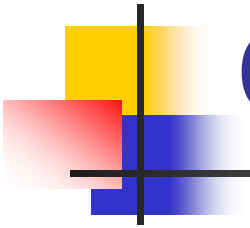
Propuestas arquitectónicas para servidores Web distribuidos con réplicas parciales

Universidad Carlos III de Madrid
Departamento de Informática
Doctorado en Ingeniería Informática

Septiembre de 2005

Autor: José Daniel García Sánchez

Directores: Jesús Carretero Pérez
Félix García Carballeira



Contenido

- Motivación
- Objetivo
- Propuestas arquitectónicas para replicación parcial
- Evaluación
- Conclusiones

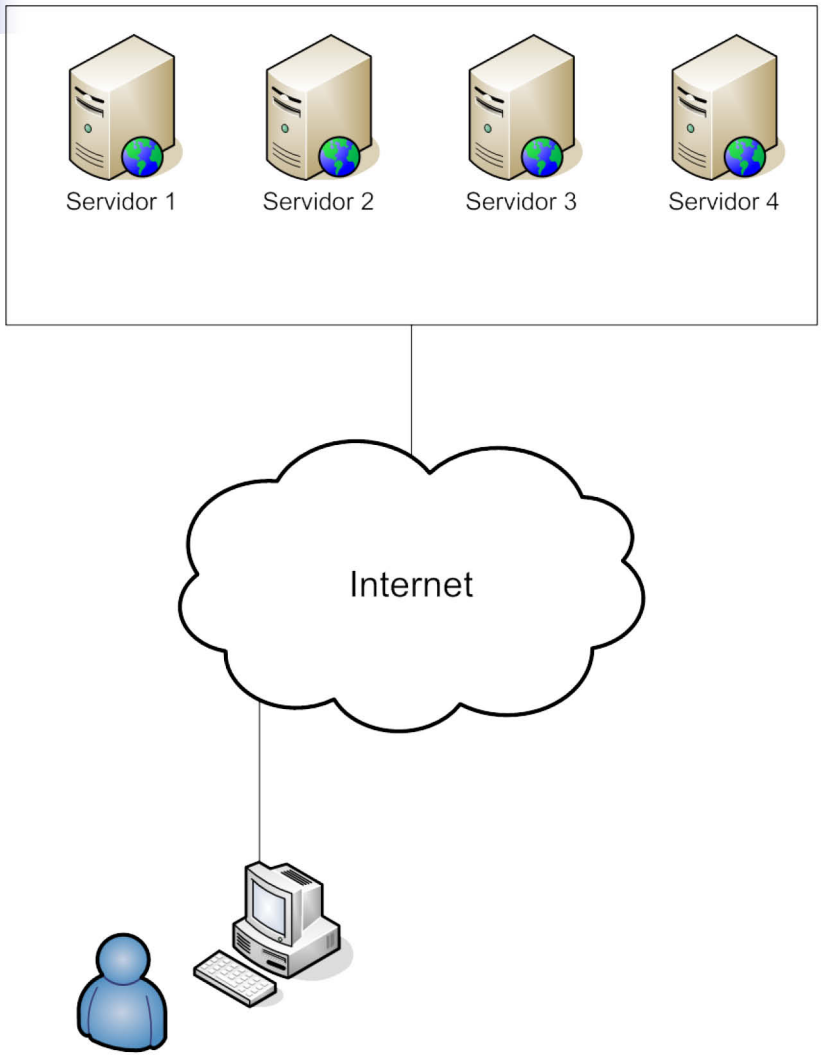
El problema de la escalabilidad

- Los servidores Web deben:
  - Satisfacer cada vez un mayor número de peticiones.
  - Alojar sitios que requieren mayor volumen de almacenamiento.
- El servidor Web es el único punto bajo control directo del proveedor de contenidos.
  - Mejora del servidor Web.
- El servidor Web contribuye en más del 40% a la latencia total.
  - Futuro ➔ Aumento del porcentaje de contribución.

Opciones para la mejora

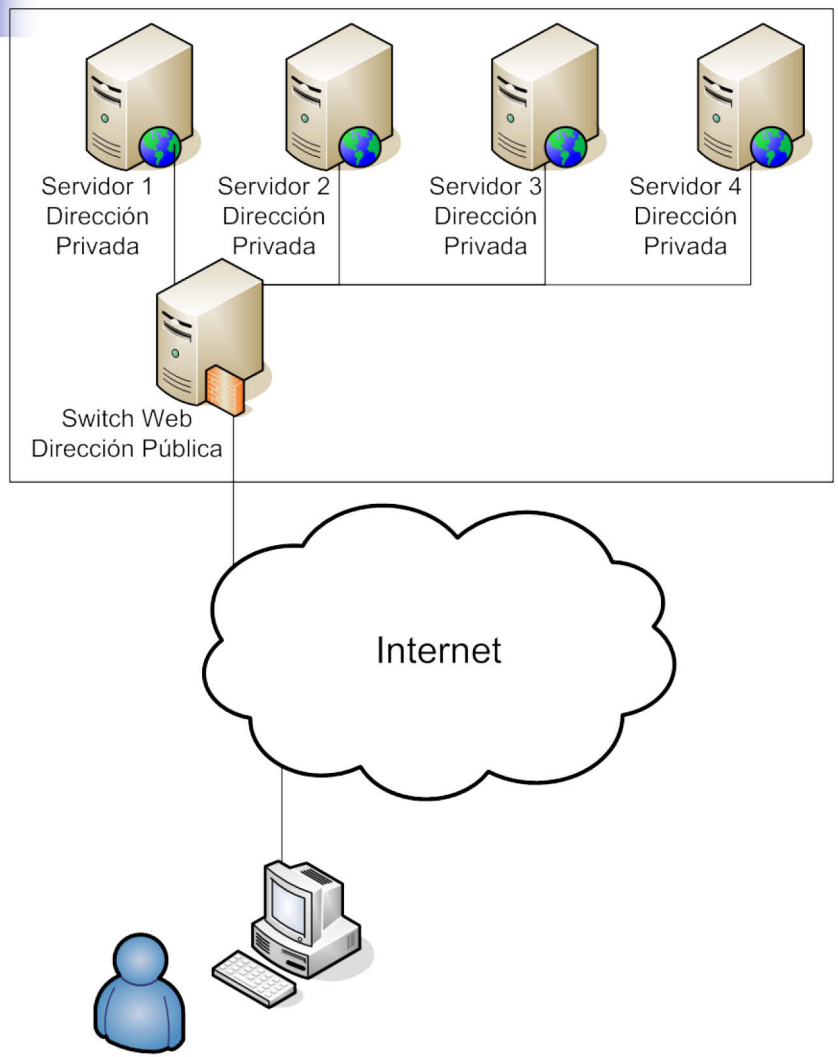
- Escalado hardware.
  - Migración a máquina de mejores prestaciones.
  - Incorporación de recursos a máquina existente: memoria, procesadores, disco, etc.
- Escalado software.
  - Mejora del sistema operativo:
    - Contenedores de recursos, mejora de llamadas al sistema, unificación de gestión de búferes y cachés.
  - Mejora del software servidor:
    - Mejora de cachés, aceleración de logs, cachés de URL, elusión de bloqueos.

Tipos de Arquitecturas distribuidas



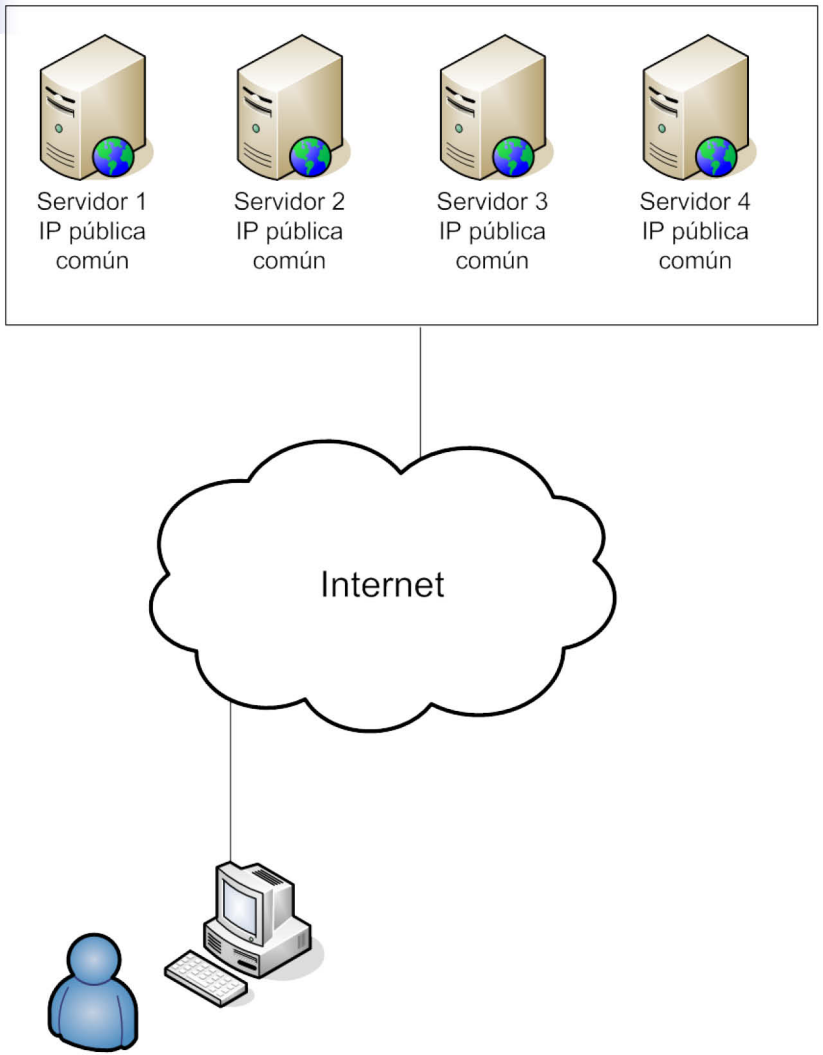
- Formadas por un conjunto de nodos servidores.
- Tipos:
  - Sistema Web basado en cluster.
  - Cluster Web virtual.
  - Sistema Web distribuido.

Sistema Web basado en Cluster



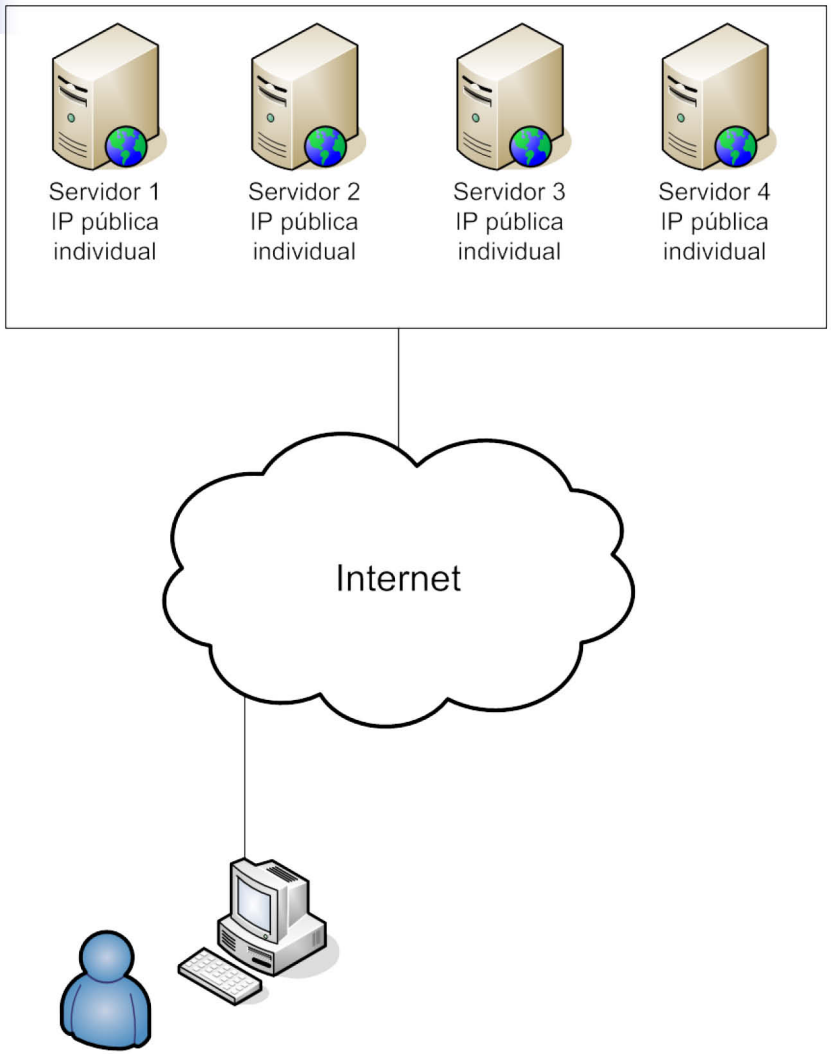
- Switch con dirección pública.
  - Distribuye las peticiones
- Nodos con direcciones privadas.
  - Sirven las peticiones.

Cluster Web virtual



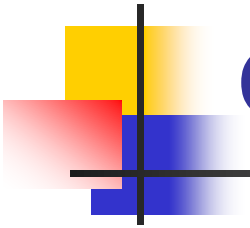
- Todos los nodos con idéntica dirección de red.
- Filtrado de peticiones en cada nodo.
- Mecanismo basado en función hash.

Sistema Web distribuido



- Direcciones de red públicas e individuales.
- Distribución de peticiones:
  - DNS dinámico.
  - Redirección de peticiones.

Tipos de replicación de contenidos



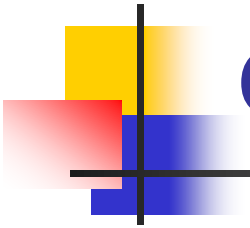
- Replicación total de contenidos.
  - Todos los archivos replicados en todos los nodos.
  - Alta fiabilidad.
  - Baja capacidad de almacenamiento.

- Distribución total de contenidos.
  - Cada archivo se aloja en un único nodo.
  - Baja fiabilidad.
  - Alta capacidad de almacenamiento.

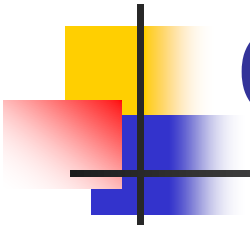
Políticas de asignación de peticiones

- Decide qué nodo debe procesar una petición.
  - Políticas sin información de estado.
    - Aleatoria estática, estática circular.
  - Políticas basadas en información del cliente.
    - Partición URL, clientes, servicios, SITA-E.
  - Políticas basadas en información del servidor.
    - Nodo menos cargado, LARD (distribución de peticiones consciente de la localidad).
  - Políticas propias de sistemas Web distribuidos.

Problemas de las soluciones existentes

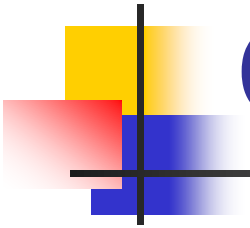


- Los escalados hardware y software no ofrecen soluciones a medio y largo plazo.
- La replicación total ofrece alta fiabilidad con bajo aprovechamiento de la capacidad de almacenamiento.
- La distribución total ofrece una alta capacidad de almacenamiento con baja fiabilidad.



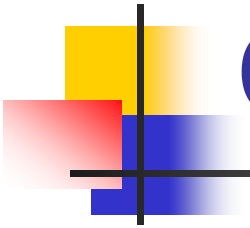
Contenido

- Motivación
- Objetivo
- Propuestas arquitectónicas para replicación parcial
- Evaluación
- Conclusiones



Objetivo

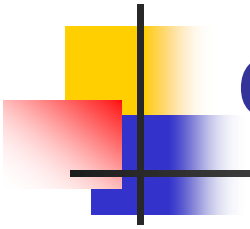
- Diseño de una arquitectura distribuida de servidor Web.
  - Basada en la replicación parcial de contenidos.
  - Alta escalabilidad en cuanto a los volúmenes de datos manipulados.
  - Sin deterioro de la fiabilidad.
  - Adaptación dinámica de la asignación de contenidos.



Contenido

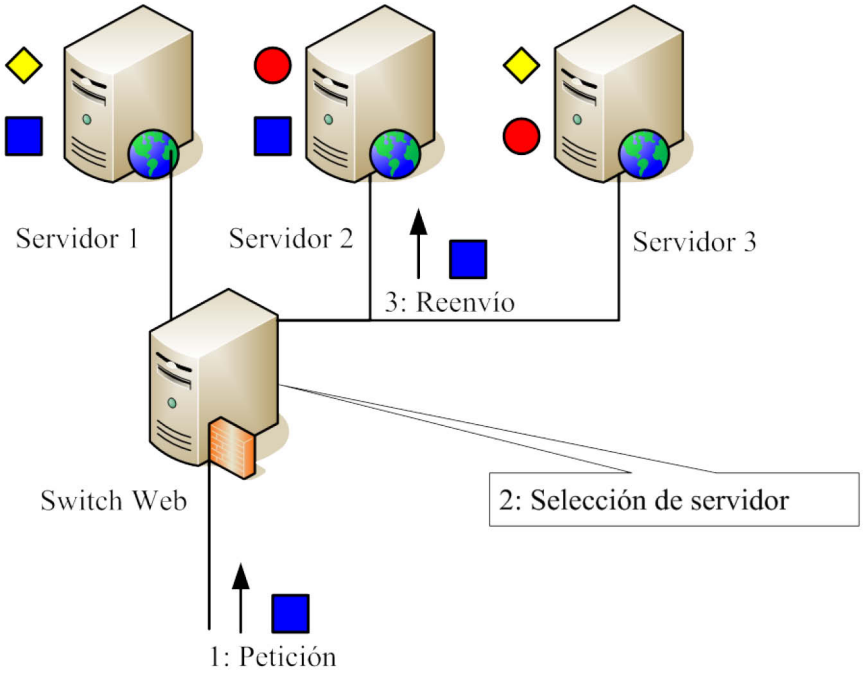
- Motivación
- Objetivo
- Propuestas arquitectónicas para replicación parcial
  - **Adaptación de arquitecturas**
  - Propuesta arquitectónica
  - Algoritmos de replicación
  - Políticas de asignación de peticiones
- Evaluación
- Conclusiones

Adaptación de las arquitecturas existentes a la replicación parcial

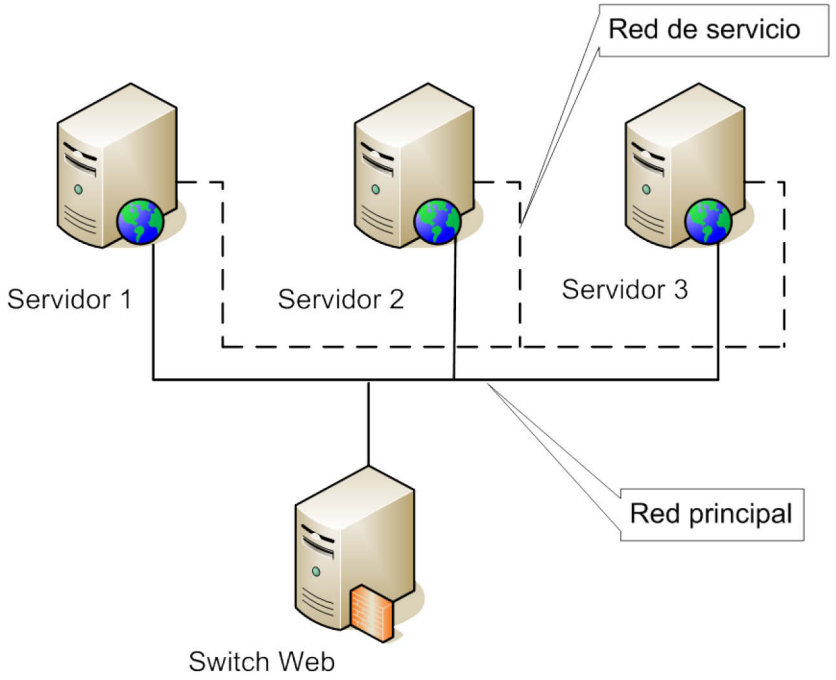


- **Sistemas Web basados en cluster.**
- Clusters Web virtuales.
- Sistemas Web distribuidos.

Sistemas Web basados en cluster

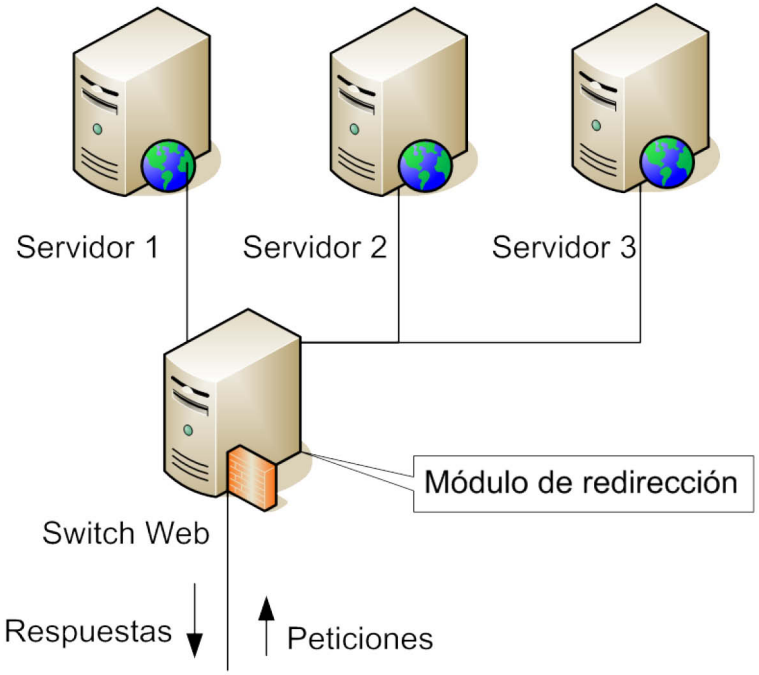


Replicación parcial

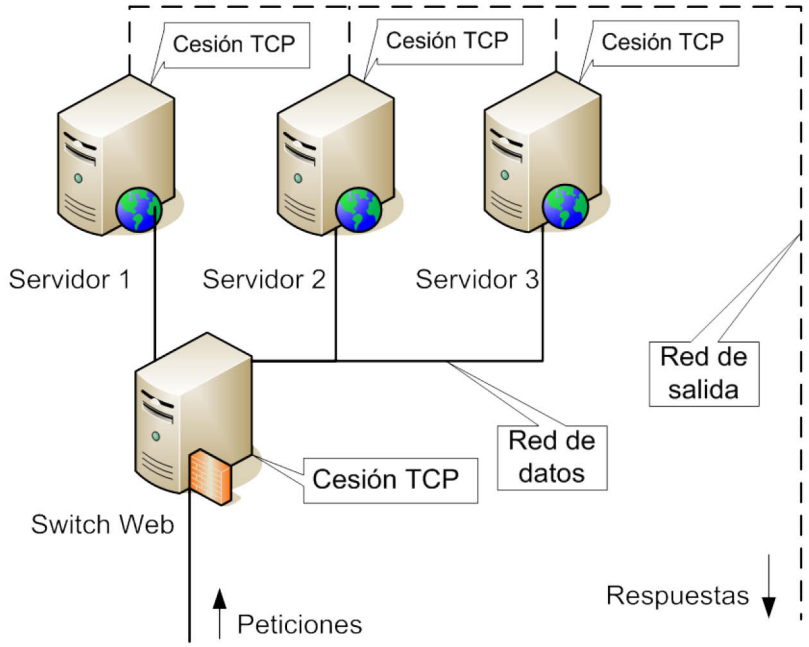


Asignación dinámica

Cluster Web: flujo de peticiones

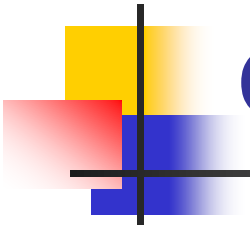


Bidireccional



Unidireccional

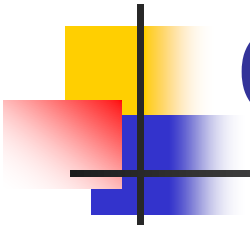
Sistema Web basado en cluster



- Switch Web: Nodo central encargado de coordinación.
- Solamente son admisibles políticas basadas en servicio de directorio.
- El Switch Web supone un punto único de fallo → limitación de fiabilidad.
- El Switch Web único supone un cuello de botella en el procesamiento de peticiones → limitación de rendimiento.

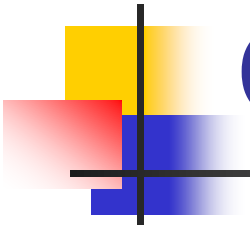
Adaptación de las arquitecturas existentes a la replicación parcial

- Sistemas Web basados en cluster.
- **Clusters Web virtuales.**
- Sistemas Web distribuidos.



Clusters Web virtuales

- Todas las peticiones llegan a todos los nodos.
  - Cada nodo descarta las peticiones que no le corresponden.
- Cada elemento puede estar alojado en un subconjunto distinto de nodos servidores.
  - No es posible utilizar una función hash general que opere exclusivamente sobre dirección y puerto de origen.



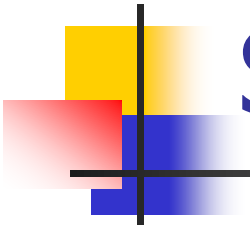
Clusters Web virtuales

- Necesidad de familia de funciones hash.
  - Una función por cada elemento.
  - Cada función con un conjunto distinto de posibles resultados.
- Necesidad de un mecanismo de selección.
- Complejo aplicar la idea a replicación parcial.
  - Dificultad de diseño de funciones hash.
  - Dificultad en selección de función hash.

Adaptación de las arquitecturas existentes a la replicación parcial

- Sistemas Web basados en cluster.
- Clusters Web virtuales.
- **Sistemas Web distribuidos.**

Sistemas Web distribuidos



- Redirección de peticiones de elementos no alojados en nodo destino a otro nodo que efectivamente contiene el elemento.

- Adaptación dinámica.
  - No existe un elemento central.
  - Basada en negociación entre nodos.

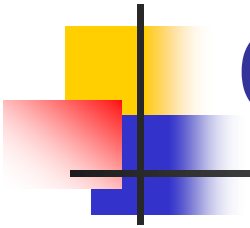
Ventajas / Inconvenientes

Cluster Web

- Ventajas
  - Escalabilidad.
  - Seguridad.
- Inconvenientes
  - Modificación del SO.
  - Punto único de fallo.

Sistema Web distribuido

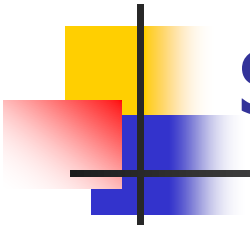
- Ventajas
  - No hay punto único de fallo.
- Inconvenientes
  - Escalabilidad (dir. IP).
  - Seguridad (IP pública).



Contenido

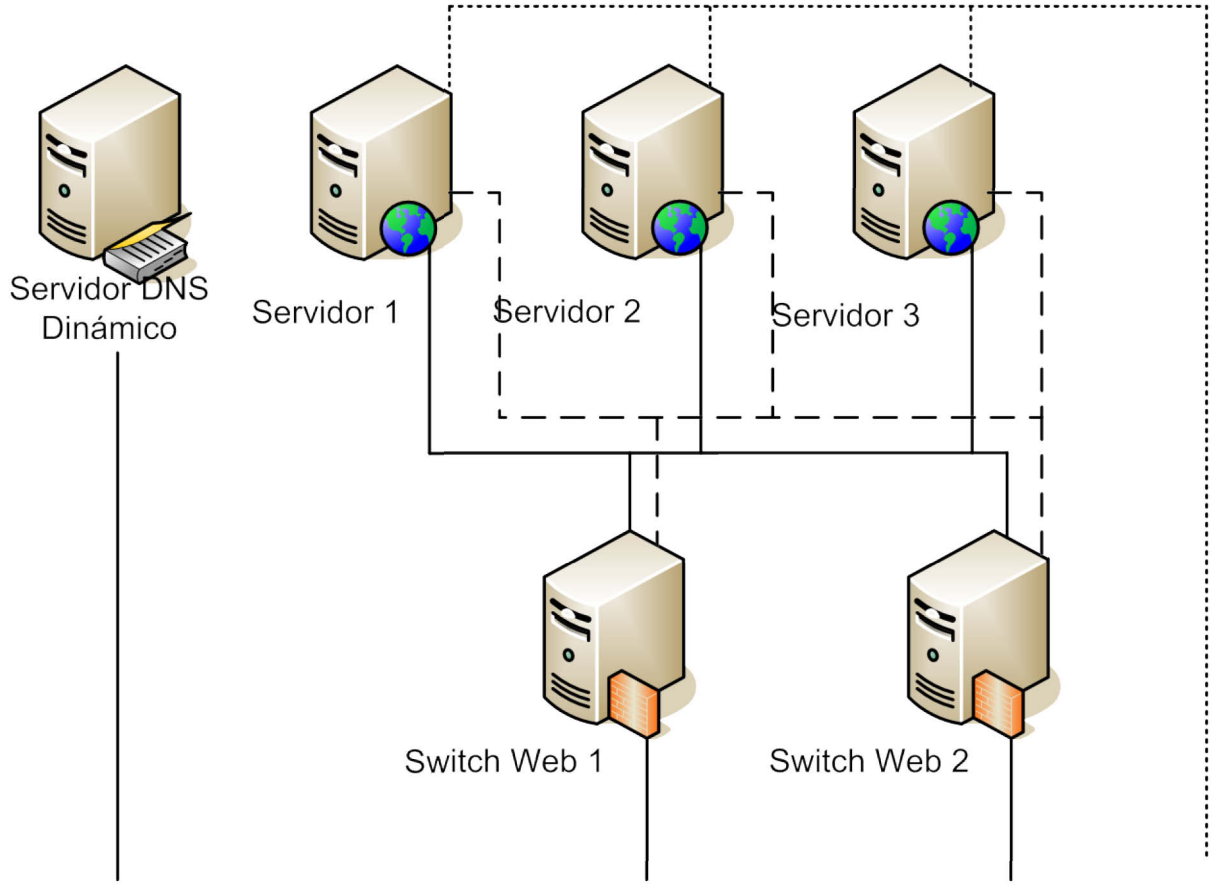
- Motivación
- Objetivo
- Propuestas arquitectónicas para replicación parcial
  - Adaptación de arquitecturas
  - **Propuesta arquitectónica**
  - Algoritmos de replicación
  - Políticas de asignación de peticiones
- Evaluación
- Conclusiones

Propuesta: Cluster Web con switch distribuido

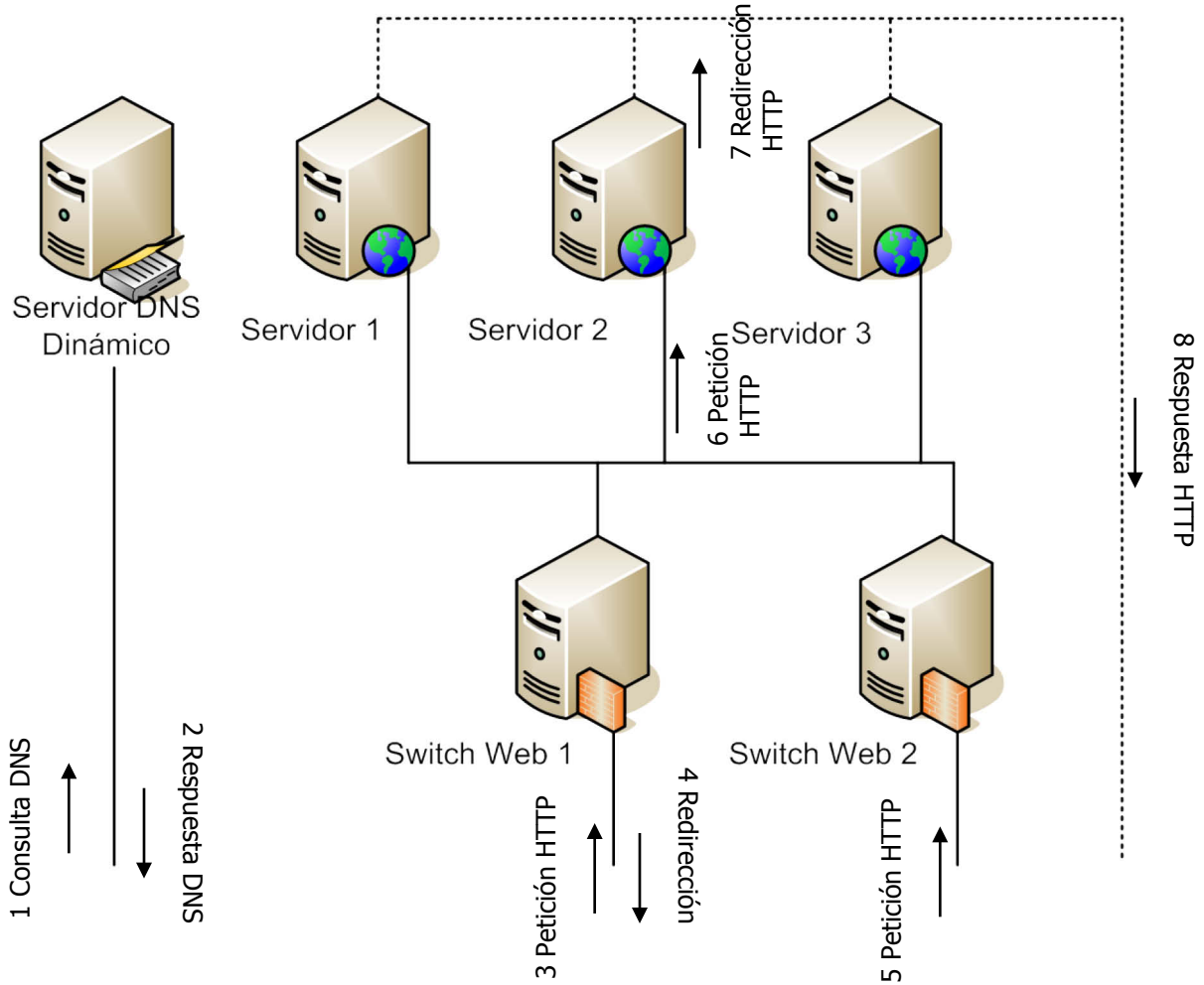


- Incorpora varios switches Web para mejorar la fiabilidad.
- Dos niveles de distribución de peticiones.
  - Entre switches.
  - Entre nodos.

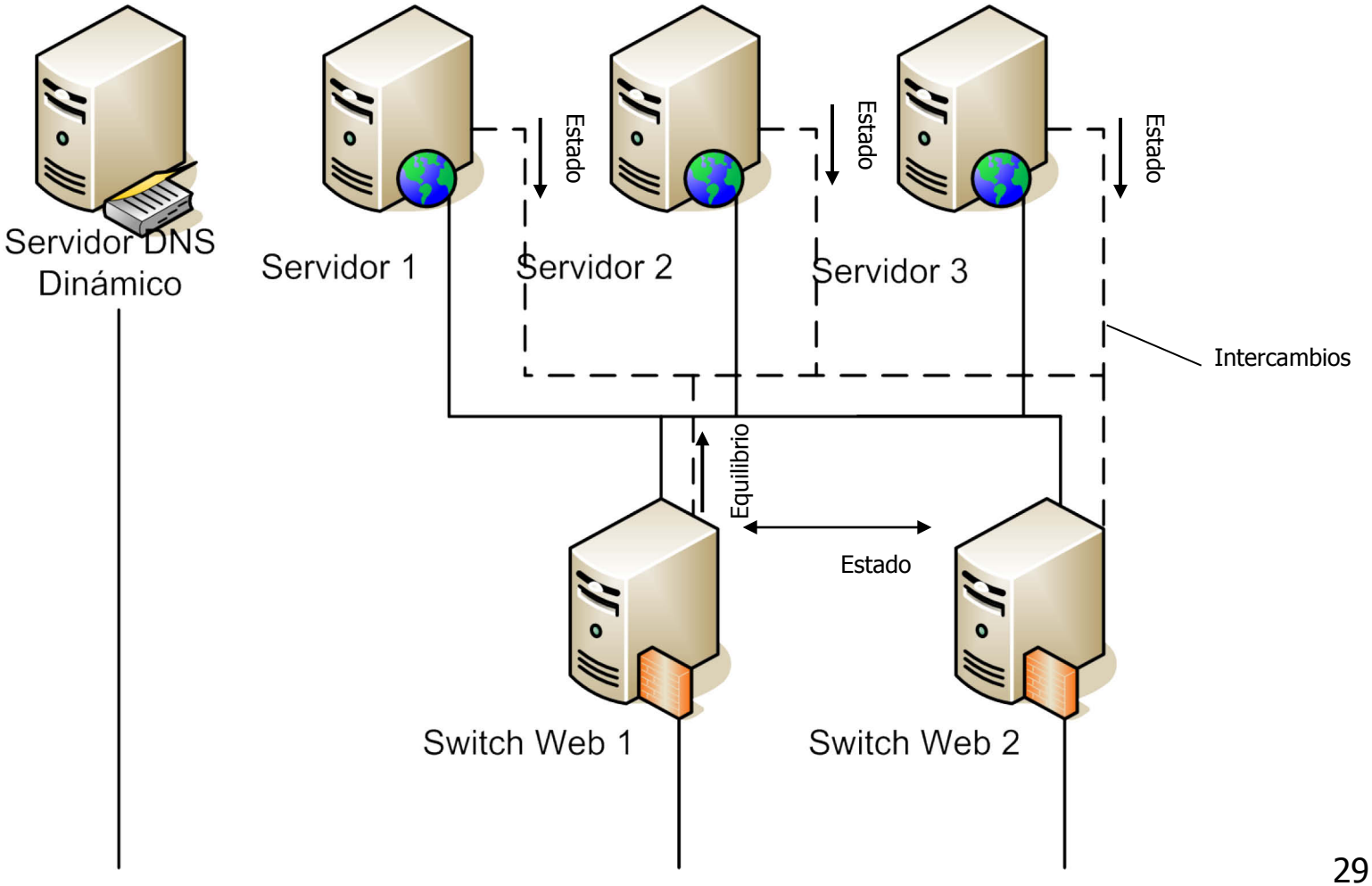
Cluster Web con switch distribuido

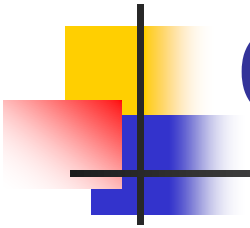


Procesamiento de una petición



Asignación dinámica de réplicas





Contenido

- Motivación
- Objetivo
- Propuestas arquitectónicas para replicación parcial
  - Adaptación de arquitecturas
  - Propuesta arquitectónica
  - **Algoritmos de replicación**
  - Políticas de asignación de peticiones
- Evaluación
- Conclusiones

Representación del problema

- Sitio Web → Conjunto de elementos.
  - E = {e_1, e_2, … , e_N}
- Servidor → Conjunto de nodos.
  - S = {s_1, s_2, … , s_M}
- Matriz de asignación: Representa la asignación de elementos a nodos servidores.

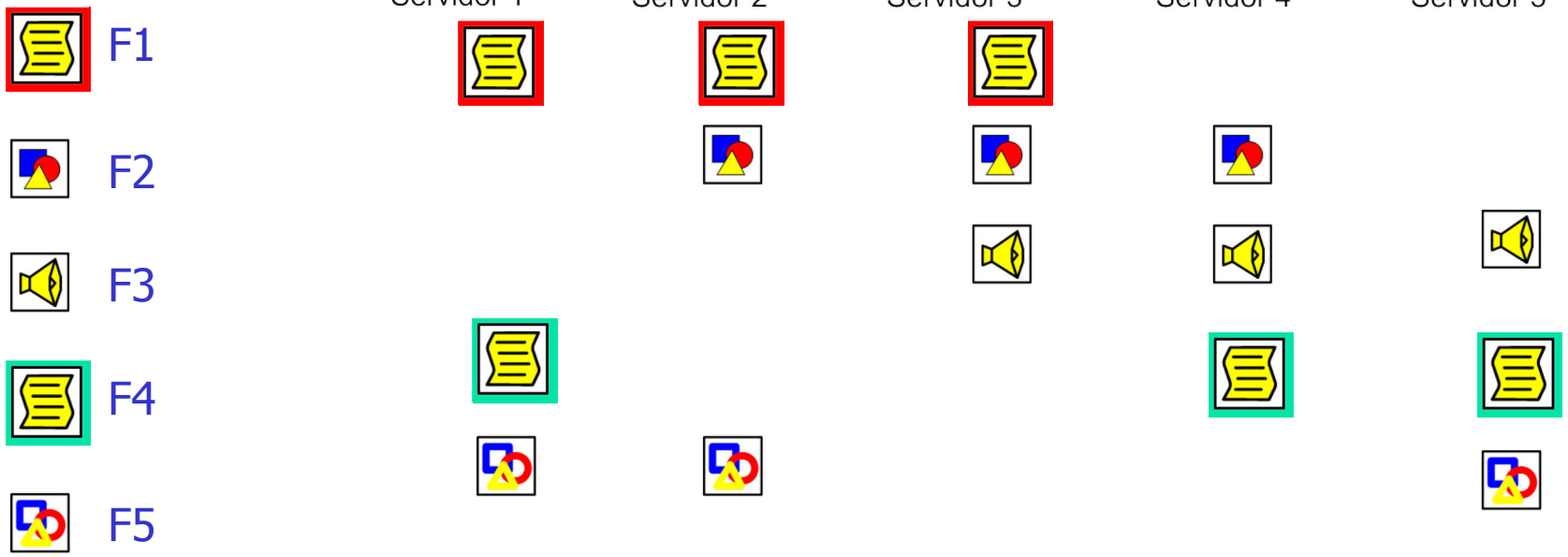
$$A = (a_{ij}) \mid a_{ij} = \begin{cases} 1 & e_i \quad \text{asignado a} \quad s_j \\ 0 & e_i \quad \text{no asignado a} \quad s_j \end{cases}$$

Algoritmos propuestos

- Asignación cíclica inicial.
  - Almacenamiento homogéneo.
  - Alto solape de elementos en nodos.
- Asignación cíclica final.
  - Almacenamiento homogéneo.
  - Menor solape de elementos en nodos.
- Asignación no equitativa.
  - Almacenamiento heterogéneo.
  - Considera frecuencias de acceso.

Asignación cíclica inicial

3 réplicas por elemento



Asignación cíclica final

3 réplicas por elemento


F1


F2


F3


F4


F5


Servidor 1


Servidor 2


Servidor 3


Servidor 4


Servidor 5





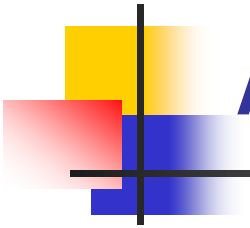












Asignación no equitativa

- Asigna un número de réplicas distinto a cada elemento.
- Basado en probabilidades de acceso a los elementos (p. ej. Zipf).
- Tiene en cuenta las restricciones de tamaño.
- Compatible con capacidades de almacenamiento heterogéneas.

Número de réplicas

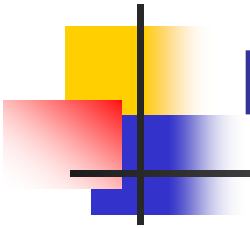
- Determinación del número de réplicas como cuota de espacio dividido por tamaño.

$$r_i = \frac{p_i \sum_{j=1}^{M} c_j}{t_i}$$

- Ajuste del número de réplicas a un número entero entre 1 y M.

$$r_i^* = \begin{cases} 1 & r_i < 1 \\ \lfloor r_i \rfloor & 1 \le r_i < M \\ M & r_i > M \end{cases}$$

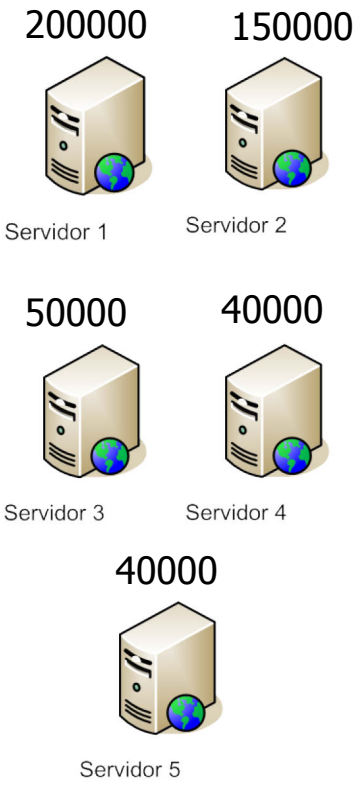
Algoritmo de asignación de réplicas



- Realiza una asignación voraz de réplicas, asignando primero los elementos de mayor tamaño.

- Si en algún momento se viola una restricción de capacidad, se reduce el número de réplicas de un elemento.

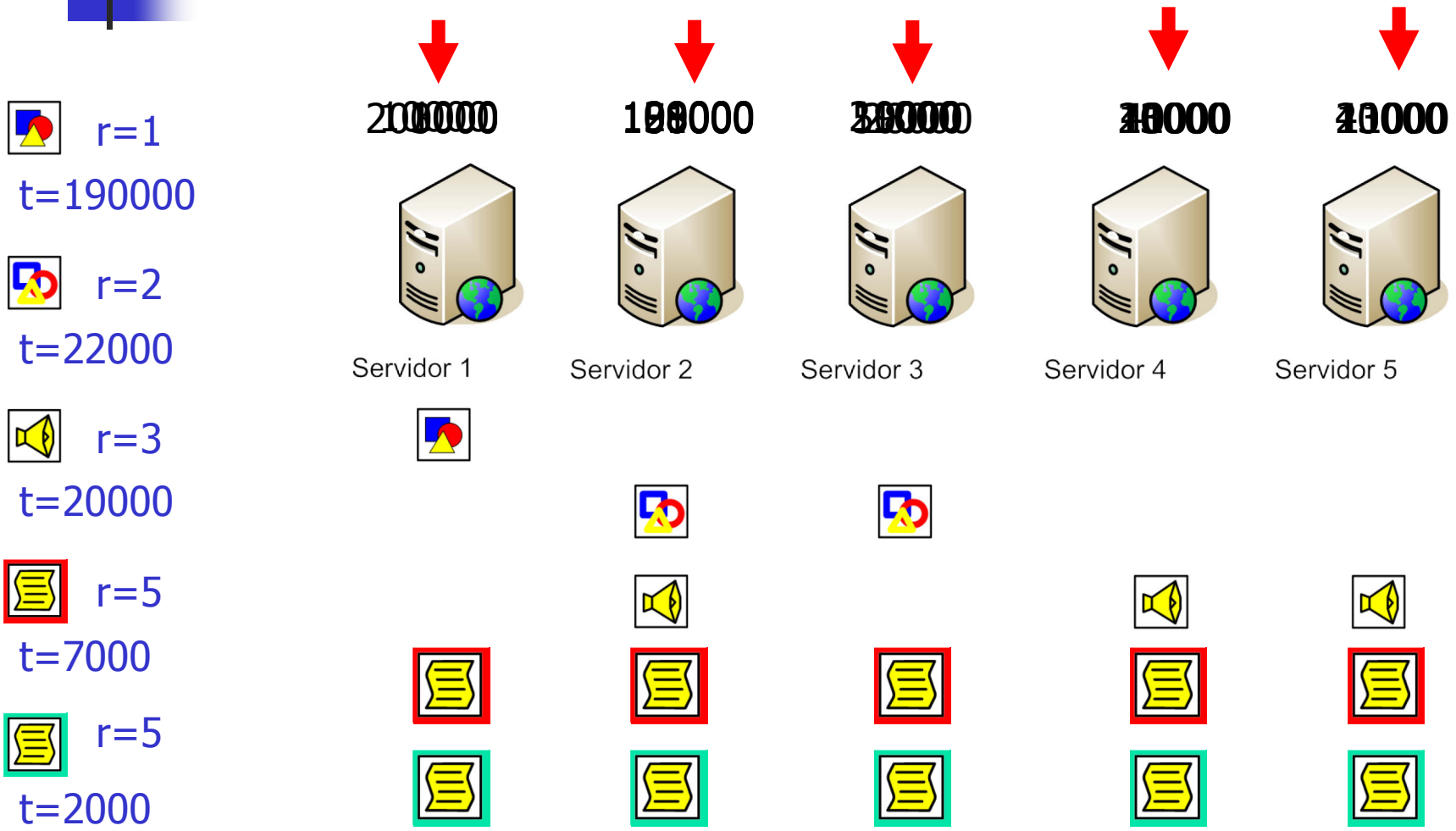
Asignación no equitativa



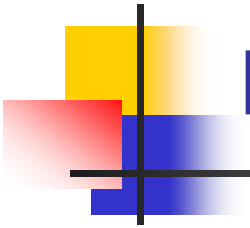
	Tamaño	Probabilidad	r	r*
	2000	0,4380	105,11	5
	190000	0,0876	0,22	1
	20000	0,1460	3,50	3
	7000	0,2190	15,02	5
	22000	0,1094	2,39	2

Volumen = 480000
Servidores = 5

Asignación no equitativa

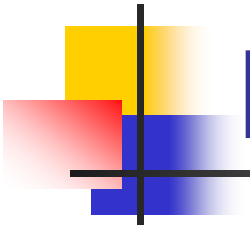


Asignación dinámica de réplicas



- Determinación del vector de probabilidades.
  - Mediante frecuencia de peticiones (estadísticas de actividad pasada).
- Condición de reasignación.
  - Basada en distancia entre vectores de probabilidad.

Registro de la frecuencia de peticiones



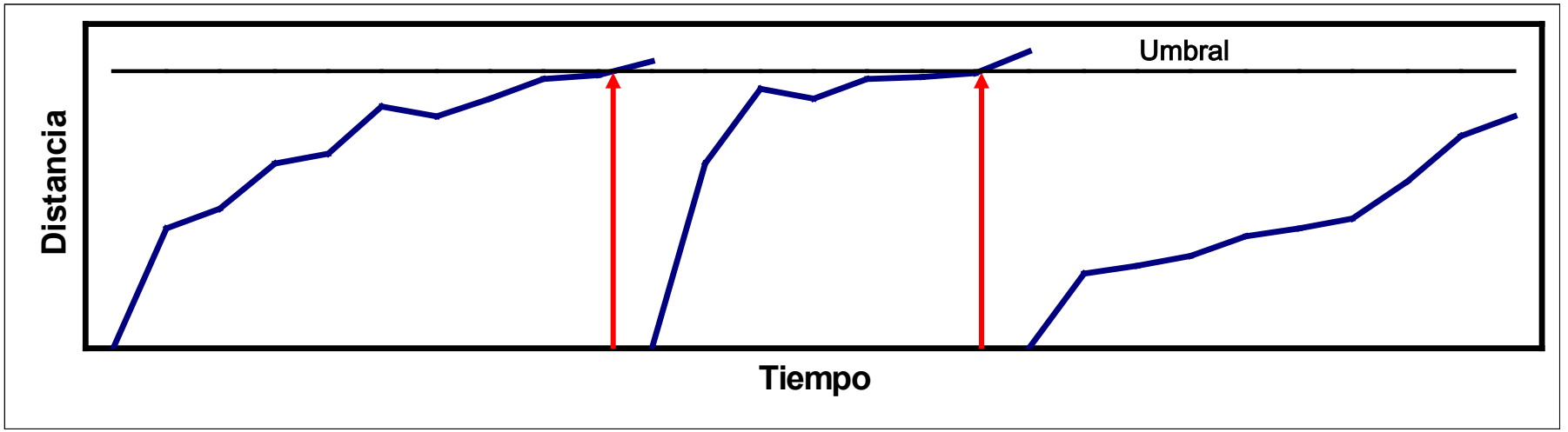
- Sistema Web basado en cluster.
  - Recogida centralizada en el switch.
- Sistema Web distribuido.
  - Recogida en cada nodo.
  - Fusión de la información en todos los nodos.
- Cluster Web con switch distribuido.
  - Recogida en cada switch.
  - Fusión de la información en switch primario.

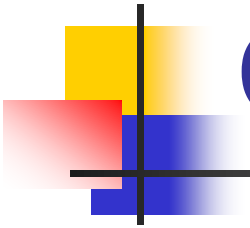
Reasignación de réplicas

- Distancia entre dos momentos como cantidad de réplicas a modificar.

$$\sum_{i=1}^{N} \left| r_i^*(t) - r_i^*(t+\Delta t) \right| > R$$

- Redistribución cuando distancia rebasa umbral.





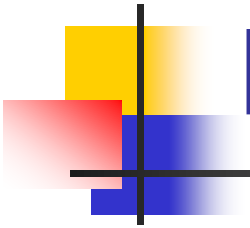
Contenido

- Motivación
- Objetivo
- Propuestas arquitectónicas para replicación parcial
  - Adaptación de arquitecturas
  - Propuesta arquitectónica
  - Algoritmos de replicación
  - **Políticas de asignación de peticiones**
- Evaluación
- Conclusiones

Adaptación de políticas

- Las políticas de asignación existentes no pueden utilizarse sin ser modificadas:
  - Es necesario tener en cuenta la replicación parcial de los elementos.

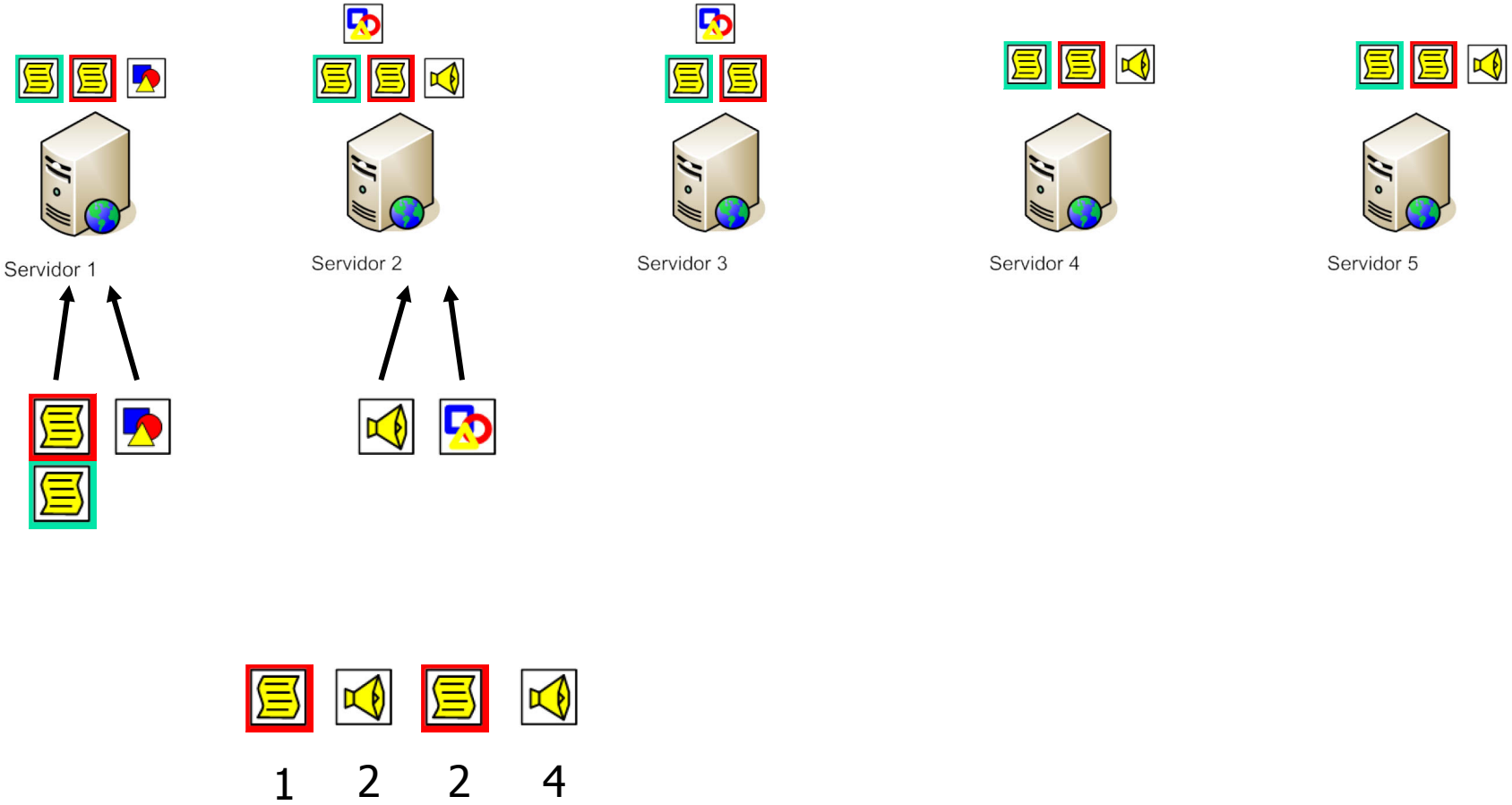
- Servicio de directorio basado en formalización de URL y tablas hash multinivel.



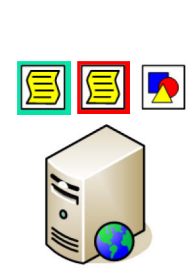
Políticas adaptadas

- Asignación estática (Round-Robin).
- Asignación al nodo menos cargado.
- LARD (Locality Aware Request Distribution).

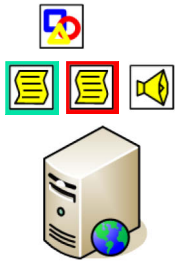
Asignación estática circular



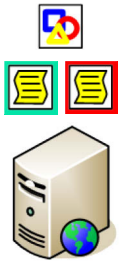
Asignación al nodo menos cargado



c=0,8
c=0,3
c=0,2



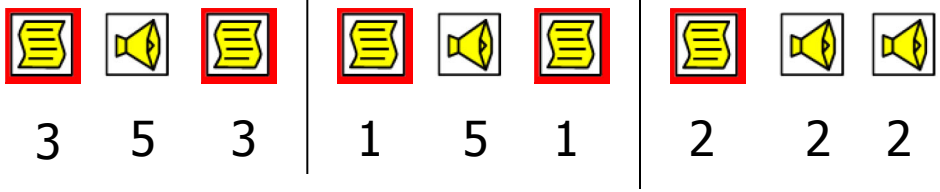
c=0,7
c=0,5
c=0,1



c=0,3
c=0,6
c=0,8

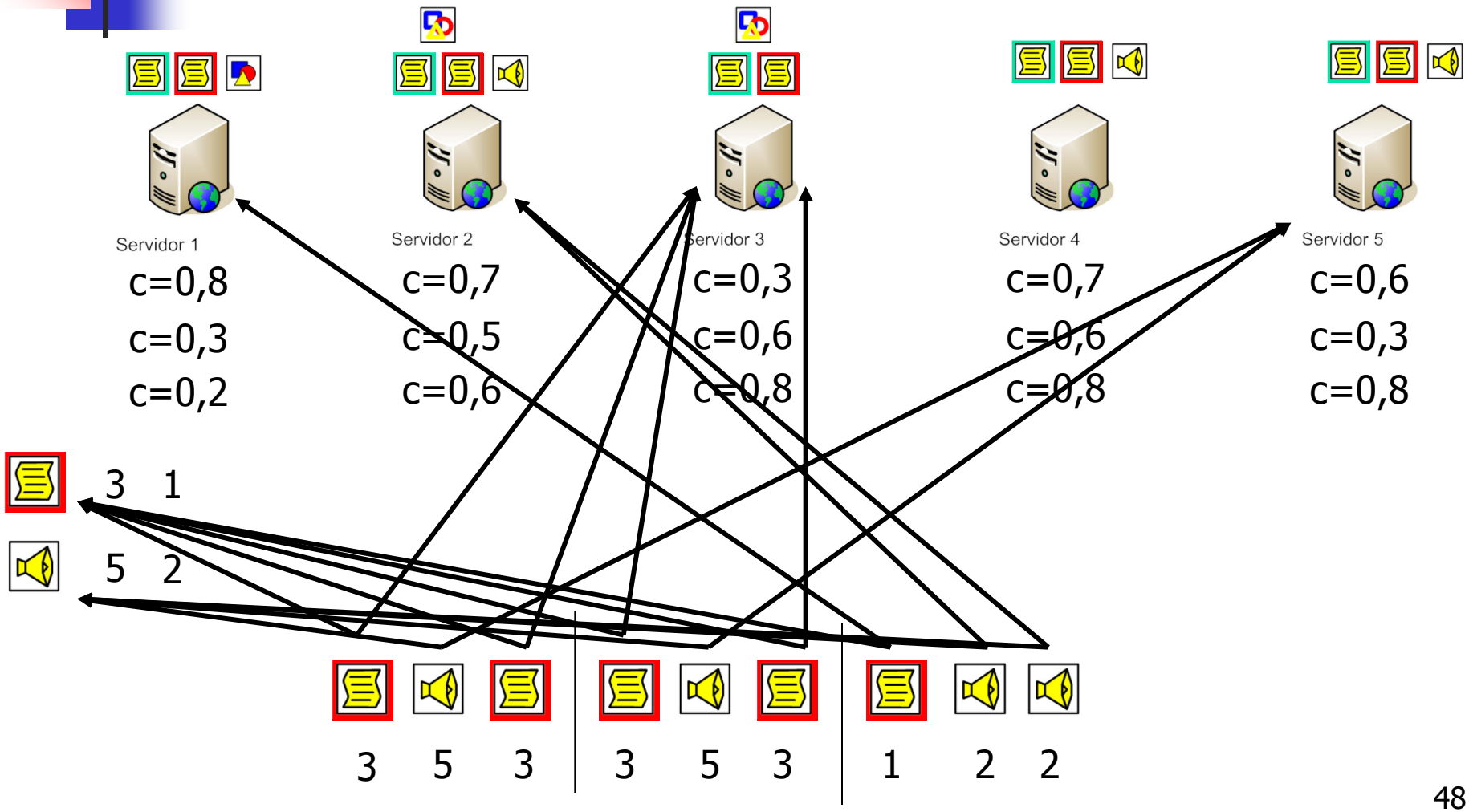
c=0,7
c=0,6
c=0,8

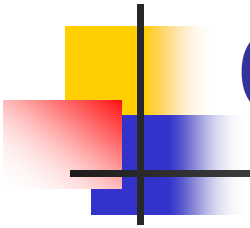
c=0,6
c=0,3
c=0,8



LARD

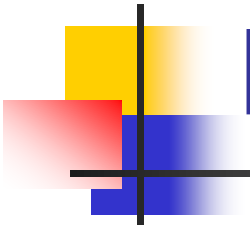
Cmax = 0,7


Servidor 1
c=0,8
c=0,3
c=0,2
Servidor 2
c=0,7
c=0,5
c=0,6
Servidor 3
c=0,3
c=0,6
c=0,8
Servidor 4
c=0,7
c=0,6
c=0,8
Servidor 5
c=0,6
c=0,3
c=0,8
3 1
5 2
3 5 3
3 5 3
1 2 2



Contenido

- Motivación
- Objetivo
- Propuestas arquitectónicas para replicación parcial
- Evaluación
  - **Rendimiento**
  - Capacidad de almacenamiento
  - Fiabilidad
- Conclusiones



Rendimiento

- Evaluación por simulación.
  - Basada en modelo estocástico.
  - 800 clientes realizando peticiones.
  - Cluster Web con 16 nodos servidores.
  - Discos de 200 GB.
  - Entorno de simulación: OMNET++.

Tipos de replicación

- **RTOT**: Replicación total.
- **RPCI2**: Replicación cíclica inicial con 2 réplicas por elemento.
- **RPCF2**: Replicación cíclica final con 2 réplicas por elemento.
- **RPCI4**: Replicación cíclica inicial con 4 réplicas por elemento.
- **RPCF4**: Replicación cíclica final con 4 réplicas por elemento.
- **RNOEQ**: Replicación no equitativa.

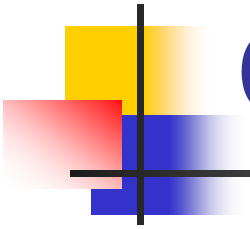
Análisis de resultados

- Evaluación del tiempo medio de respuesta.
- Comparación de tiempos medios mediante ANOVA (α=0,05).

Asignación cíclica	**Asignación al nodo menos cargado**	**Asignación LARD**
p=0,99999992	p=0,999999864	p=0,999999769

Conclusión de la simulación

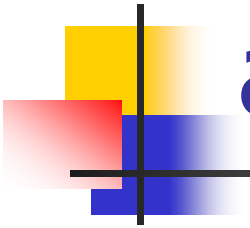
- No existe diferencia significativa en el tiempo de servicio de las peticiones Web entre un sistema totalmente replicado y un sistema parcialmente replicado.



Contenido

- Motivación
- Objetivo
- Propuestas arquitectónicas para replicación parcial
- Evaluación
  - Rendimiento
  - **Capacidad de almacenamiento**
  - Fiabilidad
- Conclusiones

Evaluación de la capacidad de almacenamiento



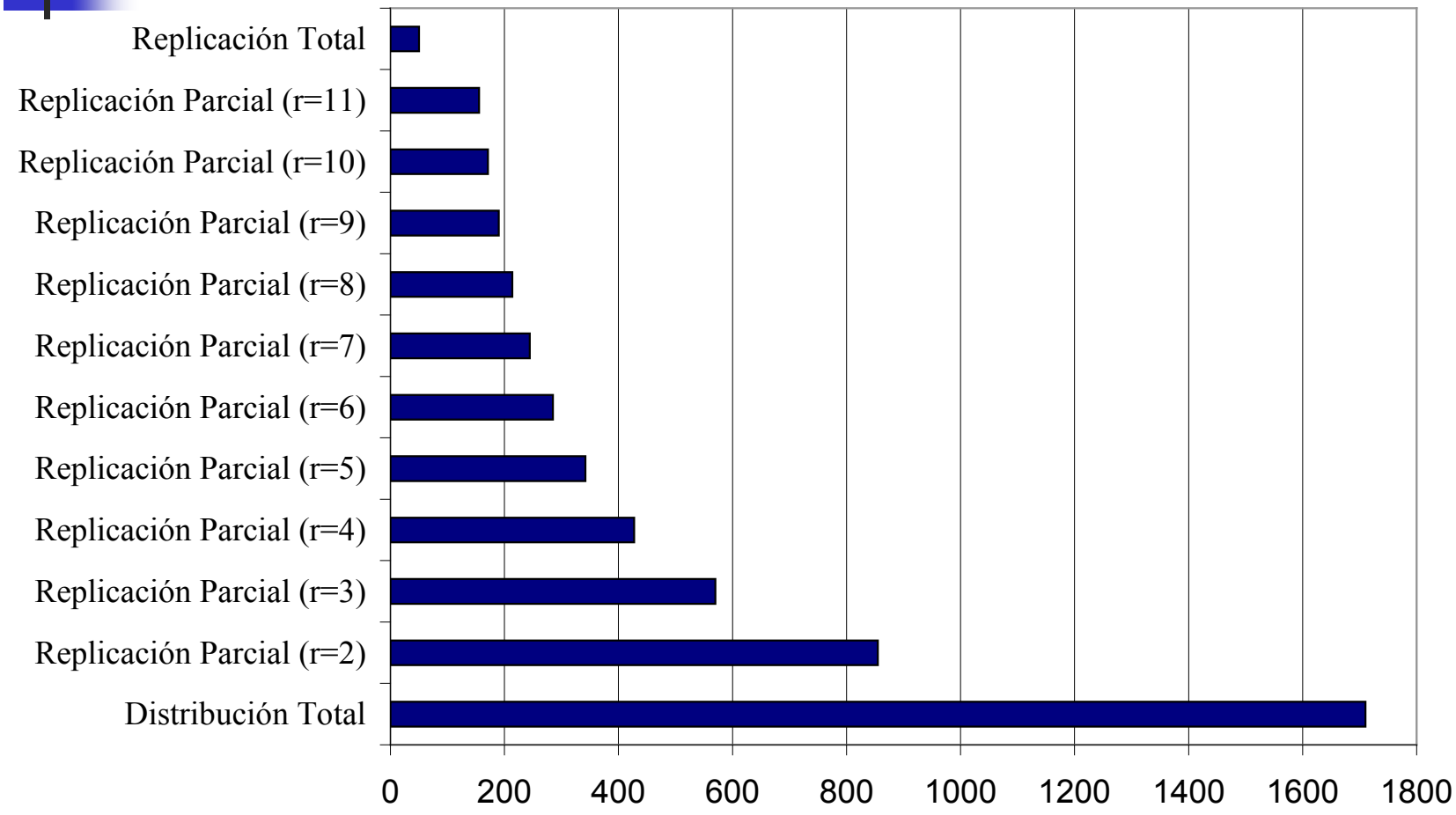
- Evaluación general de forma analítica.

- Aplicación a caso de estudio:
  - 12 nodos servidores.
  - Capacidades de entre 50 GB y 200 GB.
  - Tamaño del disco más grande que se puede adquirir: 300 GB.

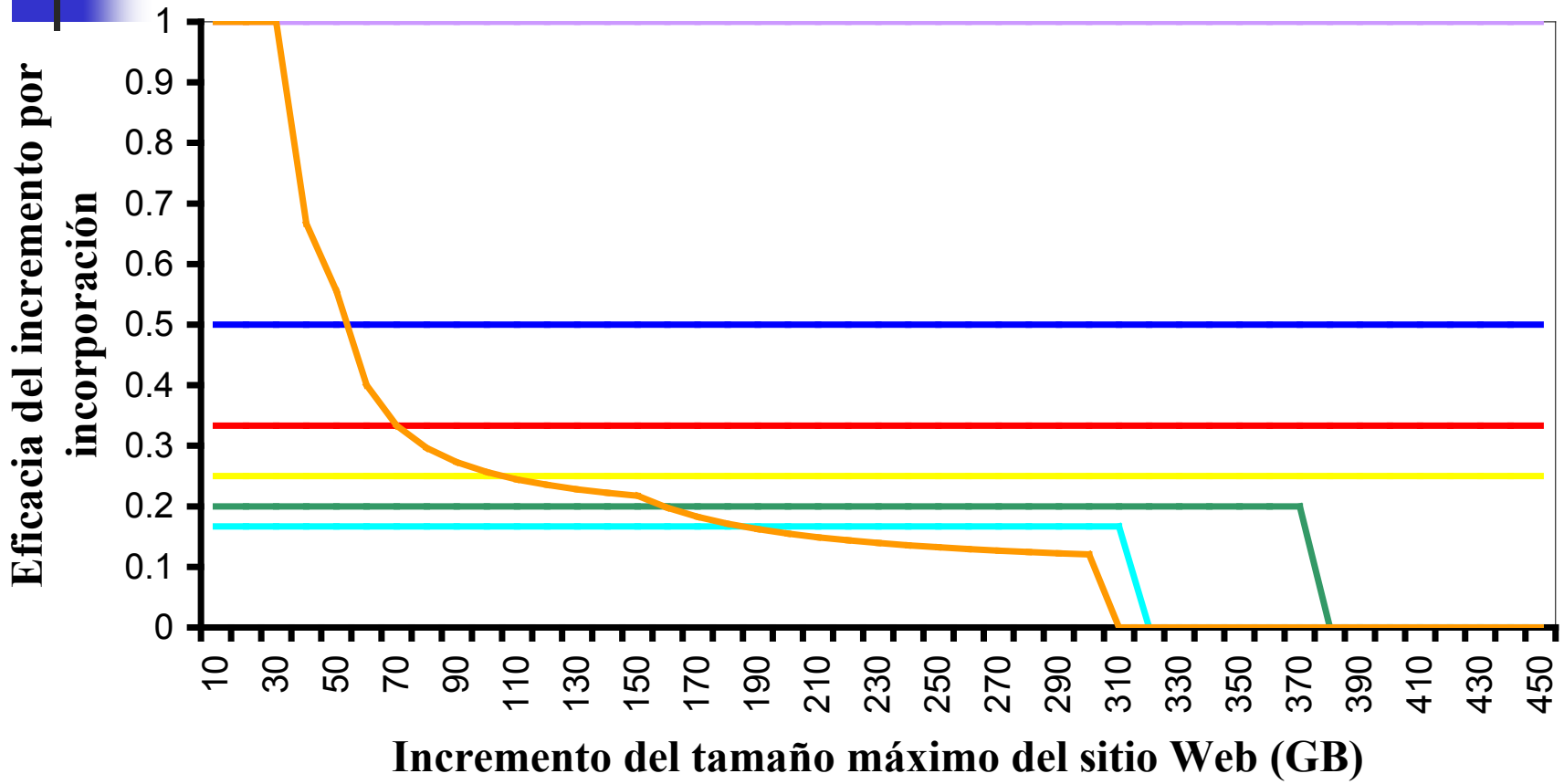
Métricas de capacidad

- Tamaño máximo del sitio Web.
- Eficacia de incremento por incorporación.
  - Tasa entre el incremento del tamaño máximo del sitio Web y el espacio físico incorporado.
- Eficacia de incremento por sustitución.
  - Tasa entre el incremento del tamaño máximo del sitio Web y el espacio físico sustituido.

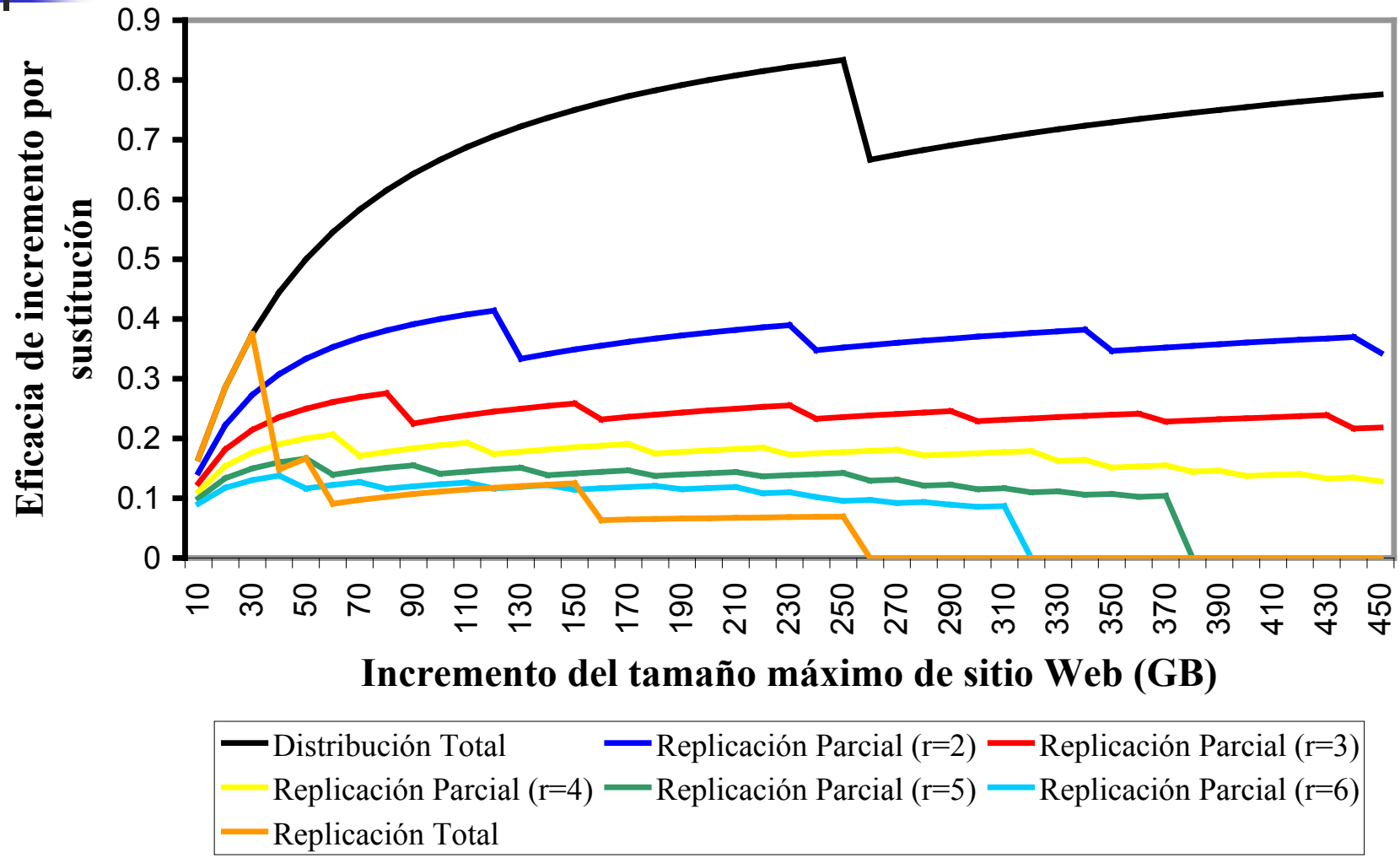
Tamaño máximo de sitio

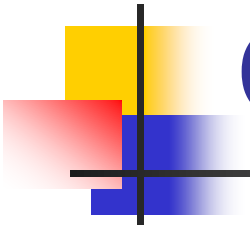


Eficacia de incremento por incorporación



Eficacia de incremento por sustitución

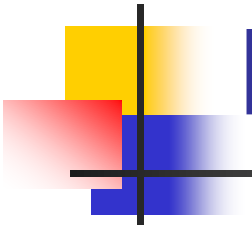




Contenido

- Motivación
- Objetivo
- Propuestas arquitectónicas para replicación parcial
- Evaluación
  - Rendimiento
  - Capacidad de almacenamiento
  - **Fiabilidad**
- Conclusiones

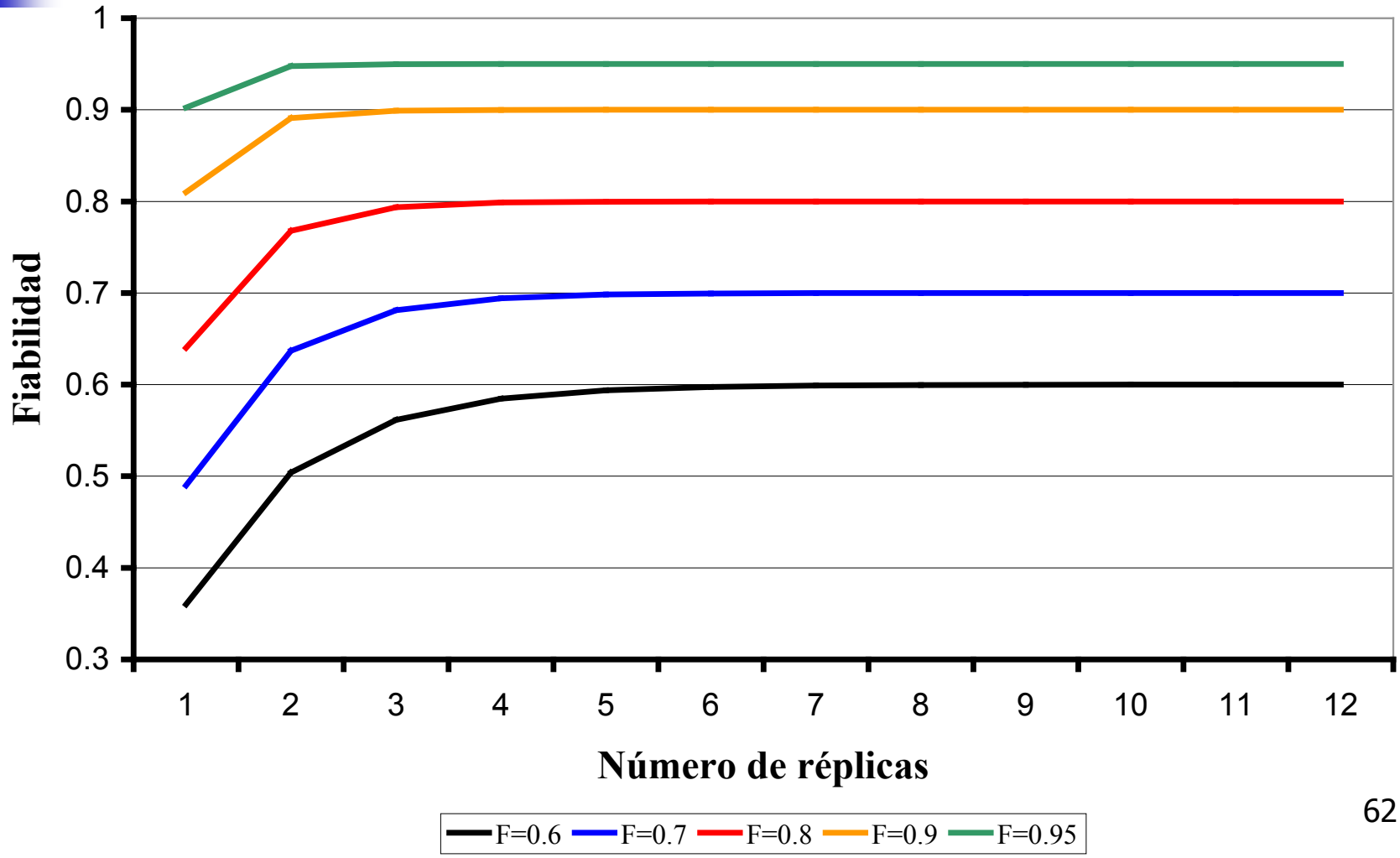
Evaluación de la fiabilidad



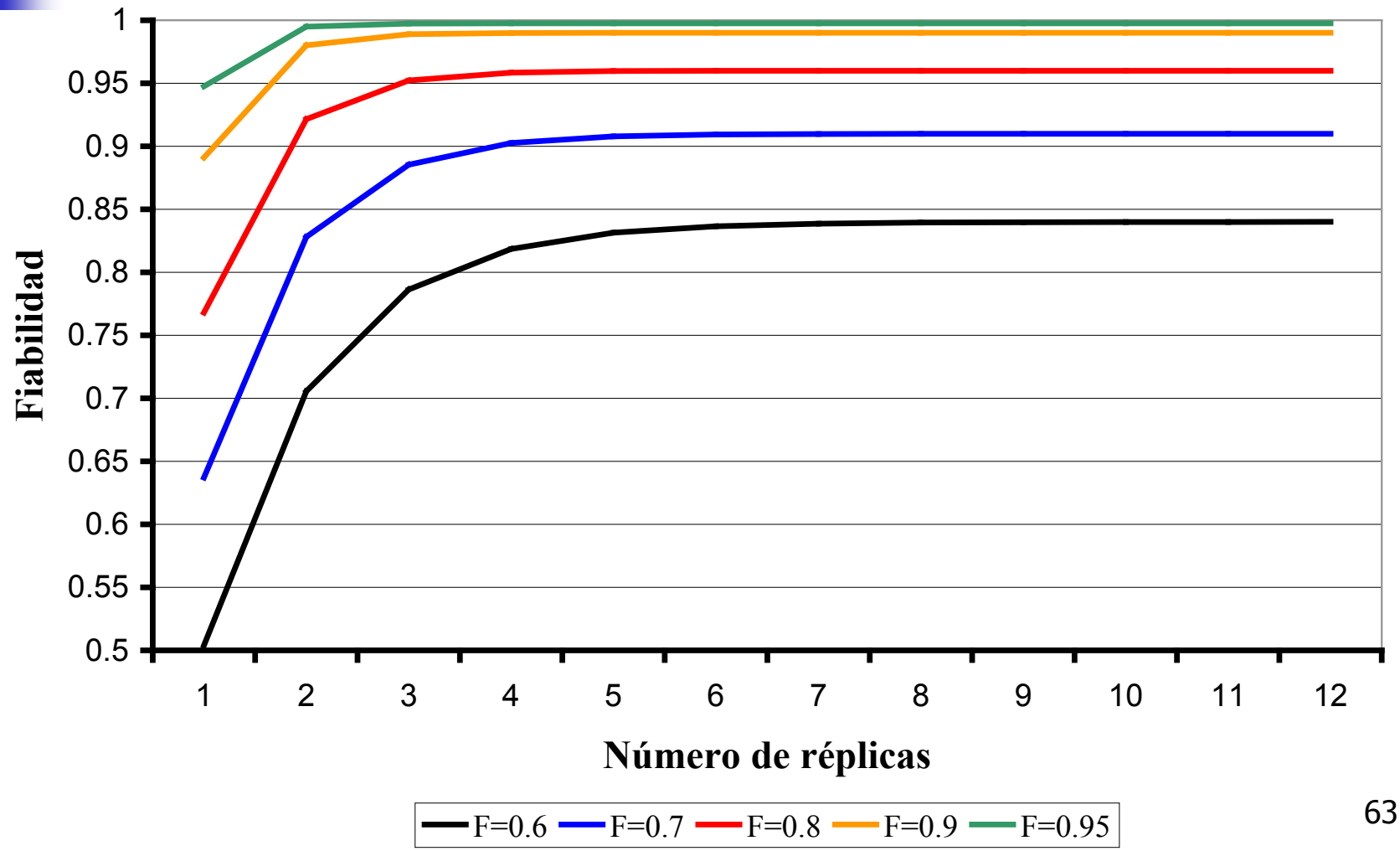
- Evaluación general de forma analítica.

- Aplicación a caso de estudio:
  - 12 nodos servidores.
  - Variación en el número de réplicas.
  - Variación en el número de switches.

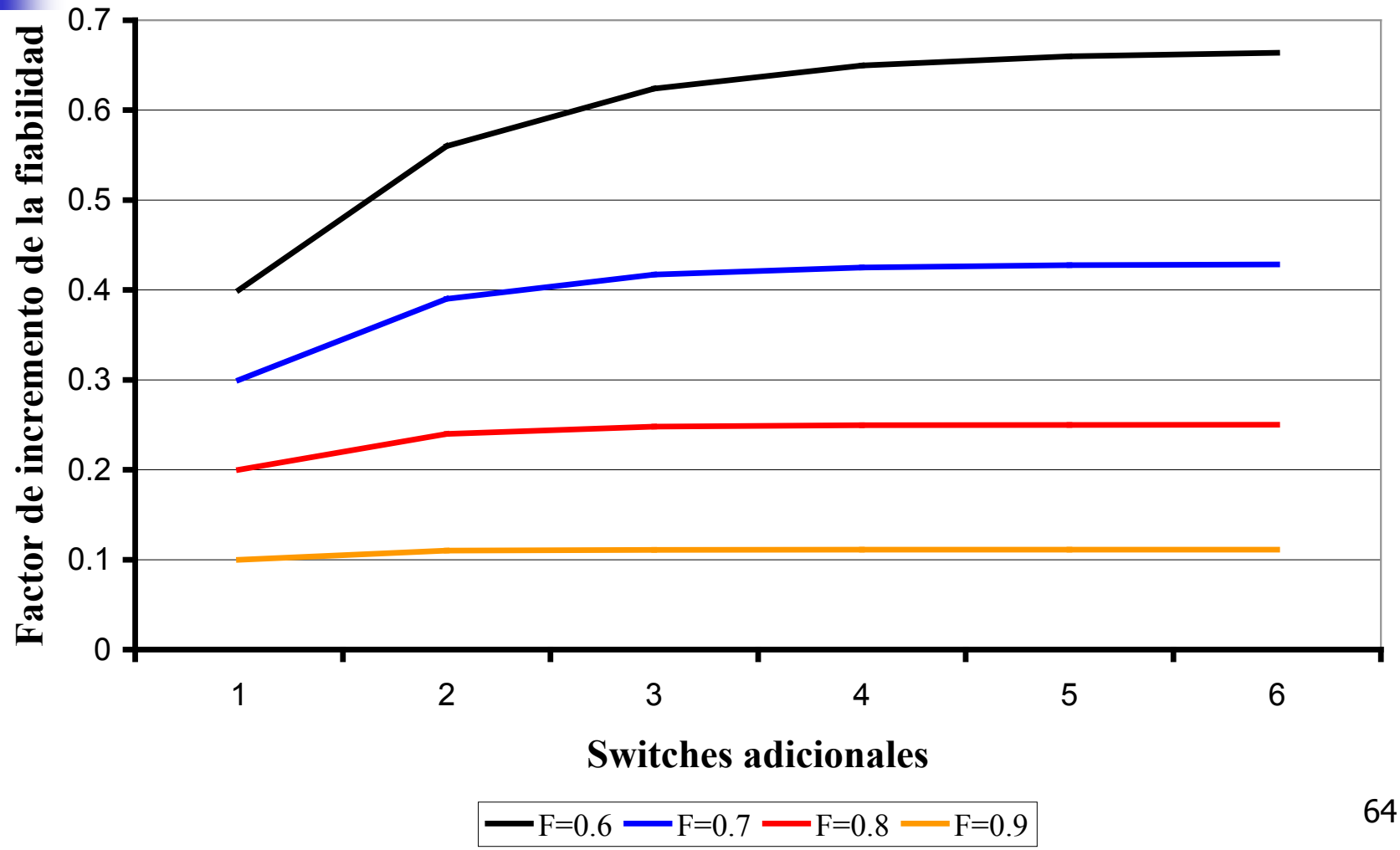
Fiabilidad de un cluster Web



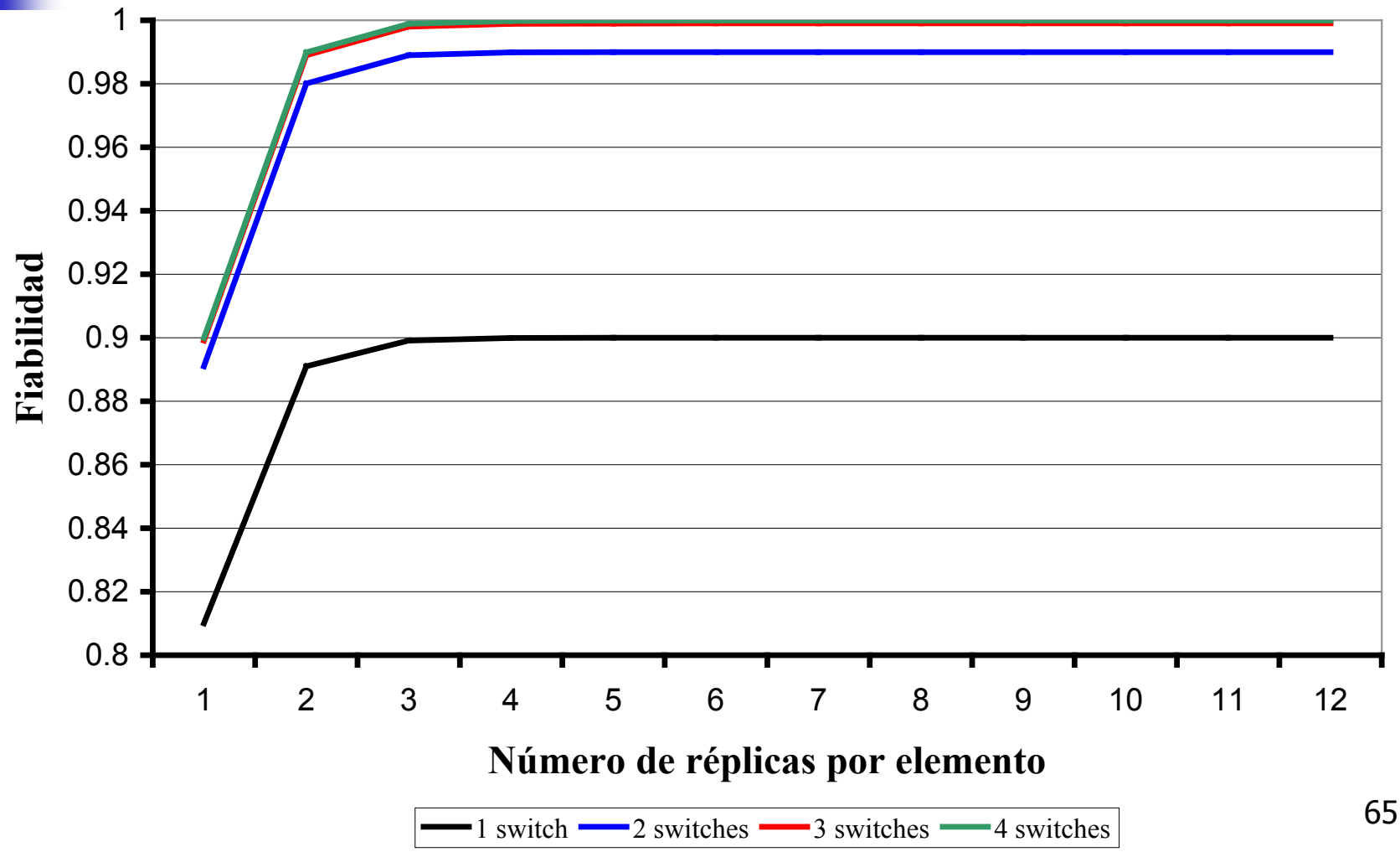
Fiabilidad de un cluster Web con 2 switches



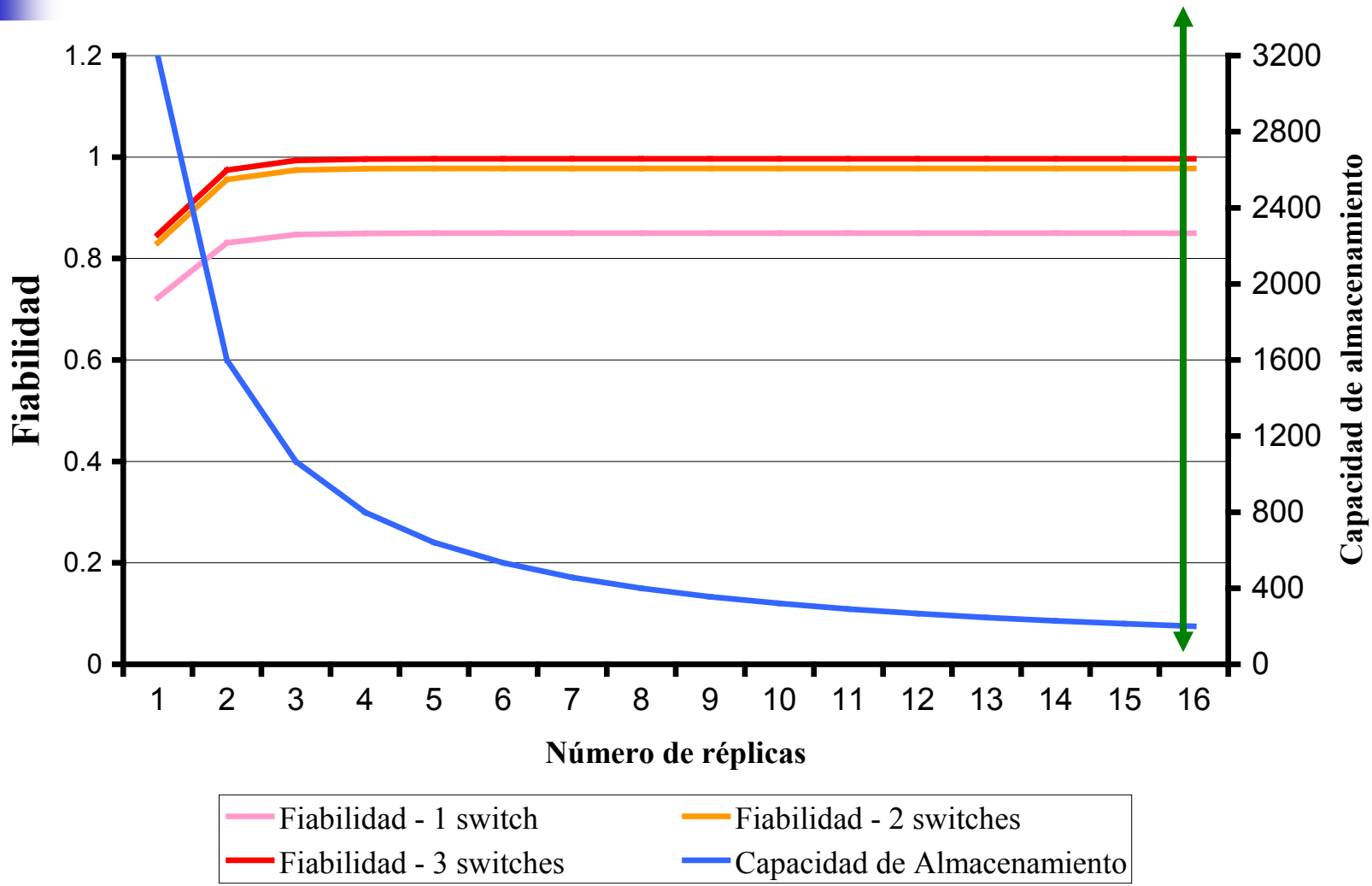
Incremento de la fiabilidad

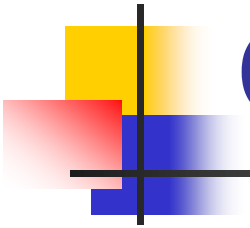


Cluster Web con switch distribuido: comparación



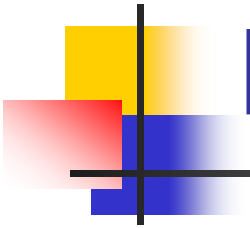
Capacidad de almacenamiento y fiabilidad





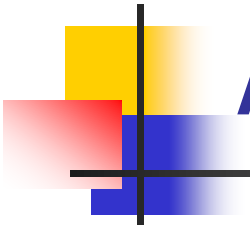
Contenido

- Motivación
- Objetivo
- Propuestas arquitectónicas para replicación parcial
- Evaluación
- **Conclusiones**



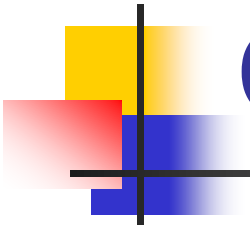
Resumen

- Se ha diseñado una arquitectura distribuida de servidor Web.
  - Basada en la replicación parcial de contenidos.
  - Alta escalabilidad en cuanto a los volúmenes de datos manipulados.
  - Sin deterioro de la fiabilidad.
  - Adaptación dinámica de la asignación de contenidos.



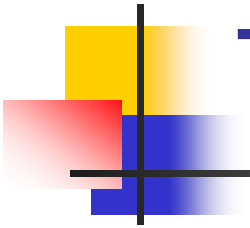
Aportaciones

- Una nueva arquitectura para servidor Web distribuido: el **cluster Web con switch distribuido**.
- Una política de asignación de réplicas basada en la frecuencia de acceso: la **replicación no equitativa**.
- La adaptación de políticas de asignación de peticiones en el contexto de **replicación parcial**.



Conclusiones

- No se produce pérdida de rendimiento en un sistema parcialmente replicado con respecto a un sistema totalmente replicado.
- La fiabilidad de un sistema Web basado en cluster está limitada por la fiabilidad de su switch Web.
- Un número relativamente bajo de réplicas (4-5) ofrece una fiabilidad equivalente a la de un sistema totalmente replicado.
- **Un sistema parcialmente replicado ofrece mayor capacidad de almacenamiento manteniendo la fiabilidad y sin pérdida de rendimiento.**



Trabajos futuros

- Determinar el efecto de la asignación dinámica de contenidos sobre el rendimiento.
- Ampliar el modelo para incluir el tratamiento de peticiones dinámicas.
- Extender la arquitectura para sistemas geográficamente distribuidos.
- Uso de técnicas no deterministas para la asignación de réplicas a nodos servidores.

Propuestas arquitectónicas para servidores Web distribuidos con réplicas parciales

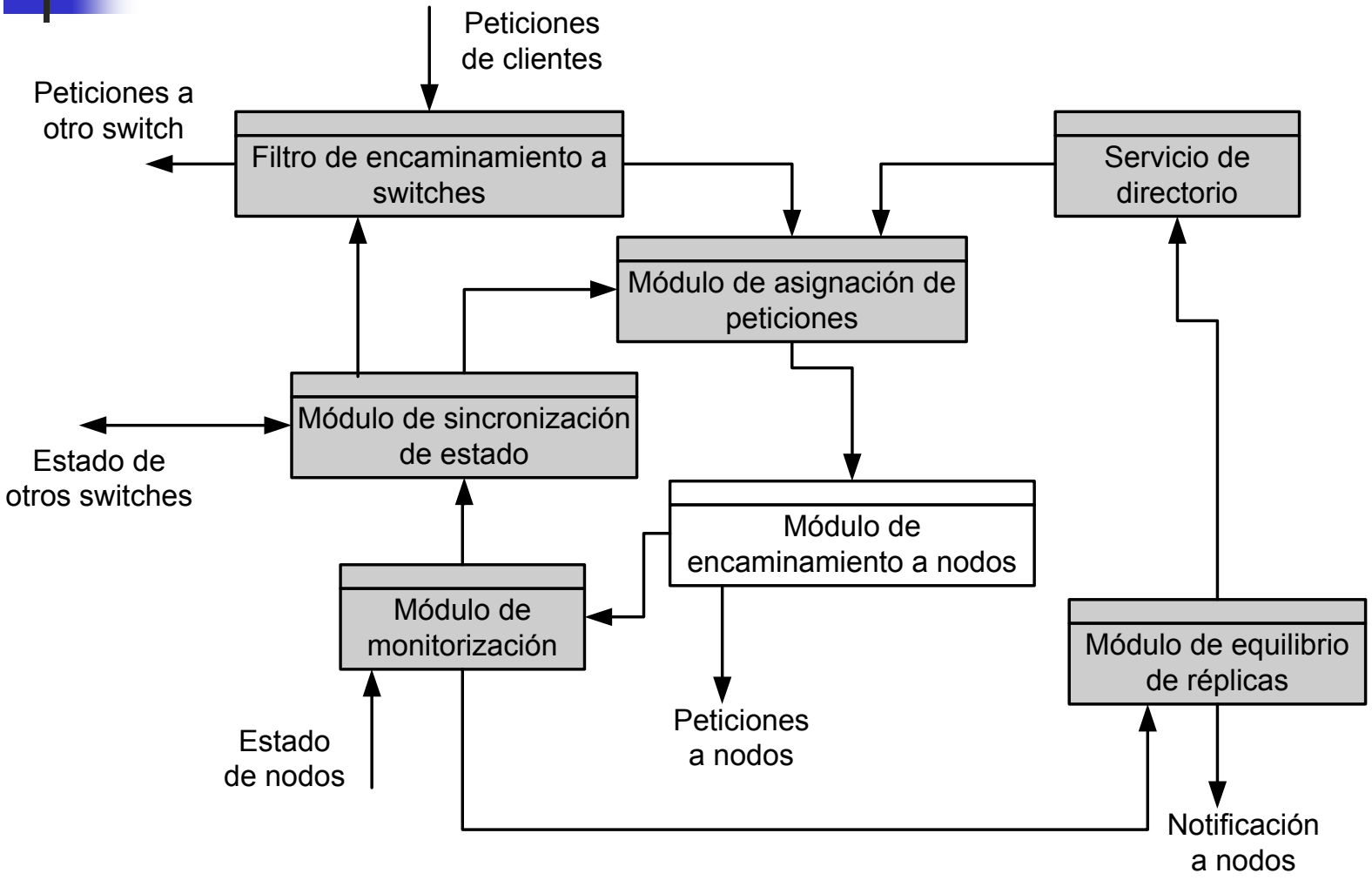
Universidad Carlos III de Madrid
Departamento de Informática
Doctorado en Ingeniería Informática

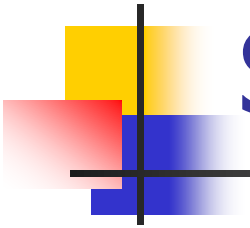
Septiembre de 2005

Autor: José Daniel García Sánchez

Directores: Jesús Carretero Pérez
Félix García Carballeira

Estructura de módulos de los switches

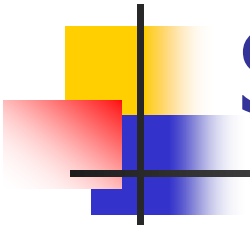




Switch: Funcionalidades (I)

- Filtrado de encaminamiento a switches.
  - Traspaso peticiones a otros switches.
- Asignación de peticiones a nodos servidores.
  - Selección del nodo encargado de servir una petición.
- Encaminamiento a nodos.
  - Envío efectivo de la petición al nodo seleccionado.
- Servicio de directorio.
  - Mantiene información de asignación de réplicas a nodos.

Switch: Funcionalidades (y II)



- Monitorización.
  - Seguimiento el estado de los nodos servidores.
- Sincronización de estado.
  - Intercambio de información de estado entre los switches.
- Equilibrio de réplicas.
  - Determinación de modificaciones en las asignaciones de réplicas.
  - Responsable de iniciar los intercambios de réplicas.

Origen del trabajo

- **Técnicas de aumento de prestaciones en clusters de servidores Web distribuidos y cooperativos.**

  Director: Jesús Carretero

  CAM 07T/0010/2003.
  Junio 2003 a Julio 2004.