

# Sistemi e Applicazioni Cloud

Appello del 18 giugno 2025 [Tempo consegna: 2h 30m]

## Parte 1: rete base

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

Il sistema è mostrato nella figura.

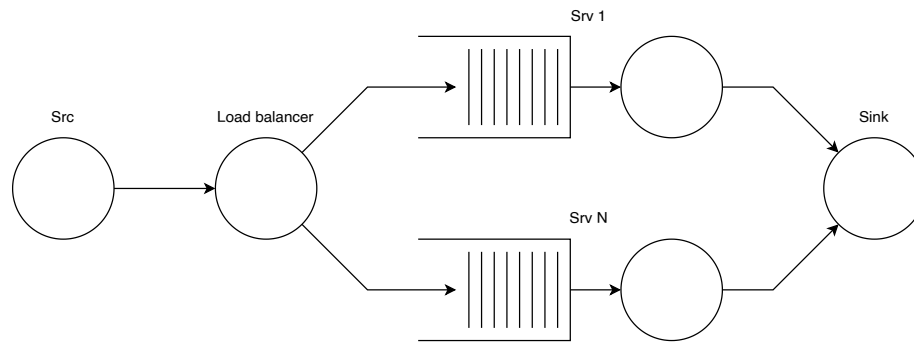


Figure 1: Modello di rete

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 stipi di server:

- Tipo 1: capacità di servizio  $\mu_1 = 10$  richieste/sec, costo = 1.5 \$ per ora
- Tipo 2: capacità di servizio  $\mu_2 = 15$  richieste/sec, costo = 2.5 \$ 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% [ $\approx 200ms$ ] per Tipo 1, [ $\approx 100ms$ ] per Tipo 2.

$N$	Tipo Srv	$T_r$	$\pm$ CI	Costo
40	Tipo 1	0.22023	$\pm 0.00042$	60 \$/hour
40	Tipo 2	0.09996	$\pm 0.000134$	100 \$/hour

## 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



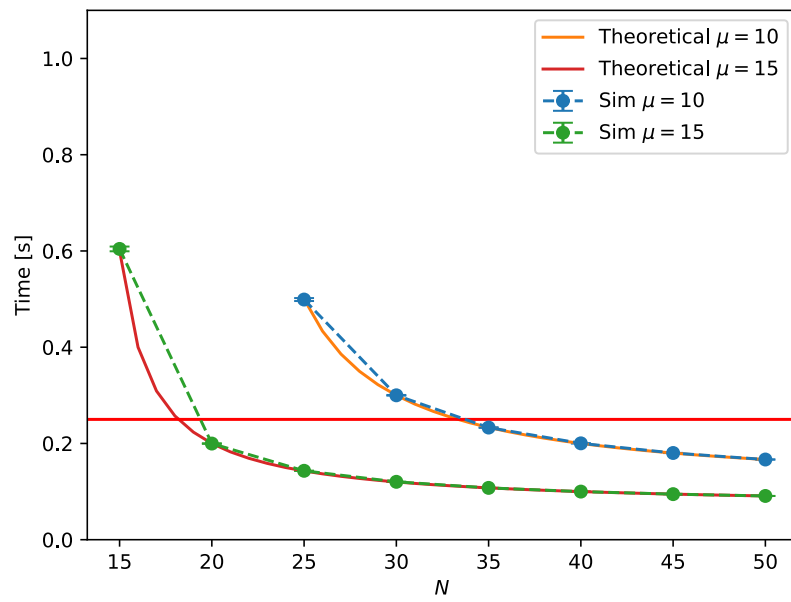


Figure 2: Plot