad**

conceptos más importantes al diseñar una granja web

4

Introducción

Nuestros servidores deben dar el mejor servicio a todos los usuarios y deben estar todo el tiempo disponible (24/7).

- Disponibilidad
- Escalabilidad
- Balanceo de carga

Índice

Alta disponibilidad

http://www.isitdownrightnow.com/

Alta disponibilidad

Mala impresión al entrar en un sitio y está caído.

Esperamos que una web esté disponible siempre.

Disponibilidad: capacidad de aceptar visitas las 24h todos los días.

Alta disponibilidad

Cuando un sitio no está disponible se dice que se ha caído o sufre un problema de no-disponibilidad:

- Tiempo de no-disponibilidad programado.
- Tiempo de no-disponibilidad no programado.

Sólo debería haber ”tiempos de no-disponibilidad programados” (y lo más cortos posibles)

actualizaciones del SO, de aplicaciones o de hardware

Alta disponibilidad

Medir la disponibilidad dando un porcentaje.

Escala “punto nueve”:

```
100 – (tiempoCaido / periodoTiempo) * 100
```

Por ejemplo:

caída de 1h en un día -> 95.83333% de disponibilidad

caída de 1h en una semana -> 99.404% de disponibilidad

Lo ideal es tener un 100% de disponibilidad.

Alta disponibilidad

Un 100% de disponibilidad es no sufrir caídas no-programadas

Los sitios web se conforman con alcanzar un 99.9% ó 99.99%

Disponibilidad (%)	Periodo de un año
90%	36.5 días
95%	18.25 días
98%	7.3 días
99%	3.65 días
99.9%	8.76 horas
99.99%	52.56 minutos
99.999%	315 segundos
99.9999%	31.5 segundos

Alta disponibilidad

¿Cómo se consigue **mejorar la disponibilidad**?

El uso de **subsistemas redundantes** y monitorizarlos mejora la disponibilidad del sistema global.

Surgen conceptos derivados:

- disponibilidad de red
- disponibilidad de servidor
- disponibilidad de aplicación

Si la disponibilidad de red es baja, quizás haya que mejorar el ancho de banda, y no tenga sentido centrar esfuerzos en mejorar las aplicaciones.

Alta disponibilidad

Ejercicio:

Buscar frameworks y librerías para diferentes lenguajes que permitan hacer aplicaciones altamente disponibles con relativa facilidad.

Como ejemplo, examina PM2
https://github.com/Unitech/pm2
que sirve para administrar clústeres de NodeJS.

Índice

Escalabilidad

Cuando una persona sufre estrés, su capacidad para afrontar tareas se ve mermada.

Cuando un sistema experimenta estrés, su capacidad para dar servicio también se ve afectada.

15

Escalabilidad

Incremento del nivel de estrés:

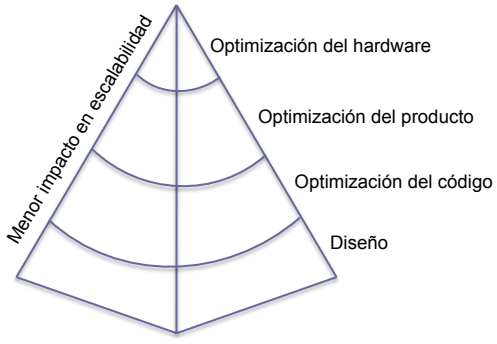
- Cambios en las aplicaciones
- Fallos o caídas de algunas partes del sistema
- Incremento del número de máquinas
- Incremento repentino del número de usuarios del sitio

La ***escalabilidad*** se refiere a la capacidad de un sistema de manejar la carga, y el esfuerzo para adaptarse al nuevo nivel de carga.

16

Escalabilidad

escalabilidad = capacidad de un sistema de manejar la carga, y el esfuerzo para adaptarse al nuevo nivel de carga



http://bit.ly/Z2II1G

Escalabilidad

Si un sitio gana popularidad, o si llega una fecha señalada, puede incrementarse su carga.

Para manejar esa carga, las empresas tienen más servidores de los necesarios normalmente.

Decidir cómo **añadir más recursos al sistema** web es crucial en el diseño inicial y en el mantenimiento.

En ocasiones, si la CPU del servidor está al 95% todo el tiempo, cambiándola puede ser suficiente para cierto nivel de carga. Pero si más adelante hay más carga, será insuficiente.

Escalabilidad

Vemos que hay dos tipos de escalado:

- Ampliación ***vertical***:
  incrementar la RAM, CPU, disco de un servidor.

- Ampliación ***horizontal***:
  añadir máquinas a algún subsistema (servidores web, servidores de datos, etc).

En ocasiones una ampliación vertical puede ser suficiente.

Escalabilidad

¿Cómo analizar la sobrecarga?

- Si la **CPU** está cerca del 100% todo el rato y el resto de subsistemas no está sobrecargado, sustituir por una CPU más potente.
- Si el uso de **RAM** es muy alto, veremos un uso alto de disco (por swapping). Incrementando la cantidad de RAM mejoraremos el rendimiento.
- Un **ancho de banda** insuficiente afectará al rendimiento. Contratando una mejor conexión será suficiente.

Escalabilidad

Ejercicio:

¿Cómo analizar el nivel de carga de cada uno de los subsistemas en el servidor?

Buscar herramientas y aprender a usarlas.

...¡o recordar cómo usarlas!

Índice

Escalar un sitio web

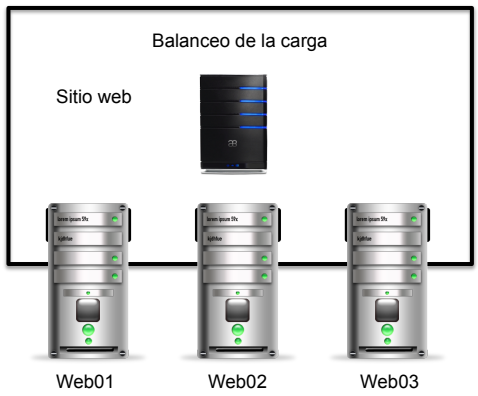
Tenemos que configurar tres niveles:

- máquinas como servidores web
- aplicaciones
- almacenamiento

Escalar un sitio web

El nivel web se puede configurar ***balanceando la carga***:

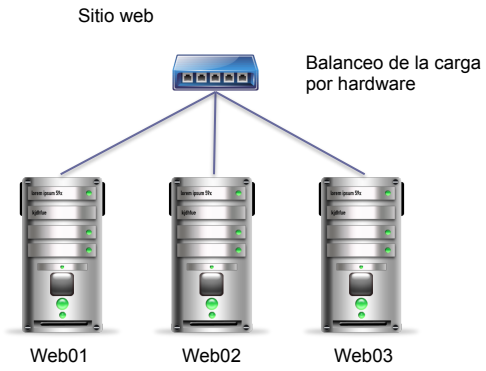
- uso de una máquina con software específico



Escalar un sitio web

También se puede usar un ***balanceador hardware***:

- Local Director (Cisco)
- ServerIron (Foundry)
- BigIP (F5)



25

Escalar un sitio web

El balanceador pasa peticiones a los servidores según el tráfico de la red.

Hay varios algoritmos para decidir qué máquina final servirá cada petición:

- Por turnos (round-robin)
- Según el menor número de conexiones
- Por ponderación
- Por prioridad
- Según el tiempo de respuesta

26

Escalar un sitio web

Escalar el ***nivel de aplicaciones*** requiere diseñar el software pensando en que se ejecute en varios servidores:

- Paralelismo
- Transparencia de ubicación: no debe haber dependencia de una máquina concreta para ejecutarse la aplicación.

Es importante diseñar las aplicaciones desde el principio para que se ejecuten en varios servidores.
Adaptar posteriormente una aplicación dependiente de cierto servidor puede ser costoso.

Escalar un sitio web

Escalar el ***nivel de almacenamiento*** es complejo y depende del tipo de servicios a ofrecer:

- LDAP: Protocolo Ligero de Acceso a Directorios
- NFS: Sistema de archivos de red
- Bases de datos

Cada uno de estos mecanismos suele requerir mecanismos y configuraciones diferentes.

Escalar un sitio web

Ejercicio:

Buscar ejemplos de balanceadores software y hardware (productos comerciales).

Buscar productos comerciales para servidores de aplicaciones.

Buscar productos comerciales para servidores de almacenamiento.

Índice

Conclusiones

Conceptos clave: escalabilidad y alta disponibilidad.

Monitorización para detectar problemas y determinar posibles mejoras del sitio web.

La escalabilidad se suele implementar **replicando servidores** para las mismas tareas.

Conseguir disponibilidad y escalabilidad mediante **balanceo de carga**.