Tesi di Dottorato di Ricerca in Informatica ed Ingegneria dell'Automazione (XIII ciclo)

Scalable Web-server Systems

Valeria Cardellini

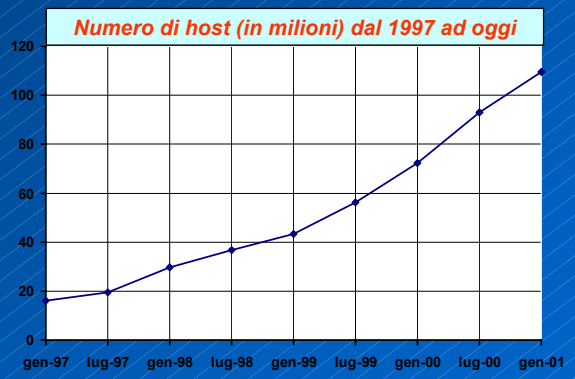
Docenti guida:
Prof. Michele Colajanni
Prof. Salvatore Tucci

Sommario

- Motivazioni e background
- Sistemi di Web server distribuiti localmente
- Sistemi di Web server distribuiti geograficamente
- Condivisione del carico mediante meccanismi basati sul Domain Name System
- Condivisione del carico mediante meccanismi basati sui Web server
- Riduzione del tempo di risposta mediante meccanismi basati sui Web server
- Conclusioni e sviluppi futuri

Motivazione 1: il successo di Internet

	Host collegati
Gennaio 1993	1.313.000
Luglio 1993	1.776.000
Gennaio 1994	2.217.000
Luglio 1994	3.212.000
Gennaio 1995	4.852.000
Luglio 1995	6.642.000
Gennaio 1996	9.472.000
Luglio 1996	12.881.000
Gennaio 1997	16.146.000
Luglio 1997	19.540.000
Gennaio 1998	29.670.000
Luglio 1998	36.739.000
Gennaio 1999	43.230.000
Luglio 1999	56.218.000
Gennaio 2000	72.340.000
Luglio 2000	93.047.000
Gennaio 2001	109.574.000



	Anni impiegati per 50M utenti
Radio	38
Televisione	13
TV via cavo	10
Internet	5

Fonte: Internet Software Consortium (http://www.isc.org)

Motivazione 2: il successo del Web

AltaVista, CNN, Microsoft, Netscape, Yahoo, … (> 50 Milioni hit/day)

Event	Period	Peak day	Peak minute
NCSA server (Oct. 1995)	-	2 Million	-
Olympic Summer Games (Aug. 1996)	192 Million (17 days)	8 Million	-
Nasa Pathfinder (July 1997)	942 Million (14 days)	40 Million	-
Olympic Winter Games (Feb. 1998)	634.7 Million (16 days)	55 Million	110,000
Wimbledon (July 1998)	-	-	145,000
FIFA World Cup (July 1998)	1,350 Million (84 days)	73 Million	209,000
Wimbledon (July 1999)	-	125 Million	430,000
Wimbledon (July 2000)	-	282 Million	964,000
Olympic Summer Games (Sept. 2000)	-	875 Million	1,200,000

[Load misurato in hit]

Motivazione 3: i servizi Web

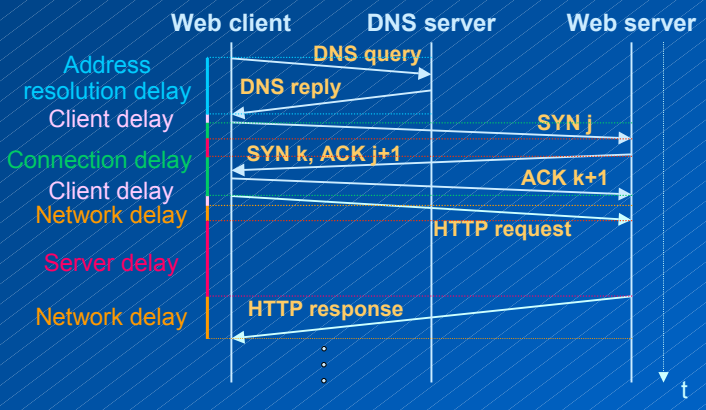
1° generazione

- Un ulteriore canale per informazione non critica
- 95% dell'informazione costituita da testo ed immagini
- Manutenzione e aggiornamenti occasionali
- Prestazioni molto variabili
- Affidabilità non garantita
- Sicurezza non necessaria

2° generazione

- Canale di informazione critica, che sta diventando privilegiato per molti utenti
- Sistema transazionale
- Contenuti dinamici ed attivi in continuo aumento
- Streaming audio e video
- Servizi a pagamento (diretto o indiretto)
- "Vetrina" importante per industrie e organizzazioni

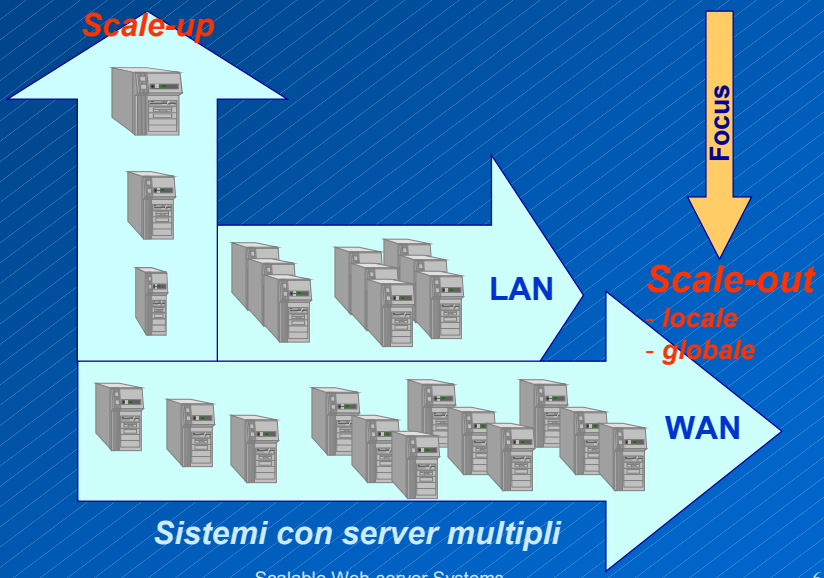
Componenti del ritardo Web



Dov'è il collo di bottiglia?

(1) DNS? (2) Client/connessione?
(3) Rete? (4) Web server?

Ottimizzazioni dal lato server



Sistemi di Web server

Sistemi Web scalabili basati su piattaforme con server multipli

- Un **meccanismo di routing** per indirizzare le richieste client al Web server "migliore"
- Un **algoritmo di distribuzione** (*dispatching*) per individuare il Web server "migliore"
- Un componente **esecutore** per eseguire l'algoritmo di distribuzione utilizzando il relativo meccanismo di routing

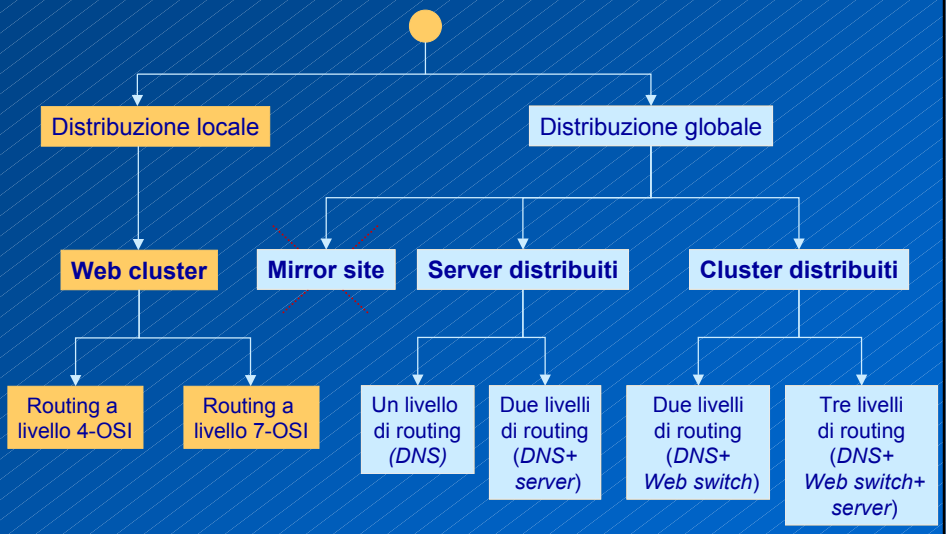
Meccanismi di routing

Meccanismo	Esecutore	Oggetto
Risoluzione hostname - *Dispatching locale* - **Dispatching globale**	DNS	Sessione utente
Ridirezione HTTP - *Dispatching locale* - **Dispatching globale**	Web server	Richiesta risorsa
Ridirezione pacchetti - *Dispatching locale (Web cluster)*	Web switch	Richiesta oggetto/risorsa *(HTTP/1.0 - HTTP/1.1)*

Algoritmi di distribuzione

- **Statici** (information-less)
- **Dinamici**
  - client state aware
  - server state aware
  - client and server state aware
- **Adattativi**

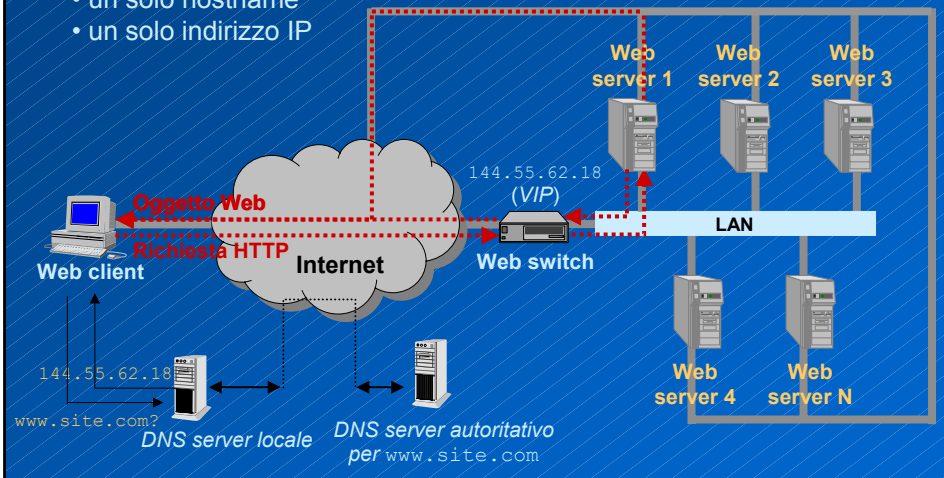
Sistemi con Web server multipli



Web cluster

- Indirizzo del sito Web
  - un solo hostname
  - un solo indirizzo IP

Architettura **one-way**

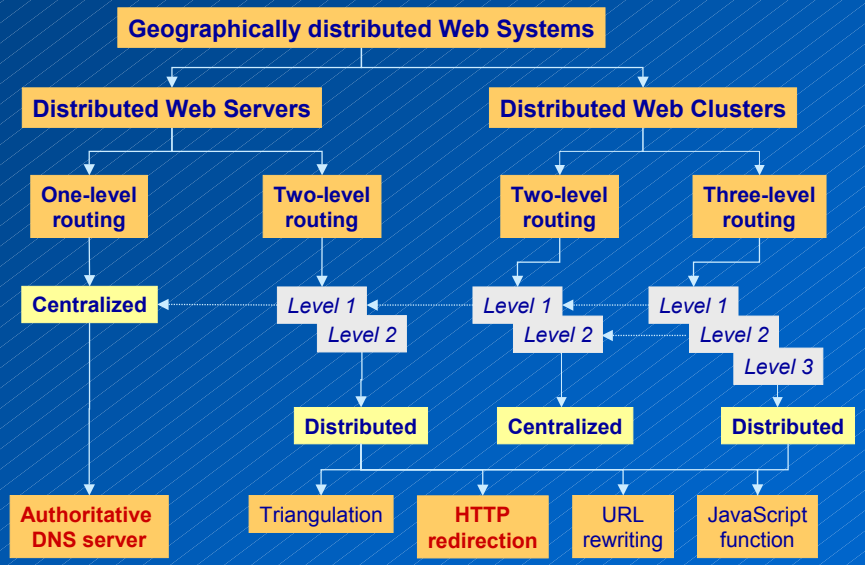


Sistemi Web distribuiti geograficamente

- Problemi dei sistemi Web distribuiti localmente
  - Scalabilità del sistema limitata dal link alla rete del sito Web (collegamento inferiore a 155 Mbps)
  - Incapacità di evitare i link di rete congestionati
  - Affidabilità della rete

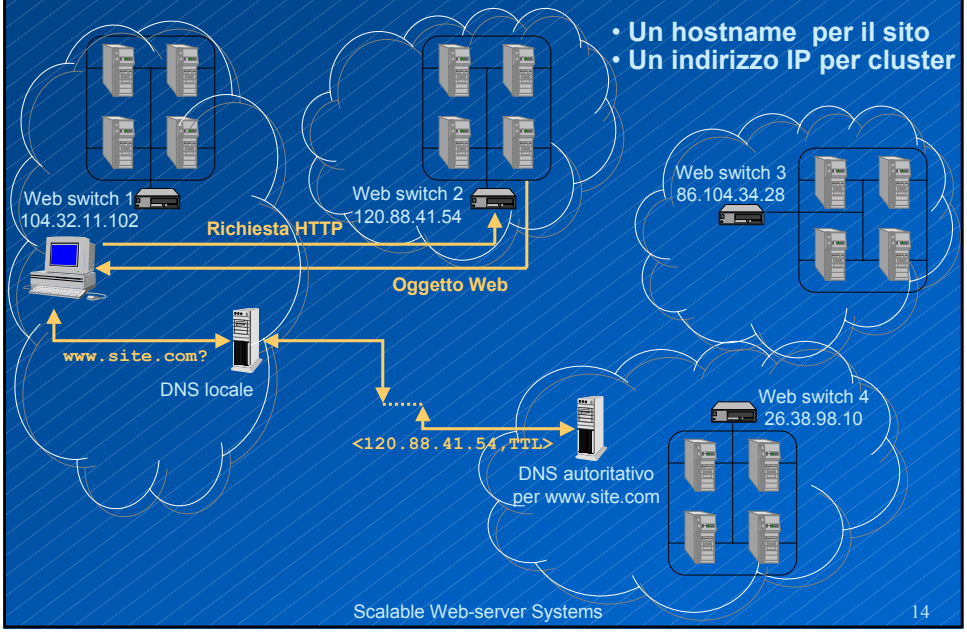
- Scale-out globale
  - Maggiore complessità dell'architettura
    - meccanismi di routing ed algoritmi di distribuzione
  - Metrica per la selezione del server "migliore"
  - Localizzazione dei server

Architetture geograficamente distribuite



Web cluster distribuito

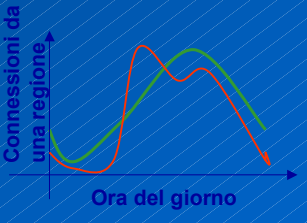
- Un hostname per il sito
- Un indirizzo IP per cluster



Problemi del dispatching geografico

Tipici problemi del dispatching in sistemi Web

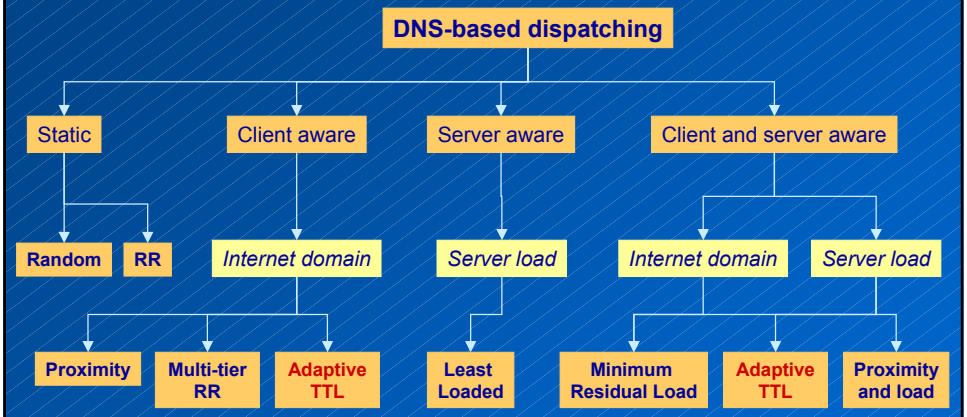
- Picchi di carico in alcune ore/giorni



Problemi aggiuntivi

- Traffico dipendente dai fusi orari
- Distribuzione non uniforme dei client tra le regioni Internet
- Prossimità Internet tra client e server
- Per DNS: Caching del mapping hostname-indirizzo IP in name server intermedi per l'intervallo definito dal *Time-To-Live* (TTL) → controllo limitato sulla distribuzione (5% in siti Web molto popolari)

Algoritmi di dispatching per DNS



Azioni sul Time-To-Live

- **TTL** costante
  - Un solo parametro di controllo (indirizzo IP)
  - Stesso valore del TTL assegnato dal DNS autoritativo per tutte le richieste di indirizzo
    - Es.: Round-Robin (RR), Two-tier Round-Robin (RR2), Minimum Residual Load (MRL)
- **TTL adattativo**
  - *Due* parametri di controllo (indirizzo IP, TTL)
    - indirizzo IP selezionato in base ad una politica a TTL costante
  - Valore del TTL adattato *dinamicamente* dal DNS autoritativo per ogni richiesta di indirizzo considerando il dominio Internet del client e/o la capacità dei Web server

Algoritmi a TTL adattativo

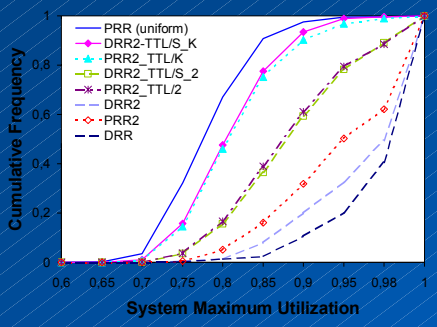
- Algoritmi probabilistici
  - eterogeneità del sistema mediante selezione del server
  - TTL inversamente proporzionale al tasso di carico del dominio del client

$$TTL_j(t) = \frac{\lambda_{\max}(t)\eta_p}{\lambda_j(t)}$$

- Algoritmi deterministici
  - eterogeneità del sistema mediante selezione del TTL
  - TTL inversamente proporzionale al tasso di carico del dominio del client e direttamente proporzionale alla capacità del server

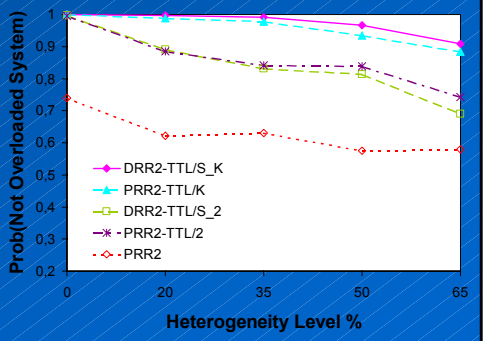
$$TTL_{ij}(t) = \frac{\lambda_{\max}(t)\eta_p\xi_i}{\lambda_j(t)}$$

Risultati simulativi



Sensibilità al variare dell'eterogeneità del sistema

Algoritmi deterministici e probabilistici (livello di eterogeneità pari a 20%)



Sommario dei risultati

- Inadeguatezza degli algoritmi a TTL costante
  - distribuzione non uniforme dei client tra i domini
  - eterogeneità del sistema Web
- Robustezza degli algoritmi a TTL adattativo rispetto a:
  - livelli crescenti di eterogeneità del sistema Web
  - contenuto dinamico
  - distribuzione dei client tra i domini Internet
  - presenza di name server non cooperativi
  - errori nella stima del tasso di carico dei domini
- Limitare l'eterogeneità del sistema
  - non eccedere il livello di eterogeneità pari al 50%

Due livelli di assegnamento

- Aumentare il controllo sulla distribuzione delle richieste
  - DNS: limitatezza del controllo
  - DNS: granularità dell'assegnamento a livello di sessione
  - Server: granularità dell'assegnamento a livello di risorsa/oggetto Web
- Impossibilità di spostare le richieste già assegnate ad un server durante il periodo definito dal TTL
  - Reazione lenta ad un server sovraccarico

Ridirezione HTTP

- Il meccanismo di ridirezione è parte del protocollo HTTP ed è supportato dagli attuali client
- **DNS**: politiche di dispatching *centralizzate*
- **Ridirezione**: politiche di dispatching *distribuite*, in cui tutti i server Web possono partecipare al (ri-)assegnamento delle richieste
- La ridirezione è completamente trasparente per l'utente (non per il client)

message header
HTTP status code
302 - "Moved temporarily" to a new location

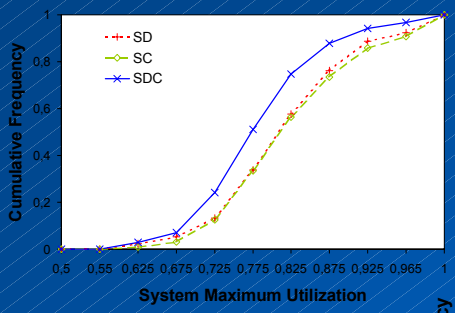
- "New location"
  - ridirezione ad un indirizzo IP (**migliori prestazioni**)
  - ridirezione ad un hostname

Algoritmi di ridirezione

Parameter	*Alternatives*		
Activation trigger (*when*)	Synchronous (*periodic*)		Asynchronous (*on Web server demand*)
Activation decision (*where*)	Centralized (*DNS*)		Distributed (*Web servers*)
Status information	Server load (*CPU queue and/or utilization*)	Alarm	Domain load (*domain hit rate*)
Server selection (*how*)	Assignment Table	Assignment Table and Server Percentage List	Available Server List
Redirected entities (*what*)	Domains (D)	Clients (C)	Domains and Clients (DC)

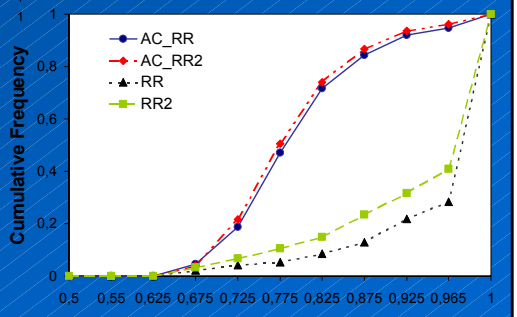
- Algoritmi sincroni
  - ridirezione dell'intero dominio
  - ridirezione dei singoli client
  - ridirezione dell'intero dominio e dei singoli client (**migliore prestazione**)
- Algoritmi asincroni
  - ridirezione dei singoli client

Risultati simulativi

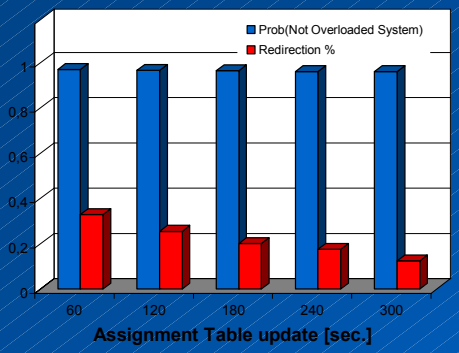


Prestazioni dei migliori algoritmi **sincroni**

Prestazioni degli algoritmi **asincroni**

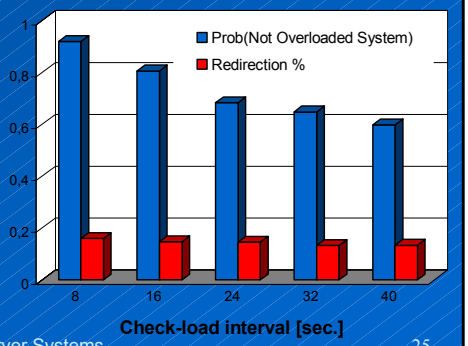


Risultati simulativi (2)



Sensibilità delle prestazioni e della percentuale di richieste ridirette per l'algoritmo **SDC** alla frequenza di aggiornamento della **Assignment Table**

Sensibilità delle prestazioni e della percentuale di richieste ridirette per l'algoritmo **AC** all'intervallo di **check-load**



Tre livelli di assegnamento

Tre livelli di routing e dispatching:

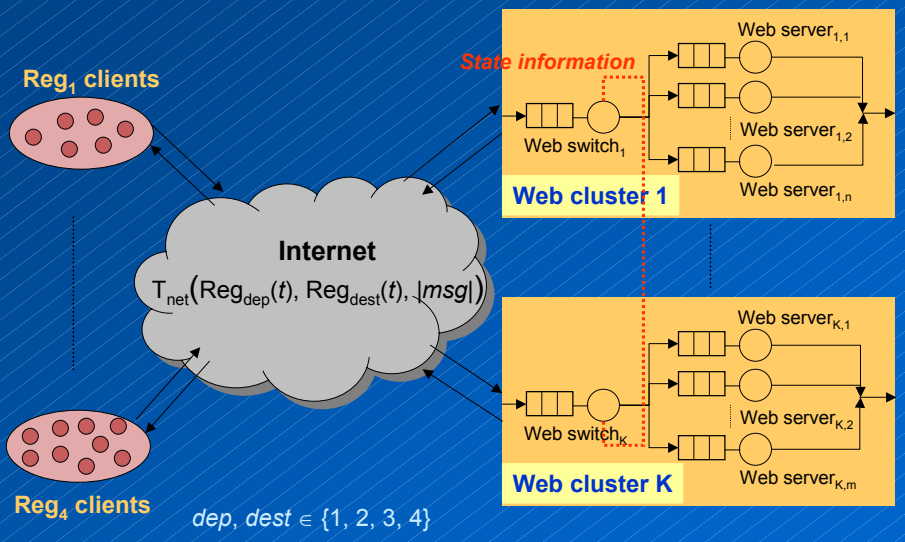
1. **DNS** (ad es., basato sulla prossimità)
2. **Web switch** del cluster: effettua l'assegnamento della richiesta ai server del suo cluster mediante la politica *Weighted Round Robin*
3. **Web server**: effettua la ridirezione (mediante il meccanismo di ridirezione HTTP) verso un altro switch per risolvere situazioni temporanee di sovraccarico

Algoritmi di ridirezione

- Tre componenti dell'algoritmo di ridirezione
  - attivazione
    - on server demand (superamento della soglia di carico)
  - selezione della richiesta da ridirigere
    - tutte le richieste, alcune richieste (dimensione, contenuto, …)
  - individuazione del server a cui ridirigere la richiesta
    - informazione sullo stato del sistema

	Name	*System information*
Selection policies	R-all	None
	R-size	Page size
	R-num	Page hit number
	R-dyn	Page content
Location policies	RR	None
	Load	Cluster load
	CluProx	Cluster-cluster network proximity
	Prox	Cluster-client network proximity

Modello del sistema



Modello della rete

$T_{net}(Reg_{dep}(t), Reg_{dest}(t), |msg|)$?

HTTP/1.1

$$T_{ij,n}=2rtt_{ij}+\sum_{k=1,n}(|request_k|/ab_{ij}(t) + (|response_k|/ab_{ji}(t))$$

dove:

$$ab_{ij}(t) = \pi_{ij}(t)\ bb_{ji}$$

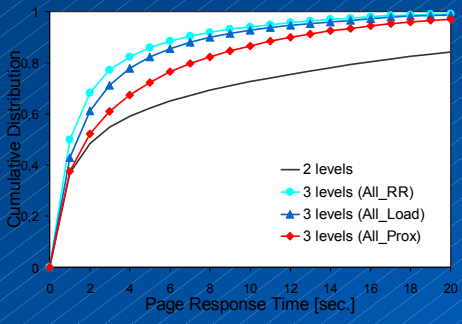
La banda disponibile tra le Regioni i e j all'istante t è pari alla frazione di banda di base a causa del traffico sulla rete.

$$\pi_{ij}(t) = 0.5\ \pi_i^Z(t)+0.5\ \pi_j^Z(t)$$

Fattore di "fortuna"

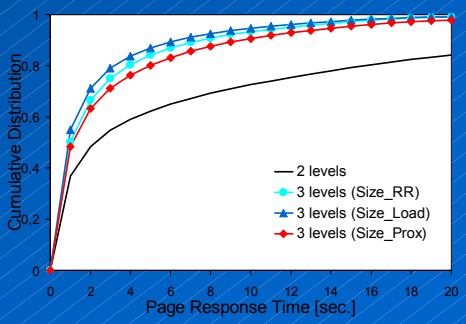
Il traffico sulla rete dipende in modo uguale dalla popolarità della Regione i e dalla popolarità della Regione j all'istante t.

Risultati simulativi

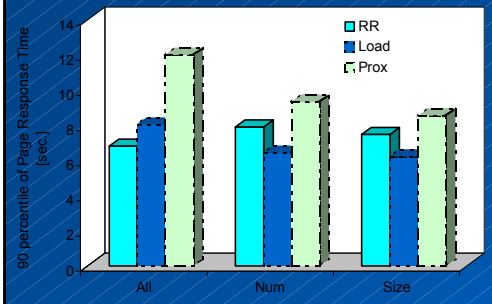


2 livelli di dispatching vs.
3 livelli di dispatching
(*Redirect **All***)

2 livelli di dispatching vs.
3 livelli di dispatching
(*Redirect **Some***)

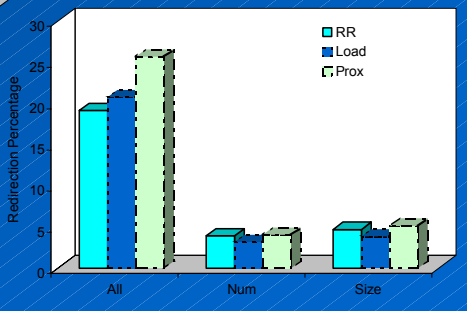


Risultati simulativi (2)



Tempo di risposta
3 livelli di dispatching (*Redirect **All***)
vs.
3 livelli di dispatching (*Redirect **Some***)

Percentuale di ridirezione
3 livelli di dispatching (*Redirect **All***)
vs.
3 livelli di dispatching (*Redirect **Some***)



Conclusioni

- Analisi delle architetture, dei meccanismi di routing e degli algoritmi di dispatching per sistemi Web distribuiti localmente e geograficamente
- Proposta di algoritmi di dispatching di primo livello basati sul DNS
  - condivisione del carico
- Proposta di algoritmi di ridirezione di secondo livello basati sui server
  - condivisione del carico
- Proposta di algoritmi di ridirezione di terzo livello basati sui server
  - riduzione del tempo di risposta
  - aumento della qualità del servizio

Sviluppi futuri

- Distribuzione geografica di servizi Web dinamici e sicuri
- Estensione delle architetture distribuite per supportare l'accesso universale al contenuto Web
  - eterogeneità dei dispositivi client usati per accedere ad Internet
  - adattamento (*transcoding*) del contenuto
- Supporto dal lato server per fornire qualità del servizio Web end-to-end
- Strumenti per valutare le prestazioni di sistemi Web distribuiti geograficamente

Pubblicazioni

[T1] M. Colajanni, P.S. Yu, V. Cardellini, “Scalable Web server systems: architectures, models, and load balancing algorithms", *ACM Sigmetrics 2000*, Santa Clara, CA, June 2000.

[RI1] V. Cardellini, M. Colajanni, P.S. Yu, “Dynamic load balancing on Web-server systems”, *IEEE Internet Computing*, Vol. 3, No. 3, pp. 28-39, May-June 1999.

[RI2] V. Cardellini, M. Colajanni, P.S. Yu, “DNS dispatching algorithms with state estimators for scalable Web-server clusters”, *World Wide Web*, Baltzer Science, Vol. 2, No. 3, pp. 101-113, Aug. 1999.

[RI3] E. Casalicchio, V. Cardellini, M. Colajanni, "Content-aware dispatching algorithms for cluster-based Web servers", *Cluster Computing*, Baltzer Science, to appear in 2001.

[RI4] V. Cardellini, E. Casalicchio, M. Colajanni, M. Mambelli, "Web switch support for differentiated services", *ACM Performance Evaluation Reviews*, to appear in 2001.

[M1] V. Cardellini, M. Colajanni, P.S. Yu, “Impact of workload models on evaluating the performance of distributed Web-server systems”, *System performance evaluation: methodologies and applications*, E. Gelenbe ed., CRC Press, pp. 397-417, March 2000.

[CI1] M. Colajanni, P.S. Yu, V. Cardellini, “Dynamic load balancing in geographically distributed heterogeneous Web servers”, *Proc. of IEEE 18th Int'l Conf. on Distributed Computing Systems*, Amsterdam, The Netherlands, pp. 295-302, May 1998.

Pubblicazioni (2)

[CI2] V. Cardellini, M. Colajanni, P.S. Yu, "Efficient state estimators for load control policies in scalable Web server clusters”, *Proc. of IEEE 22th Int'l Computer Software and Application Conf.*, Vienna, Austria, pp. 449-455, Aug. 1998.

[CI3] V. Cardellini, M. Colajanni, P.S. Yu, "High performance Web-server systems", *Proc. of 13th Int'l Symp. on Computer and Information Sciences*, Ankara, pp. 288-293, Oct. 1998.

[CI4] V. Cardellini, M. Colajanni, P.S. Yu, "Redirection algorithms for load sharing in distributed Web-server Systems", *Proc. of IEEE 19th Int'l Conf. on Distributed Computing Systems*, Austin, TX, pp. 528-535, June 1999.

[CI5] V. Cardellini, M. Colajanni, P.S. Yu, "Redirecting strategies for load sharing policies in distributed Web systems", *Proc. of 14th Int'l Symp. on Computer and Information Sciences*, Kusadasi, Turkey, pp. 91-98, Oct. 1999.

[CI6] V. Cardellini, M. Colajanni, P.S. Yu, "Impact of workload models on evaluating the performance of distributed Web-server systems", *Proc. of IFIP/W.G.-7.3 Performance'99*, Istanbul, Turkey, Oct. 1999.

[CI7] V. Cardellini, E. Casalicchio, M. Colajanni, S. Tucci, "Parallel and distributed architectures for fast and dependable Web services", *Proc. of 7th Int'l Congress on New Technologies and Computing Applications*, Cuba, May 2000.

Pubblicazioni (3)

[CI8] V. Cardellini, M. Colajanni, P.S. Yu, "Geographic load balancing for scalable distributed Web systems", *Proc. of 8th IEEE Int'l Symp. on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, San Francisco, pp. 20-27, Aug. 2000.

[CI9] V. Cardellini, P.S. Yu, Y.-W. Huang, "Collaborative proxy system for distributed Web content transcoding", *Proc. of 9th Int'l ACM Conf. on Information and Knowledge Management*, Washington D.C., pp. 520-527, Nov. 2000.

[CI10] V. Cardellini, E. Casalicchio, M. Colajanni, "Performance evaluation of distributed architectures for the quality of Web services", *Proc. of Hawaii Int'l Conf. on System Sciences* (HICSS-34), Maui, Hawaii, Jan. 2001. IEEE Computer Society.

[CI11] V. Cardellini, E. Casalicchio, M. Colajanni, S. Tucci, “Mechanisms for quality of service in Web clusters", *Proc. of TERENA Networking Conference 2001*, Antalya, Turkey, May 2001.

[CI12] V. Cardellini, E. Casalicchio, M. Colajanni, M. Mambelli, "Web switch support for differentiated services", *Proc. of Performance and Architecture of Web Servers Workshop*, Cambridge, MA, June 2001.

[CN1] V. Cardellini, "Architetture di Web server distribuiti", *Workshop sui Sistemi Distribuiti: Algoritmi, Architetture e Linguaggi* (WSDAAL'98), Pontignano (Siena), Sept. 1998.