Tesi di Laurea

Cluster di server Web a Qualità del Servizio garantita

Candidato:

Marco Orazi

Relatore:

Chiar.mo Prof. Salvatore Tucci

Correlatore:

Chiar.mo Prof. Michele Colajanni

Sommario

- Qualità del Servizio e Qualità dei Servizi Web
- Cluster di server Web: architetture e algoritmi
- Politiche per la Qualità dei Servizi Web
- Architettura del prototipo
- Risultati sperimentali
- Conclusioni e sviluppi futuri

Qualità del Servizio (QoS): principi

- Il sistema Web è composto dal *lato rete* e dal *lato server*
- La QoS nasce nel lato rete
- Partendo dai principi di QoS:

Classificazione dei servizi
Isolamento delle prestazioni
Elevato utilizzo delle risorse
Richiesta di ammissione

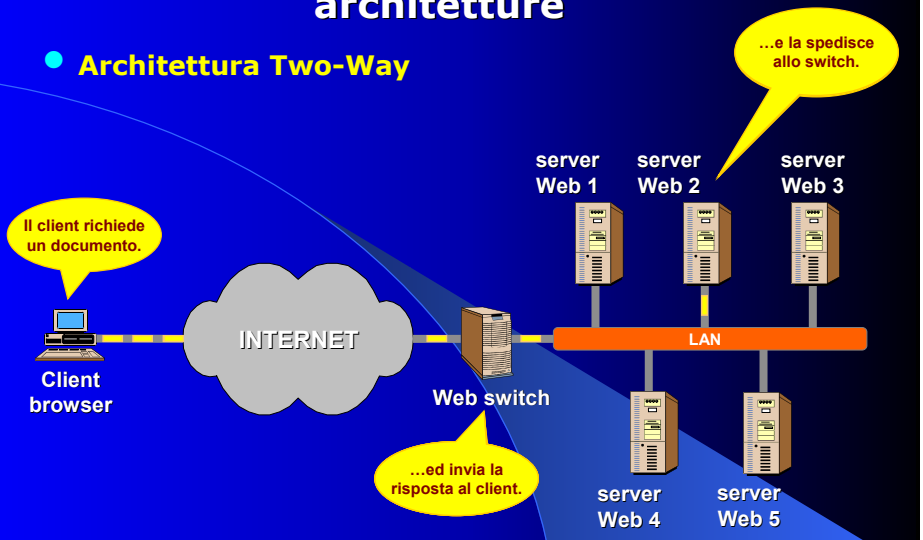
Qualità dei Servizi Web (QoWS)

Qualità dei Servizi Web (QoWS): principi

- **Classificazione**
  - Identificazione utenti e servizi
  - Classificazione utenti e servizi
- **Isolamento delle prestazioni**
  - Politiche di scheduling con priorità
  - Partizionamento delle risorse
- **Elevato utilizzo delle risorse**
  - Partizionamento dinamico delle risorse
- **Richiesta di ammissione**
  - Stima della richiesta di risorse
  - Controllo di accesso

Cluster di server Web: architetture

- **Architettura Two-Way**



- **Architettura One-Way, la risposta viene spedita direttamente al client**

Cluster di server Web: algoritmi

- **Quarto livello OSI**
  - Protocollo TCP/IP
  - Content blind
  - Statici e dinamici

- **Settimo livello OSI**
  - Livello Applicativo (HTTP)
  - Content aware
  - Statici e dinamici

Politiche per la QoWS: *SwitchAdm*

- Le politiche sono classificate in base al numero crescente di principi di QoWS che soddisfano
  - classificazione e richiesta di ammissione
  - aggiunta di isolamento delle prestazioni
  - aggiunta di elevato utilizzo delle risorse
- Classificazione *(High*, *Low)* e richiesta di ammissione

  SwitchAdm

  - servizio negato dal Web switch agli utenti *Low*
  - meccanismo di rifiuto della richiesta sulla base del carico del cluster (soglia di carico)

Politiche per la QoWS: *StaticPart*

- Aggiunta di isolamento delle prestazioni

  StaticPart

  - server Web **staticamente** partizionati in *High Set* (HS) e *Low Set* (LS)
  - richieste di classi diverse vengono assegnate a differenti insiemi
  - servizio negato dal Web switch agli utenti *Low*
  - meccanismo di rifiuto della richiesta sulla base del carico dell' insieme LS (soglia di carico)

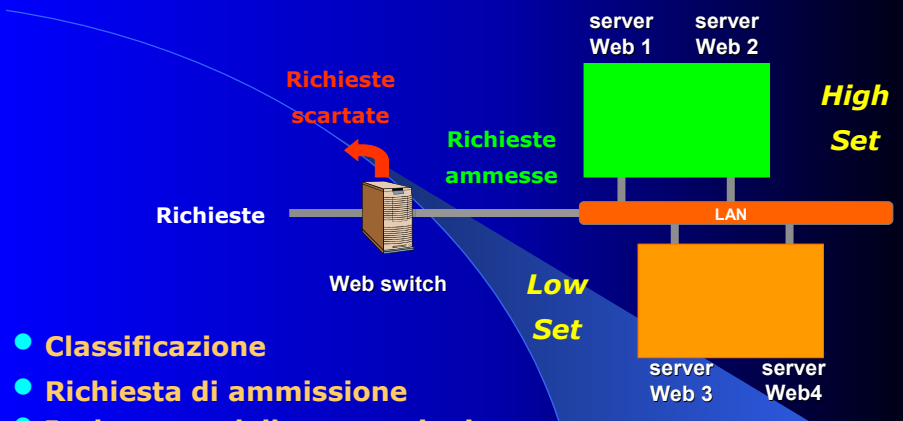
Politiche per la QoWS: *DynamicPart*

- Aggiunta di elevato utilizzo delle risorse

  DynamicPart

  - server Web **dinamicamente** partizionati in *High Set* (HS) e *Low Set* (LS)
  - richieste di classi diverse vengono assegnate a differenti insiemi
  - servizio negato dal Web switch agli utenti *Low*
  - meccanismo di rifiuto della richiesta sulla base del carico dell' insieme LS (soglia di carico)

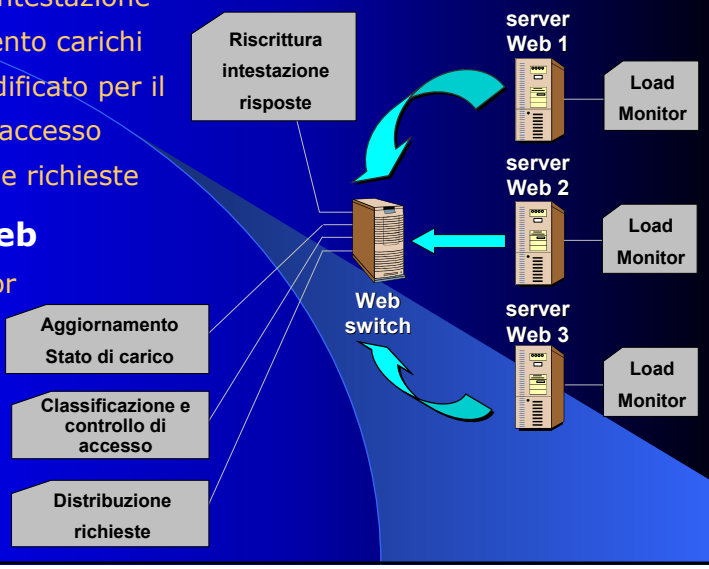
Architettura del prototipo



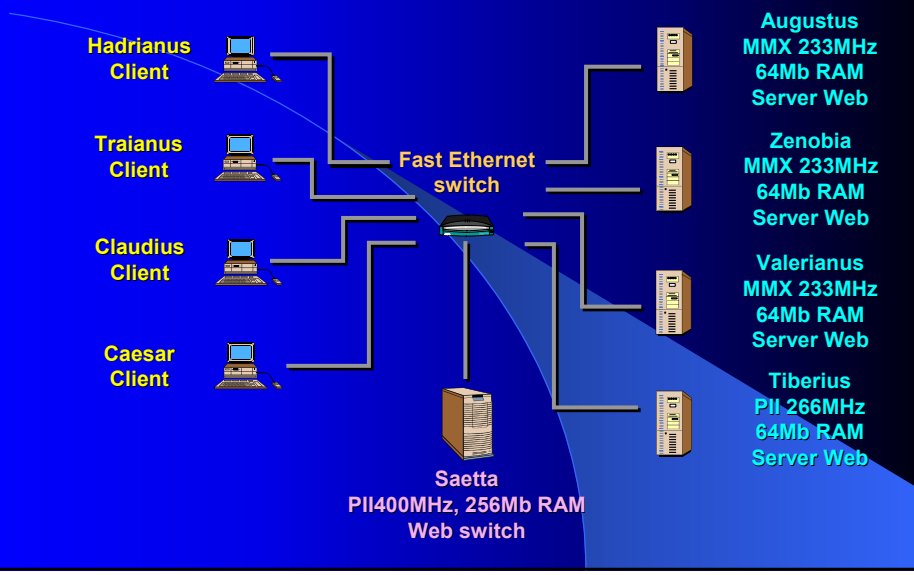
- **Classificazione**
- **Richiesta di ammissione**
- **Isolamento delle prestazioni**
- **Elevato utilizzo delle risorse**

Componenti del prototipo

- **Web switch**
  - Riscrittura intestazione
  - Aggiornamento carichi
  - Apache modificato per il controllo di accesso
  - Distribuzione richieste
- **server Web**
  - Load Monitor



Ambiente di test

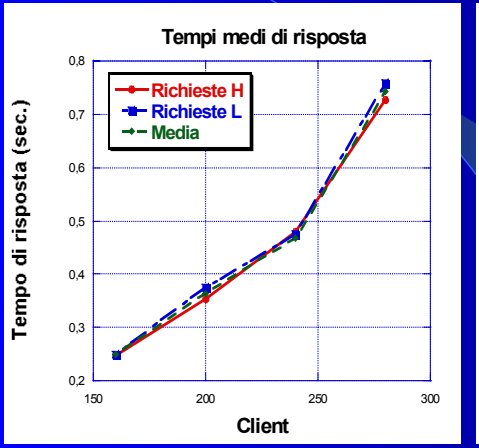


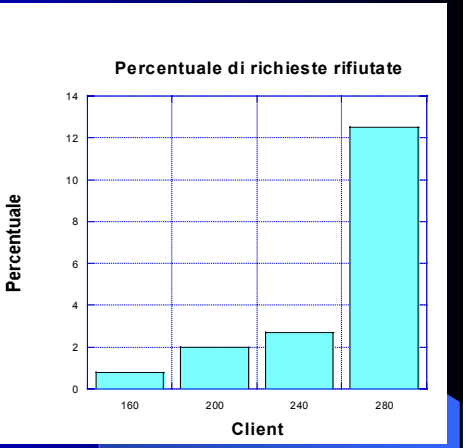
Ambiente di test (2)

- **Webstone: benchmark per server Web**
- **Workload**
  - Richieste statiche e dinamiche
  - Percentuale di richieste di classe *High* 20%
  - Percentuale di richieste di classe *High* 40%
- **Misurazioni effettuate**
  - Utilizzazioni medie CPU server e Web switch
  - Connessioni / sec. aperte dal cluster
  - Throughput medio in uscita dal cluster
  - **Tempi medi di risposta per ogni classe di servizio**
  - **Percentuale di richieste *Low* rifiutate**

Risultati sperimentali

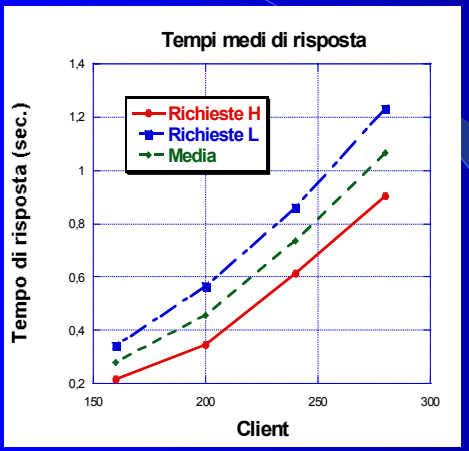
- ***SwitchAdm*, percentuale richieste *High* 20%**
  - Tempi di risposta non differenziati
  - Percentuale di rifiuto limitata fino a 240 client

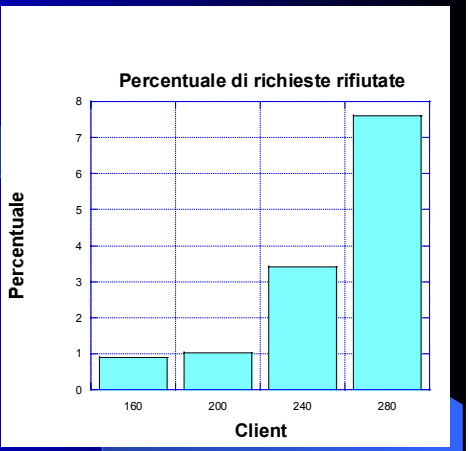




Risultati sperimentali (2)

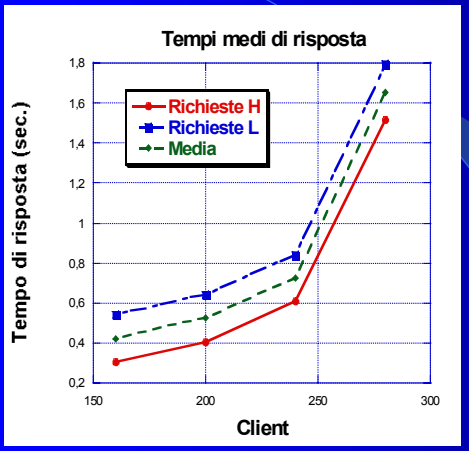
- ***StaticPart*, percentuale richieste *High* 20%**
  - Tempo di risposta utenti *High* inferiore
  - Percentuale di scarto limitata fino a 200 client

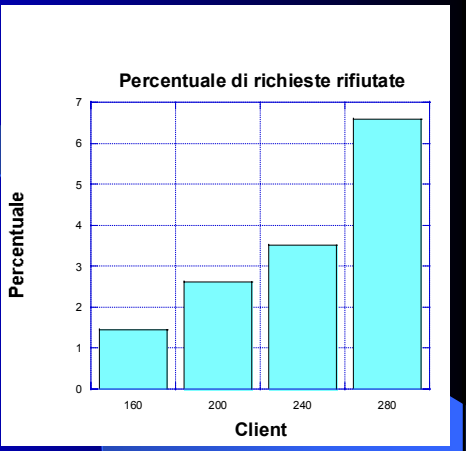




Risultati sperimentali (3)

- ***DynamicPart*, percentuale richieste *High* 40%**
  - Tempo di risposta utenti *High* inferiore
  - Percentuale di scarto crescente in modo uniforme





Conclusioni e sviluppi futuri

- **Conclusioni**
  - Realizzazione di un cluster di server Web a Qualità del Servizio garantita
  - Implementazione di politiche per la QoWS
  - Le politiche con partizionamento dei server Web garantiscono una differenziazione del livello di servizio
- **Sviluppi futuri**
  - Nuove politiche per la QoWS, a livello applicativo e del sistema operativo
  - Nuove metriche per lo stato di carico dei server Web
  - Implementazione di un Web switch One-Way di settimo livello