Traffic Engineering of Anycast Services
(Engenharia de Tráfego de Serviços *Anycast*)

Modelação e Desempenho de Redes e Serviços

Prof. Amaro de Sousa (asou@ua.pt)

DETI-UA, 2025/2026

Unicast vs. Anycast Services

Unicast Service:

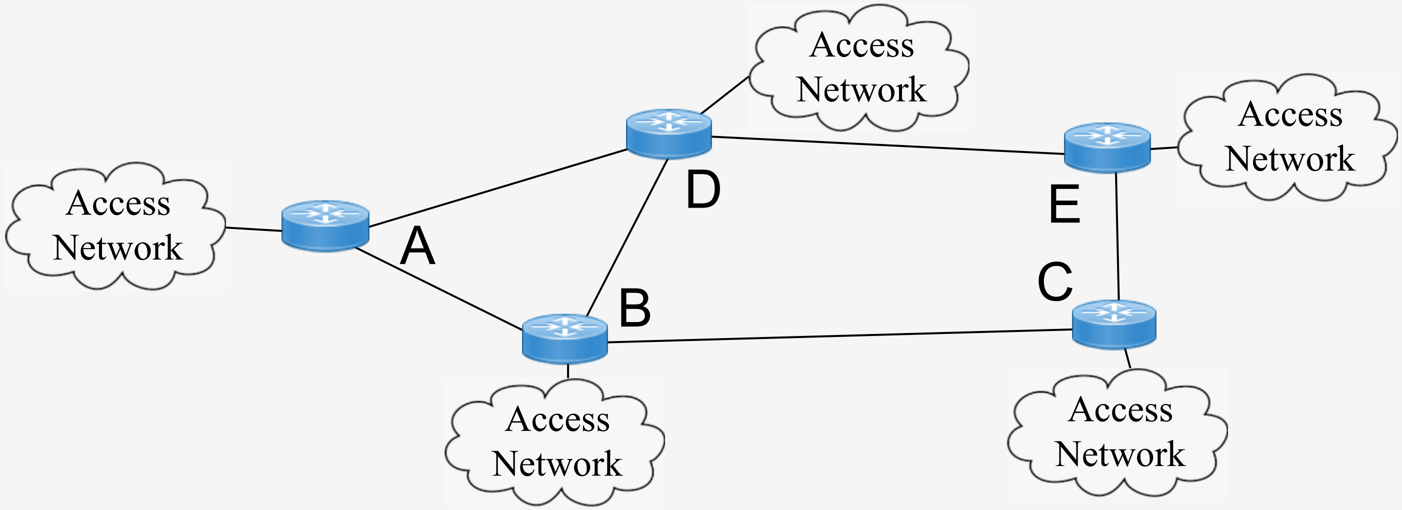
- The traffic of a unicast service is defined by a set of flows, such that each flow is defined with an origin node and a destination node.

Anycast Service:

- Examples: video or music *streaming* services (such as Netflix, Youtube, Amazon Prime Video, Spotify)
- In an anycast service, a set of nodes (named anycast nodes) are associated with the service, i.e., the network nodes where the servers of the anycast service, typically hosted in Data Centres (DC) are connected to.
- The traffic of an anycast service is defined by a set of flows, such that each flow is defined with an origin node.
- The destination node of each traffic flow can be any of the anycast nodes, i.e., it depends on the routing paths set in the network.

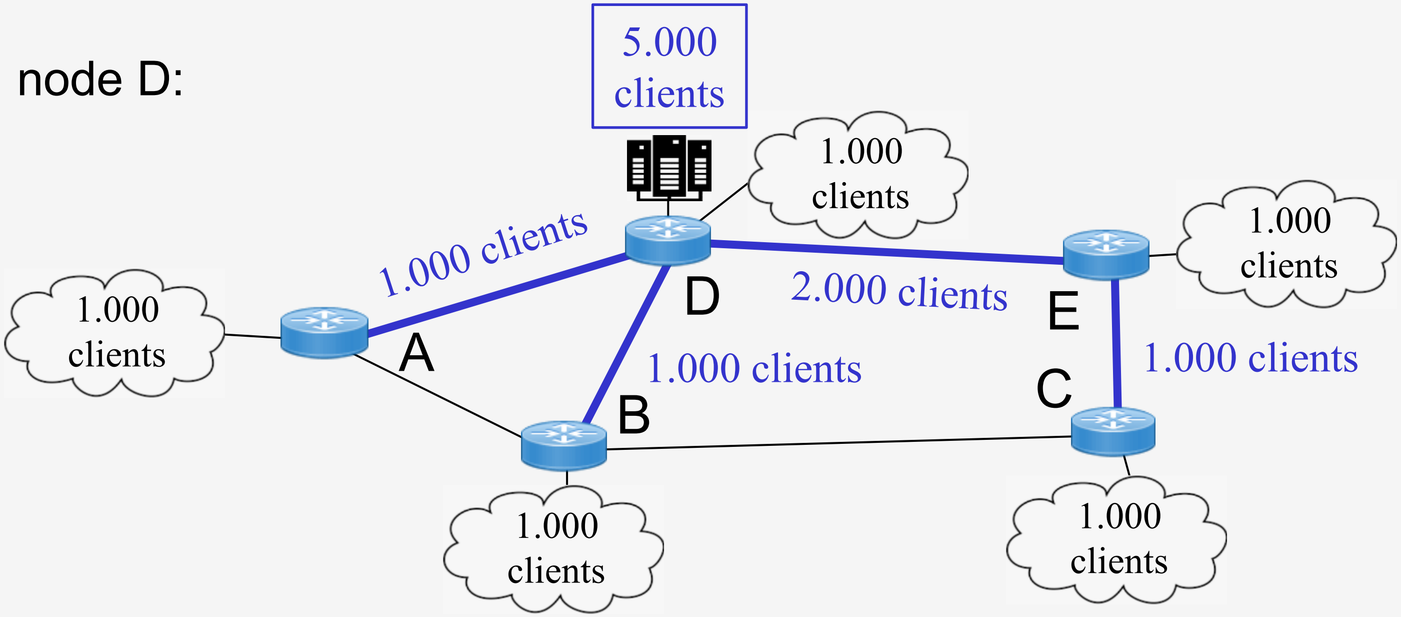
Anycast services

- When a client connects to an anycast service, by default, the network routes the communication towards the closest anycast node (in terms of number of hops or delay).
- The number of anycast nodes and their location in the network influence:
  - the required resource to support the service;
  - the network performance in terms of delay and service availability.
- Consider the following network example supporting an anycast service with 100 clients connected on each access network:



Anycast service illustration

1 DC in node D:

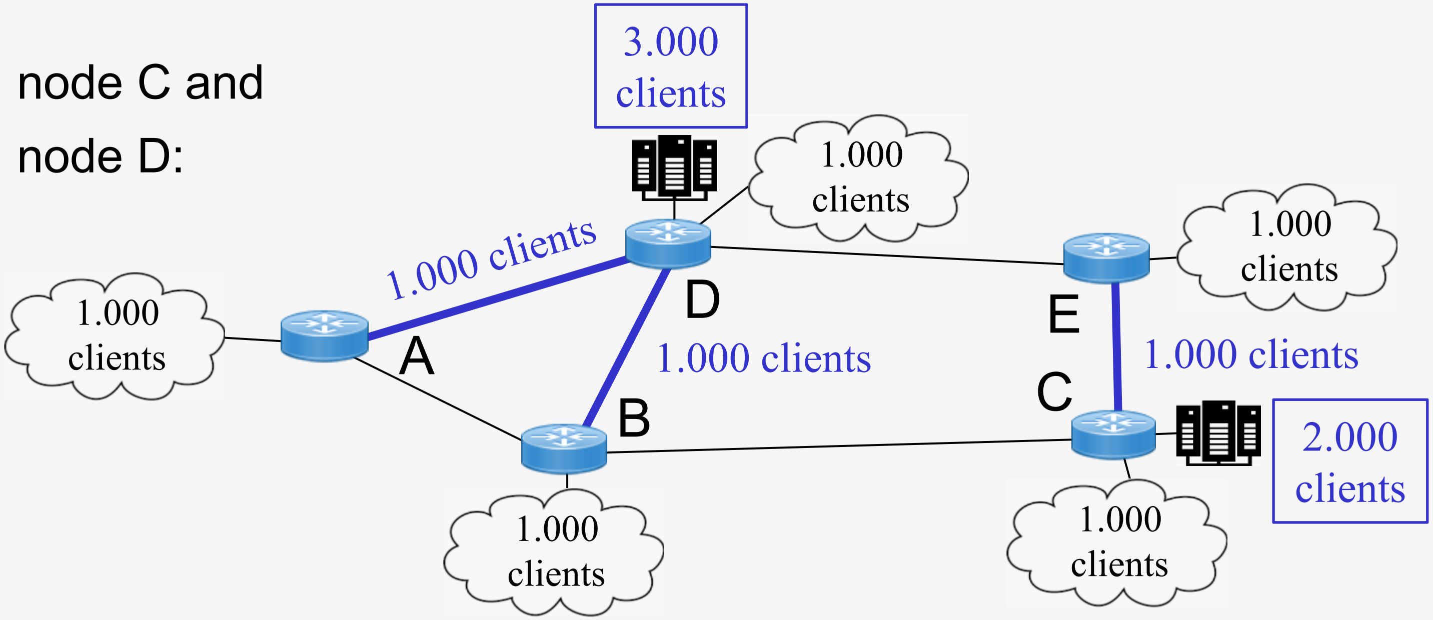


- 3 links must have enough capacity to support the traffic flow generated by 1000 clients and 1 link must have enough capacity to support the traffic flow generated by 2000 clients.
- The Data Centre (DC) must have the capacity to provide the service to 5000 clients.
- If the DC fails, the anycast service fails completely.

Anycast service illustration

1 DC in node C and

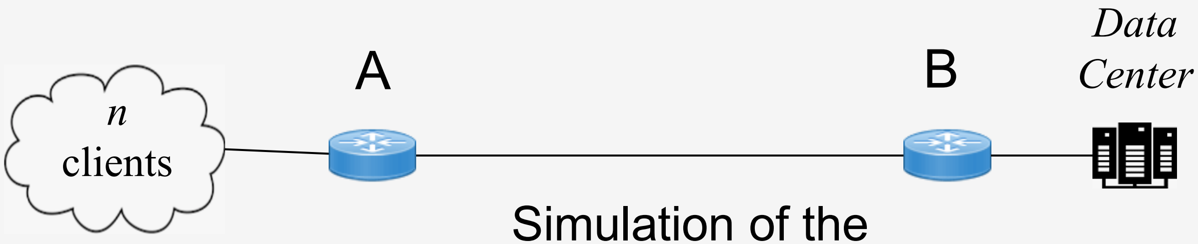
1 DC in node D:



- 3 links must have enough capacity to support the traffic flow generated by 1000 clients (it requires less link resources than the previous case).
- One DC must have the capacity to provide the service to 3000 clients and the other to 2000 clients.
- If one of the DCs fail, all flows start being routed towards the other DC: there is a service degradation but the service does not completely fail.

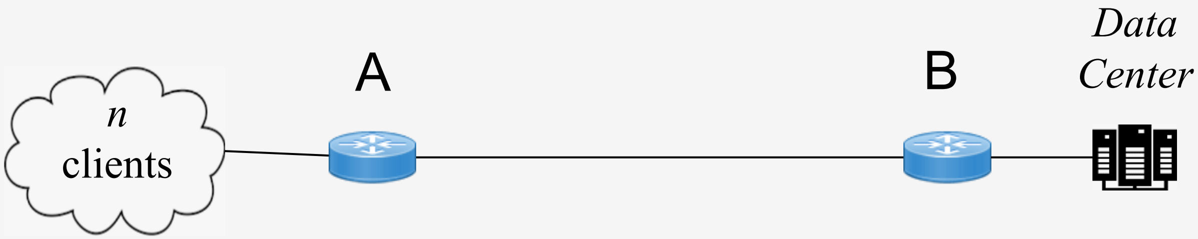
Traffic aggregation

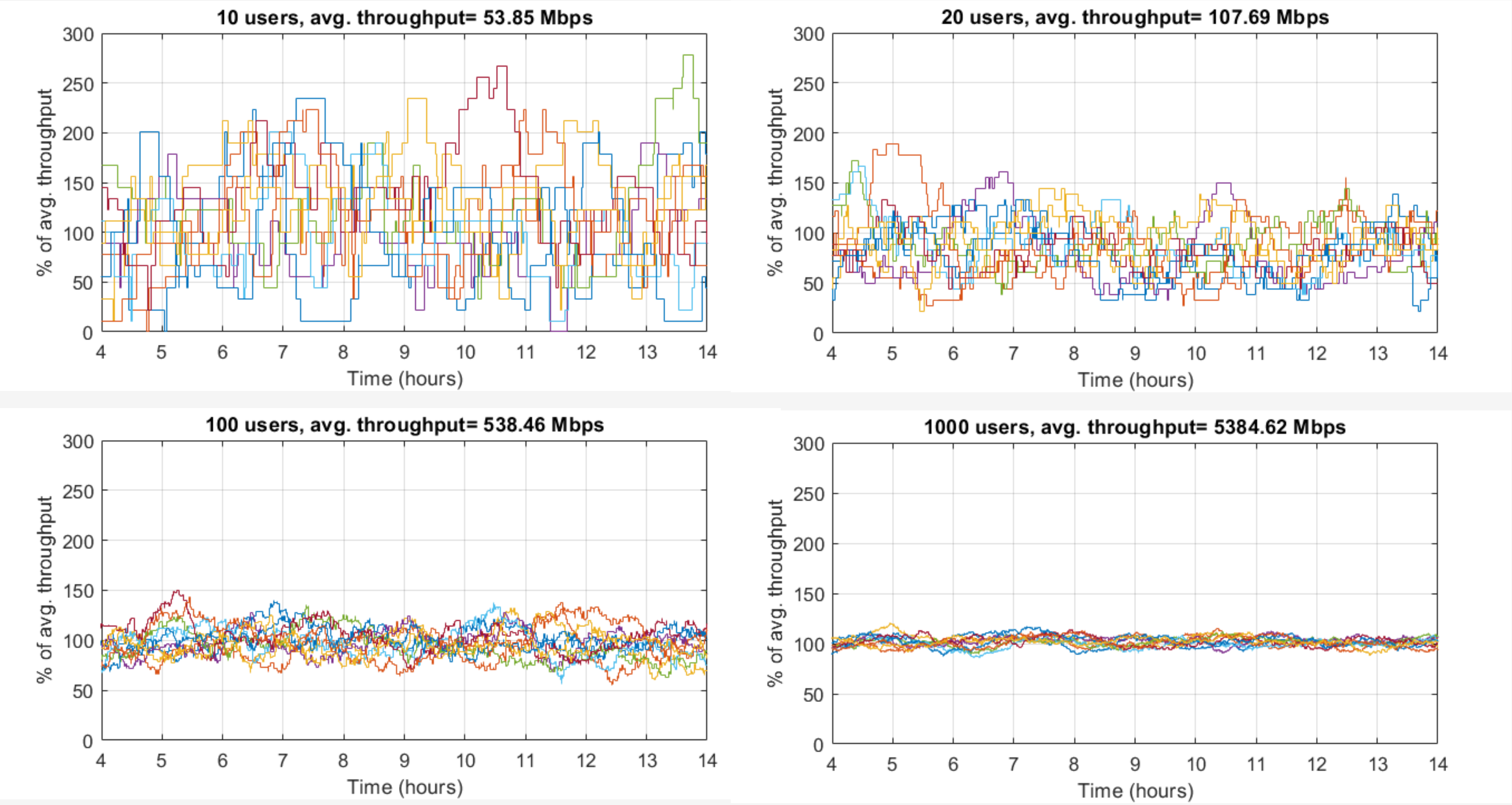
- Consider the simulation of the downstream traffic of *n* clients in an access network from a movie streaming server (hosted in a remote DC). The simulated case is characterized as:
  - depending on the client terminal device, the movie is transmitted in a stream of 6, 12 or 24 Mbps (all formats with equal probability)
  - the service access continuous time duration of each client is between 0.5 and 2 hours (with a uniform distribution)
  - the time interval between service accesses of each client is exponentially distributed with an average of 2 hours
  - in each service access, a pause between 2 and 8 minutes (with a uniform distribution) can happen with probability *p*.
- In the following slides, 10 simulation runs are visualized.



Simulation of the aggregated traffic in the link from B to A

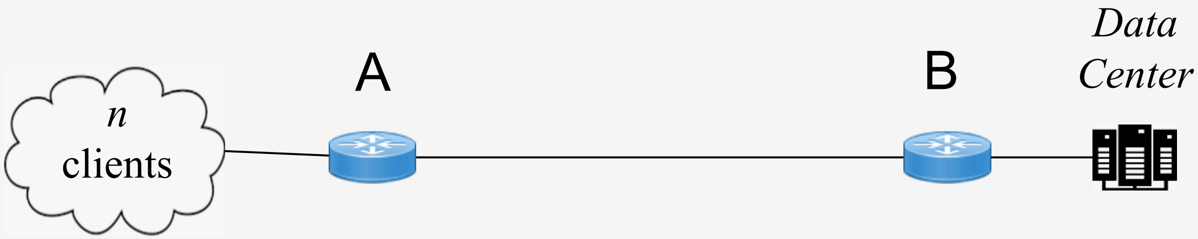
Aggregated traffic without pauses



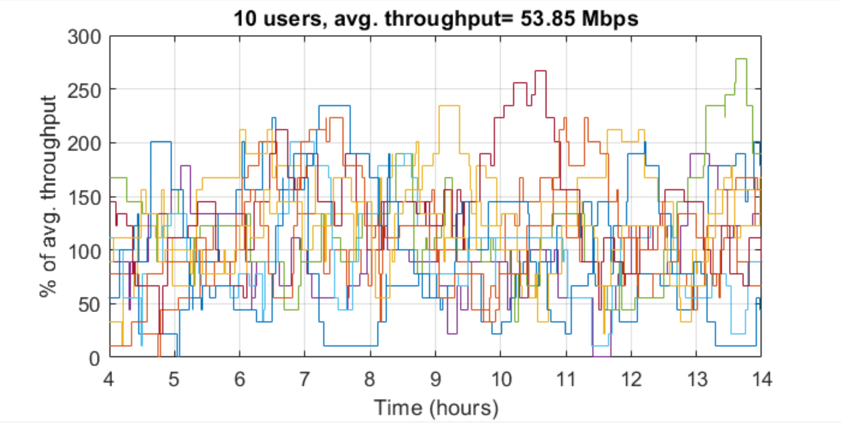


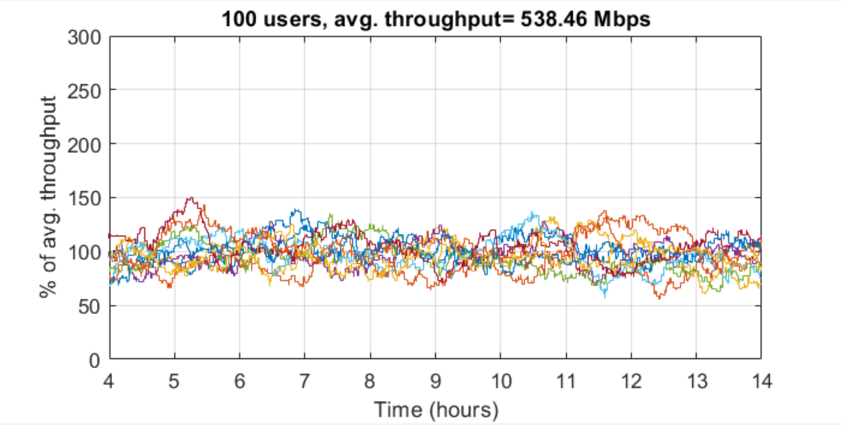
For a large no. of clients, the aggregated traffic throughput variation is very small (in percentage terms) around the average aggregated traffic throughput.

Aggregated traffic

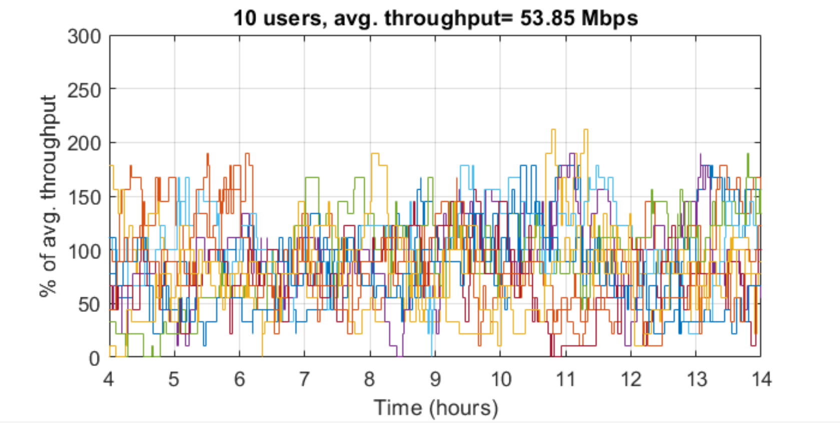


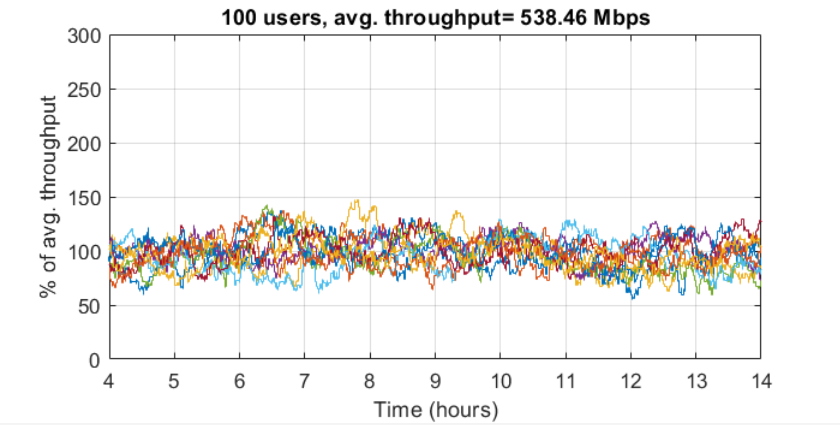
Without pauses (p = 0.0)





With pauses (p = 0.99)





The pauses reduce a little the throughput variation around the average and the reduction is observable only when the no. of clients is small.

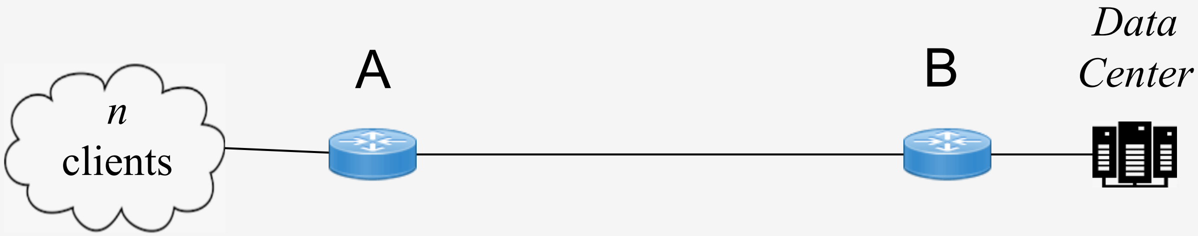
Average throughput of aggregated flows

The previous simulations show that the traffic throughput supported by the network is approximately equal to the average traffic throughput for services with a large number of clients.

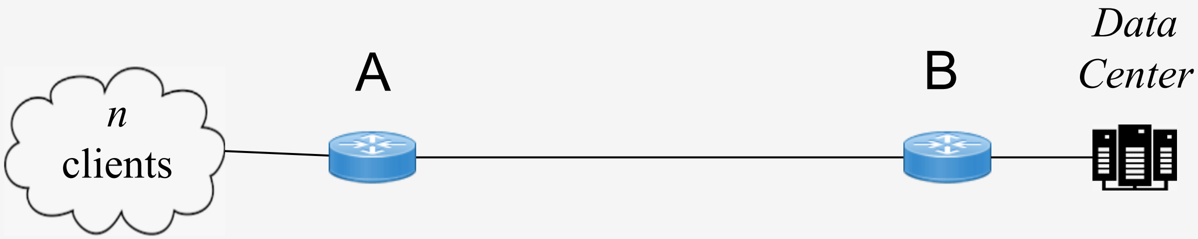
Consider the downstream traffic on *n* clients in an access network from a movie streaming server.

- The movies have an average duration of 90 minutes and are available in 3 formats (for example, HD, FHD and 4K) whose transmission throughput is 6, 12 and 24 Mbps, respectively.
- Each client gets on average 2 movies per day.
- Movies in HD, FHD and 4K formats are requested by 20%, 30% and 50% of clients, respectively.

1. Determine the average downstream throughput *r* per client.
2. Determine the average aggregated throughput *R* for *n* = 1000 clients.



Average throughput of a traffic aggregate



1. Determine the average downstream throughput *r* per client.

The avg. throughput of a movie per client when he is getting a movie is:

$0.2 \times 6 + 0.3 \times 12 + 0.5 \times 24 = 16.8$ Mbps

The average fraction of time each client is getting movies is:

$2 \times 1.5 / 24 = 0.125$ (=12.5%)

Finally:

$r = 16.8 \times 0.125 = 2.1$ Mbps

2. Determine the average aggregated throughput *R* for *n* = 1000 clients.

$R = n \times r = 1000 \times 2.1 = 2100$ Mbps = 2.1 Gbps

Blocking probability of a video streaming server

- Consider a server with a capacity of attending *m* clients. What is the blocking probability of the server (i.e., the probability of a movie request being refused because the server is attending *m* clients)?
- If the movie request arrivals are a Poisson process with rate λ (no. of requests per unit of time) and the duration of the movies is exponenctially distributed with average 1/μ (units of time), the server performance can be modelled by a *M*/*M*/*m*/*m* queuing system.
- Probability of the server being attending *n* clients:

$$P_n = \frac{(\lambda/\mu)^n / n!}{\sum_{i=0}^{m} (\lambda/\mu)^i / i!} \qquad n = 0, 1, \ldots, m$$

- Blocking probability (ErlangB formula):

$$P_m = \frac{(\lambda/\mu)^m / m!}{\sum_{i=0}^{m} (\lambda/\mu)^i / i!}$$

- It is possible to demonstrate that these mathematical expressions are valid for any movie duration statistics as long as the movie duration is statistically independent of the time instants of movie request arrivals.

Blocking probability and availability of an anycast service

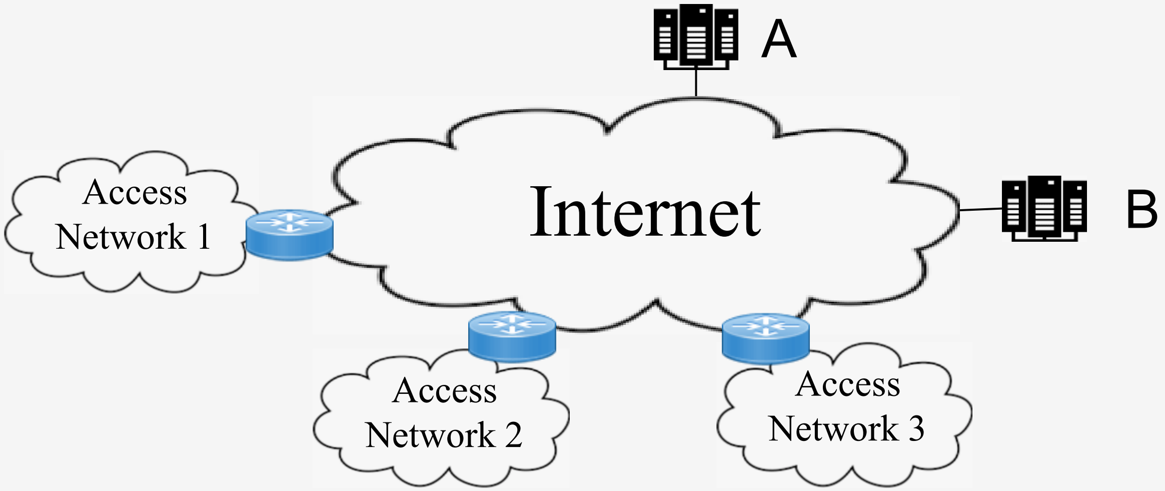
- Consider an anycast service such that each client can be connected to any available server in operation.
- In this case, if the closest server is very busy, the service request can be routed towards another available server:
  - in general, there is a load balancing module that distributes the service requests among the different servers taking into consideration the origin node of the requests and the service load of each server.
- Therefore, in terms of availability, it is conceptually equivalent to consider a single server whose capacity is the sum of the capacities of all servers.

- If each server has an associated availability value (which represents the probability of the server being operational)
- then the service blocking probability is the weighted sum of the blocking probability of the available servers in all possible failure cases, where the weights are the probabilities of all possible failure cases.
- Example in the next slide.

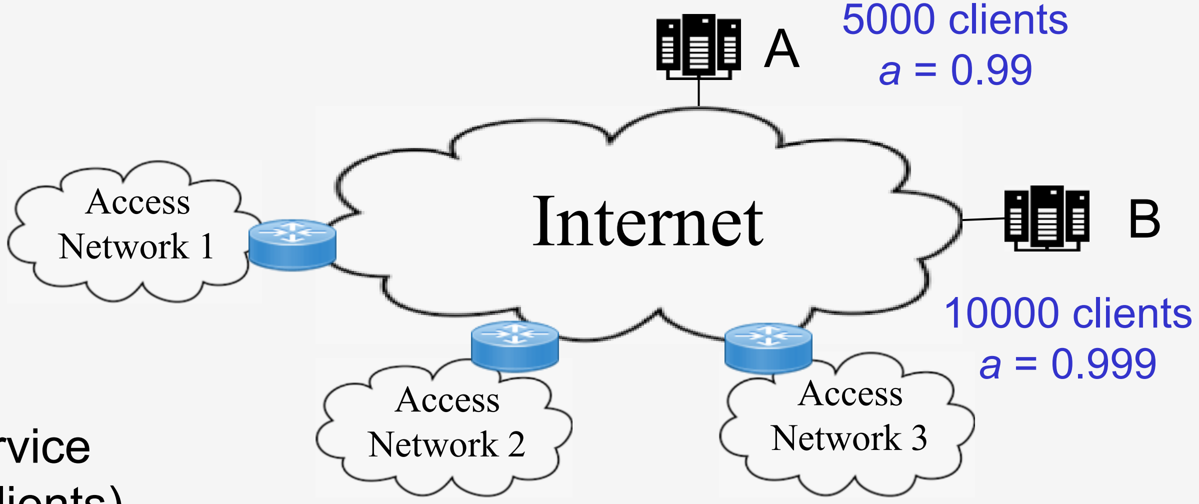
Example

Consider a movie streaming anycast service with 2 servers (shown in the figure): a server hosted in Data Centre A with availability 0.99 and a capacity of 5000 clients and a server hosted in Data Centre B with availability 0.999 and a capacity of 10000 clients.

1. Determine the average service capacity *C* (in number of clients).
2. Determine the service blocking probability *P* when the client request rates are 5000, 3000 and 2000 requests per hour (in the access networks 1, 2 and 3, respectively) and the movies have an average duration time of 90 minutes.



Example



1. Determine the average service capacity C (in number of clients).

Probability of the 2 servers being operational

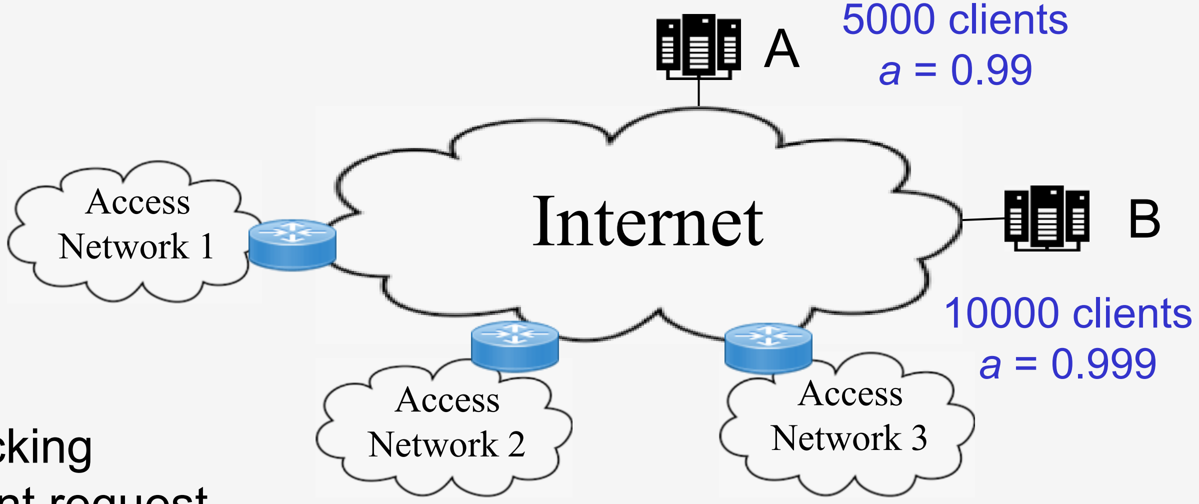
Probability of server in A being operations and server in B being down

Probability of server in A being down and server in B being operational

$$C = (5000+10000)\times 0.99\times 0.999 + 5000\times 0.99\times(1-0.999) + 10000\times(1-0.99)\times 0.999 =$$

$$= 15000\times 0.989 + 5000\times 0.00099 + 10000\times 0.00999 = 14939.85 \text{ clients}$$

Example


A

5000 clients
a = 0.99

B

10000 clients
a = 0.999

2. Determine the service blocking probability P when the client request rates are 5000, 3000 and 2000 requests per hour (in the access networks 1, 2 and 3, respectively) and the movies have an average duration time of 90 minutes.

$$E(\lambda/\mu, m) = \frac{(\lambda/\mu)^m / m!}{\sum_{i=0}^{m} (\lambda/\mu)^i / i!}$$

λ = 5000+3000+2000 = 10000 request/hour

$1/\mu$ = 1.5 horas $\qquad \lambda/\mu = 10000 \times 1.5 = 15000$

$P = E(\lambda/\mu,15000)\times0.989 + E(\lambda/\mu,5000)\times0.00999 + E(\lambda/\mu,10000)\times0.00099 =$

$= 0.0065\times0.989 + 0.6667\times0.00099 + 0.3335\times0.00999 = 0.0104 = 1.04\%$

Blocking probability when the capacity is 15000 clients

Blocking probability when the capacity is 5000 clients

Blocking probability when the capacity is 10000 clients