Load Balancer

Programmable Networks A.Y. 24/25

Obiettivo

- Il contesto è quello di una rete Datacenter nella quale vengono forniti diversi servizi
- Ogni servizio
  - è esposto all'esterno tramite un indirizzo IP
  - è fisicamente espletato da un pool di macchine fisiche
- Gli utenti inviano continuamente richieste di servizio al Datacenter
  - se l'infrastruttura fisica è sotto stress, il livello di qualità percepito dagli utenti sarà basso
- L'obiettivo del lavoro è realizzare un'applicazione SDN per bilanciare il carico di lavoro sui vari server

Modello di riferimento

- La rete Datacenter è di tipo Openflow capable ed è rappresentata dal grafo **G(N,L)**, in cui
  - **N** è il set di switch SDN
  - **L** è il set di link
- Si assume che i link della rete abbiano capacità illimitata
- Il servizio $\sigma_i \in \Sigma$ è rappresentato dalla tupla $\langle IP_i, [s_k]_{k=1,...K}, d_i, t_i \rangle$, in cui
  - IP_i è l'indirizzo IP col quale il servizio σ_i viene esposto verso l'esterno
  - $[s_k]$ è il set di macchine fisiche che erogano il servizio σ_i
  - d_i è la domanda di traffico (in bps) relativa al servizio σ_i
  - t_i è la durata (in secondi) di ciascuna richiesta relativa al servizio σ_i
- Ciascun server s_k ha una capacità di rete b_k
  - il modello non tiene conto di altre risorse fisiche (CPU, memoria)
- Al tempo **t**, il numero di richieste del servizio σ_i è rappresentato da $r_i(t)$
- Al tempo **t**, il numero di richieste del servizio σ_i in esecuzione sul server s_k è rappresentato dal numero $n_{i,k}(t)$

Modello di riferimento

- Il carico del server $\mathbf{s_k}$ è calcolabile come il rapporto tra
  - la domanda complessiva che il server sta gestendo
  - la capacità del server stesso

$$\rho_k(t) = \frac{\sum_{i=1}^{\Sigma} n_{i,k}(t) d_i}{b_k}$$

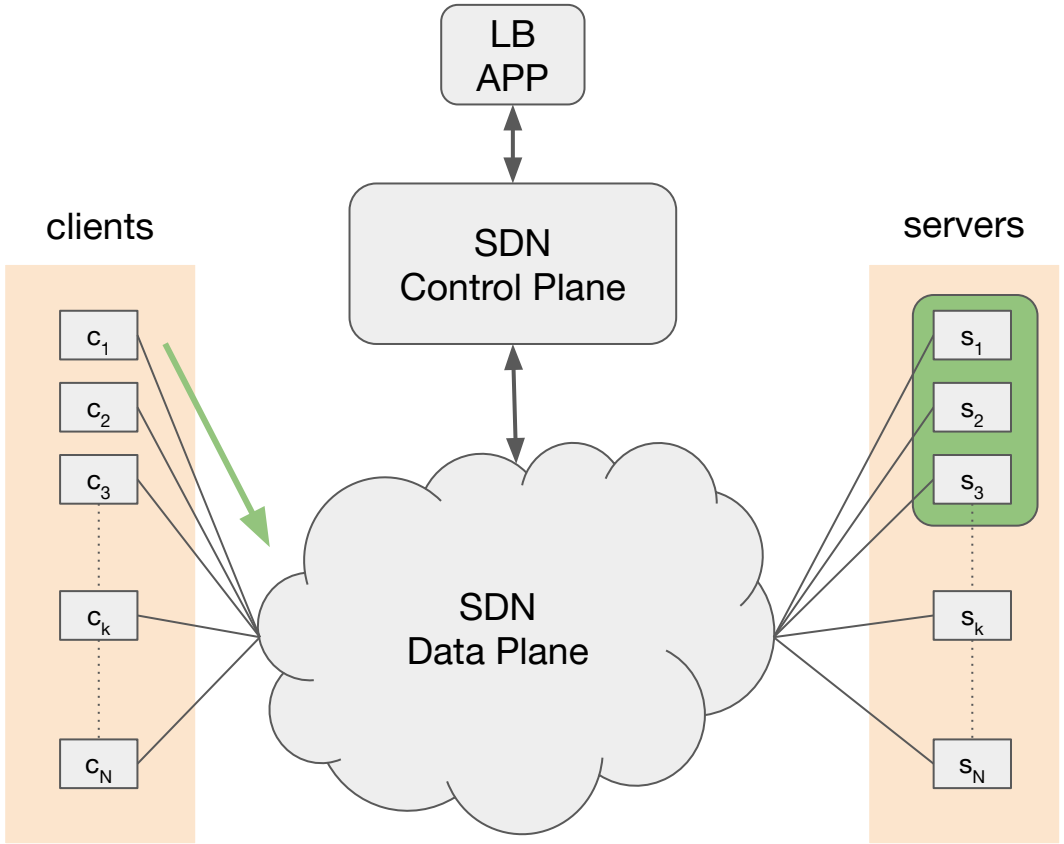
- Indicando con $\boldsymbol{\rho_{max}(t)}$ e $\boldsymbol{\rho_{min}(t)}$ l'utilizzazione del server più carico/meno carico rispettivamente, si ha che l'obiettivo è

$$\min_{\forall t} \rho_{\max}(t) - \rho_{\min}(t)$$

- Il vincolo è che ogni domanda deve essere soddisfatta

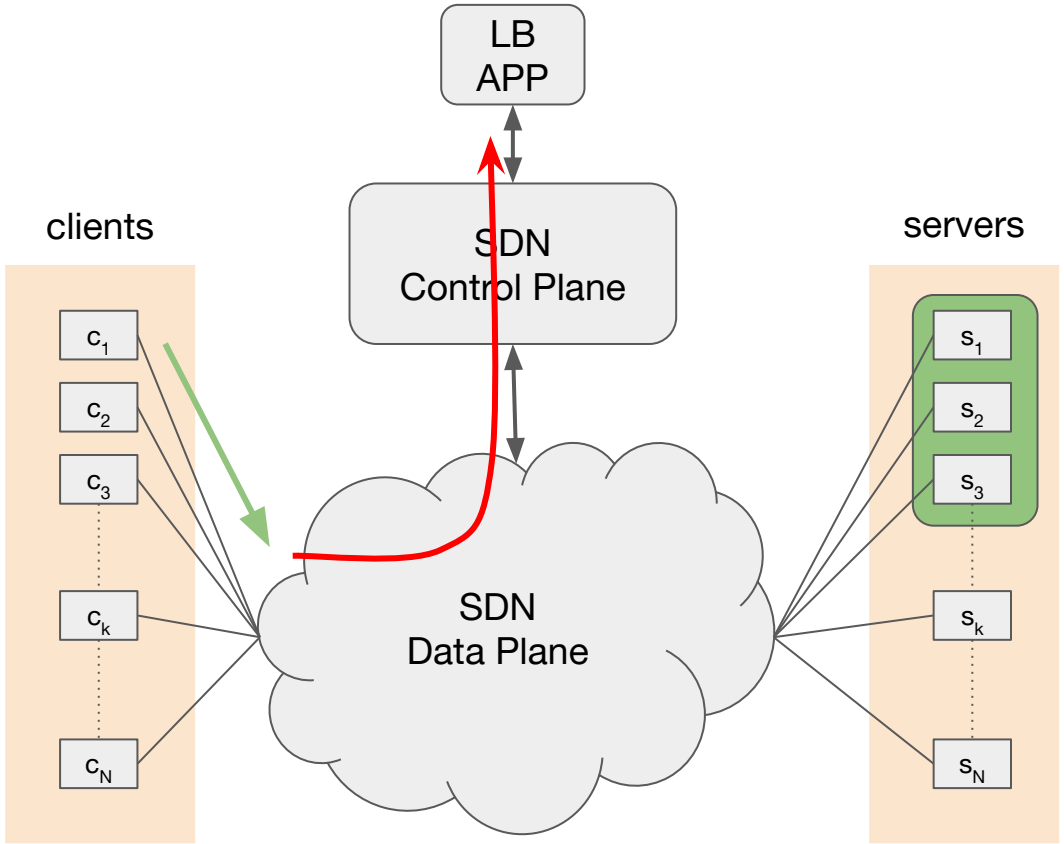
LB App

1. il client c1 genera una richiesta per il servizio verde



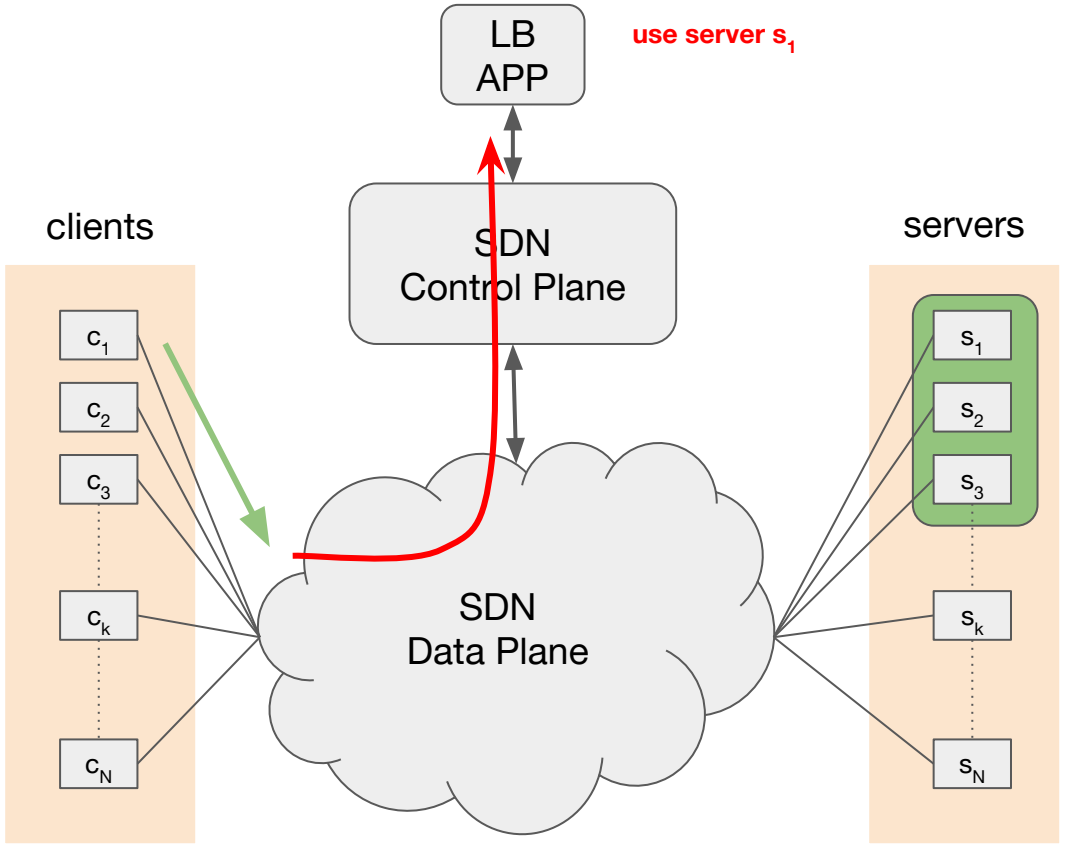
LB App

1. il client c1 genera una richiesta per il servizio verde
2. tramite un packetIN la richiesta viene inoltrata alla LB App



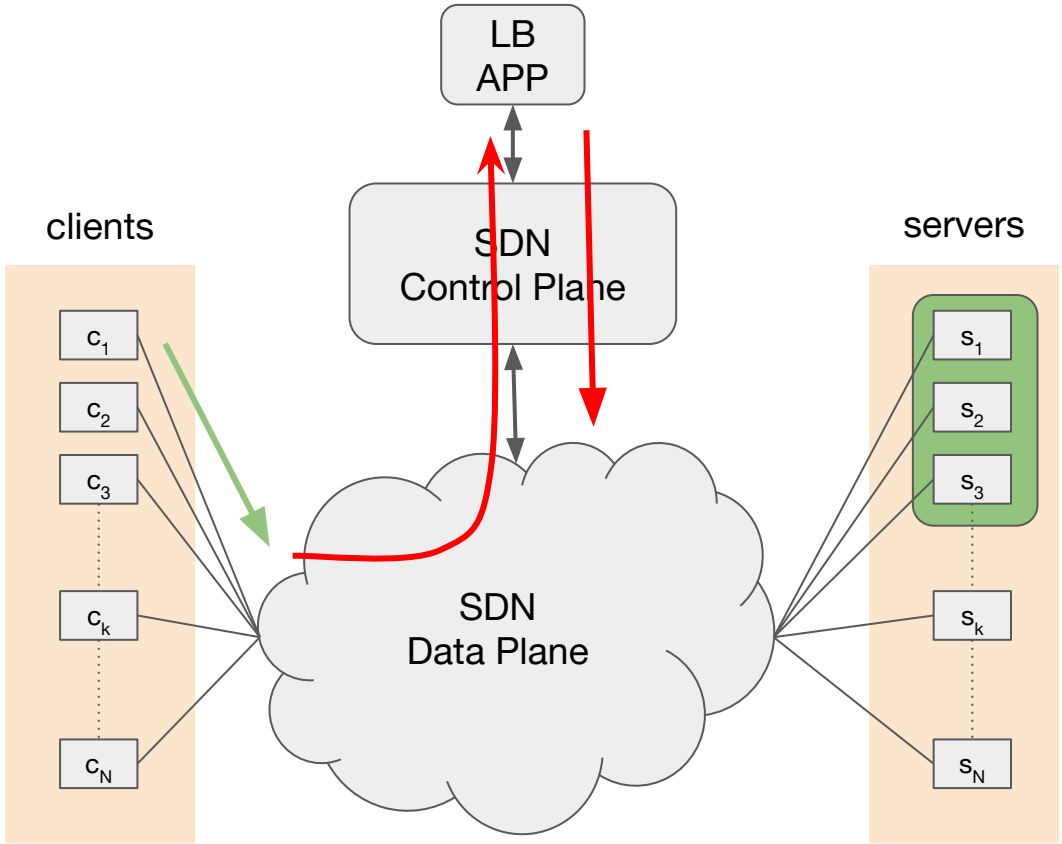
LB App

1. il client c1 genera una richiesta per il servizio verde
2. tramite un packetIN la richiesta viene inoltrata alla LB App
3. l'applicazione decide il server a cui assegnare la richiesta



LB App

1. il client c1 genera una richiesta per il servizio verde
2. tramite un packetIN la richiesta viene inoltrata alla LB App
3. l'applicazione decide il server a cui assegnare la richiesta
4. il percorso viene configurato



LB App

1. il client c1 genera una richiesta per il servizio verde
2. tramite un packetIN la richiesta viene inoltrata alla LB App
3. l'applicazione decide il server a cui assegnare la richiesta
4. il percorso viene configurato
5. la richiesta viene espletata