

Sistemi e Applicazioni Cloud

Appello del 9 settembre 2025 [Tempo consegna: 2h 30m]

Parte 1: rete base

Si usa un simulatore per studiare il comportamento di un sistema in grado di parallelizzare il traffico su diversi nodi.

Il sistema è mostrato nella figura.

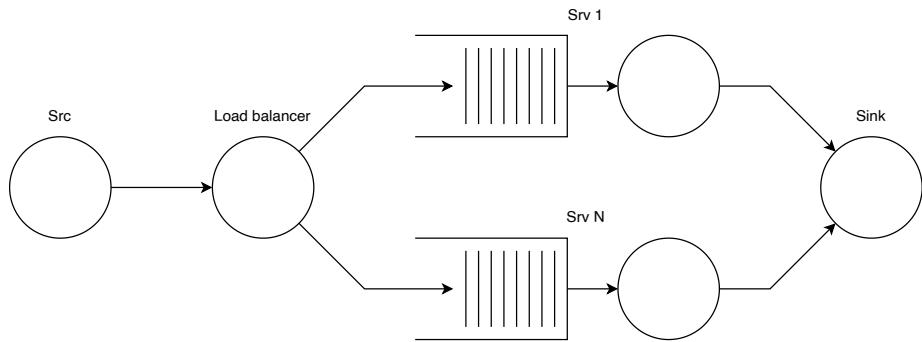


Figure 1: Modello di rete

Il carico in ingresso è $\lambda = 200$ richieste al secondo e viene ripartito equamente tra gli N server (politica *round-robin* o *random* a piacere). Il datacenter cloud mette a disposizione due tipi di server:

- Tipo 1: capacità di servizio $\mu_1 = 8$ richieste/sec, costo = 1.5 \$ per ora
- Tipo 2: capacità di servizio $\mu_1 = 16$ richieste/sec, costo = 3 \$ per ora

Il tempo di servizio segue una distribuzione esponenziale per entrambi i server. Il processo di servizio delle richieste è vincolato ad un SLA sul tempo di risposta medio T_r che deve restare al di sotto di 250 ms.

Testare il tempo di servizio per $N = 40$ indicando anche l'intervallo di confidenza del 65% per ciascun tipo [$\approx 281 \pm 0.5ms$, $\approx 87 \pm 0.08ms$].

N	Tipo Srv	T_r	\pm CI	Costo
45	Tipo 1			
45	Tipo 2			

Parte 2: dimensionare il bilanciamento

Identificare mediante la teoria delle reti di code il valore di N^* tale per cui il requisito di SLA soddisfatto per ciascuno dei due tipi di server

Tipo Srv	N	T_r	Costo
Tipo 1			
Tipo 2			

Indicare quale scelta è più conveniente da un punto di vista economico.

Parte 3: verifica

Eseguire un'analisi del tempo di risposta e del costo per i seguenti range di valori:

- $N \in [30, 35, 40, 45, 50, 55]$ per server di Tipo 1
- $N \in [15, 20, 25, 35, 45, 55]$ per server di Tipo 2

Tipo Srv	N	T_r	\pm CI	Costo
Tipo 1	30			
Tipo 1	35			
Tipo 1	40			
Tipo 1	45			
Tipo 1	50			
Tipo 1	55			
Tipo 2	15			
Tipo 2	20			
Tipo 2	25			
Tipo 2	35			
Tipo 2	45			
Tipo 2	55			

Punto bonus: realizzare plot dei dati sulla base dell'esempio fornito

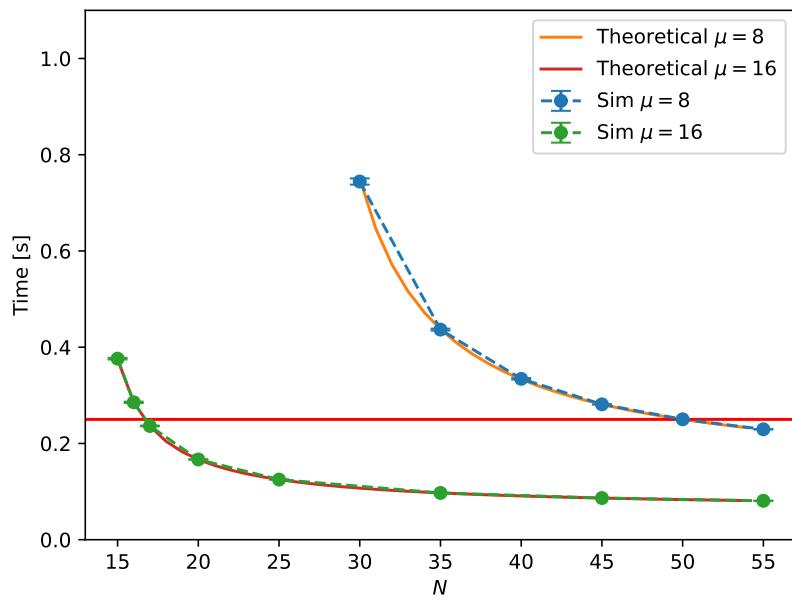


Figure 2: Plot