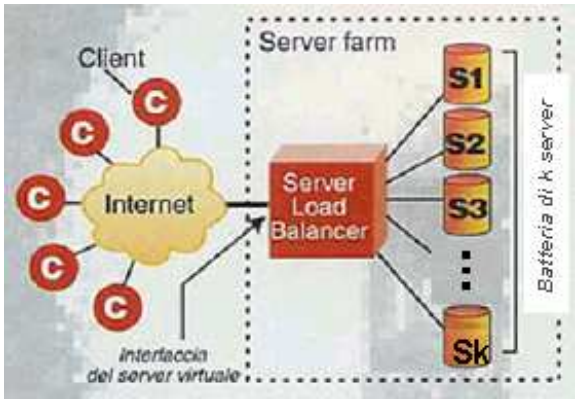
Network Technologies

Un *Server Load Balancer* (SLB) (vedi figura) è utilizzato da un *Content Service Provider* per bilanciare il carico sui k server di una *server farm*. L'algoritmo adottato dal SLB per distribuire le richieste di servizio dei client è il seguente: una nuova richiesta viene inoltrata al server che sta rispondendo al minor numero di richieste e, nel caso in cui più server stiano rispondendo allo stesso ed, al contempo, più basso numero di richieste, la richiesta è inoltrata ad uno qualsiasi tra questi server.

Si supponga:

a) che le richieste di servizio arrivino al SLB secondo un processo indipendente di Poisson a velocità λ richieste/sec;
b) che sia trascurabile l'intervallo temporale tra l'istante in cui una richiesta arriva al SLB e l'istante in cui il server selezionato riceve tale richiesta;
c) che i server rispondano ad una richiesta non appena la ricevono e con tempi di servizio che sono variabili casuali indipendenti e distribuite esponenzialmente a valor medio $1/\mu$ secondi;
d) che un generico server non possa rispondere <u>contemporaneamente</u> a più di M richieste e che sia <u>persa</u> qualsiasi richiesta in arrivo quando la server farm è satura (ciascun server sta rispondendo a M richieste);

Nel caso in cui k=2 e M=2

1) Disegnare il diagramma dei tassi di transizione di una CMTC che modella il sistema e discutere l'ergodicità della catena;
2) Calcolare:
   - la probabilità che una richiesta di servizio non sia persa;
   - il numero medio di richieste a cui un server risponde contemporaneamente;