March 2nd, 2018

Exam of Switching technologies for data centers (2017/18)

Rules for the exam. It is **forbidden** to use notes, books or calculators. Use only draft paper provided by the professor. When needed, use approximations. The answers must be provided in correct English. Any notation must be defined.
Time available: 70 minutes.

Problem A

Consider an input queued switch with Virtual Output Queueing fed by admissible Bernoulli i.i.d. (ABIID) traffic. The adopted scheduler computes the *maximum size matching* at each timeslot.

1. Explain in details the meaning of "admissible" traffic
2. Explain in details the meaning "Bernoulli i.i.d." traffic
3. Describe an ABIID traffic pattern under which the scheduler achieves 100% throughput
4. Describe an ABIID traffic pattern under which the scheduler does not achieve 100% throughput and prove such property, by explaining all the required assumptions.

Problem B

Consider the topology of a data center with 2 layers, built only with switches with 50 ports running at 1 Gbps, interconnecting 800 blade servers, each with a single port at 1 Gbps. The adopted oversubscription ratio is 4:1. The rack is a standard 19-inch one. Assume that the average total traffic generated at each server interface is λ Gbps, with $\lambda \in [0, 1]$.

1. Design the overall interconnection network connecting the servers.
2. Compute the required number of switches and racks.
3. Draw the corresponding Clos network
4. Apply the Lee method to evaluate the blocking probability on the Clos network in function of λ.
5. How it is possible to interpret the blocking probability for the original data center scenario? Under which assumptions?

Problem C

Given the following routing table:

```
IP routing table
Destination/Netmask     Gateway (Rule)
16.0.0.0/5              A
28.0.0.0/6              B
26.0.0.0/7              C
128.0.0.0/5             D
193.0.0.0/8             E
129.0.0.0/8             F
default                 G
```

1. Build the corresponding binary trie, after having defined the fields within each node.
2. Build the corresponding Patricia trie, after having defined the fields within each node.

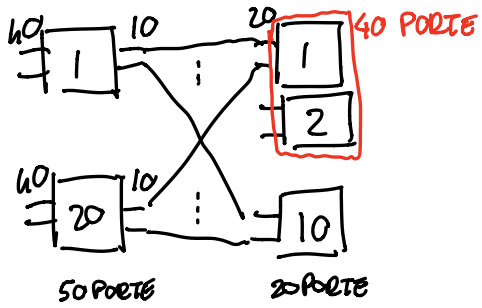
Hints for the solution

Problem A

Same as ex. 86.

Problem B

1. The topology is shown in figure.

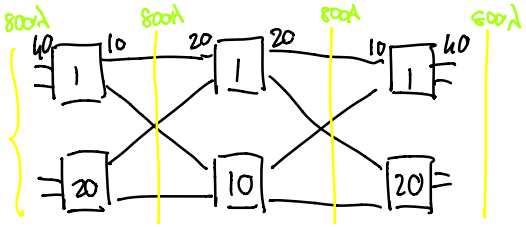


2. Let S_x be a switch with x ports. The total number of switches is:

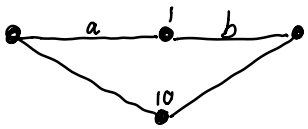
$$20S_{50} + 10S_{20} = 20S_{50} + 5S_{50} = 25S_{50}$$

The required number of racks is 20 for the leaf layer and 1 for the spine layer, thus a total of 21 racks are needed.

3. The Clos network is shown in the figure:



4. To apply the Lee method, the pi-graph is shown below:



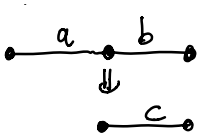
Let ρ be the normalized load on a link defined as:

$$\rho = \frac{\lambda[\text{Gbps}]}{1[\text{Gbps}]}$$

Now:

$$a = b = \frac{800\rho}{10 \times 20} = 4\rho$$

Note that it must $\rho \leq 0.25$ to ensure admissible traffic. The first reduction is shown below:



where

$$c = 1 - (1 - a)(1 - b) = 1 - (1 - 4\rho)^2 = 8\rho - 16\rho^2$$

The second reduction is shown below:

where

$$d = c^{10} = (8\rho - 16\rho^2)^{10}$$

In summary, the total blocking probability for the Clos network is:

$$P_b = (8\rho - 16\rho^2)^{10}$$

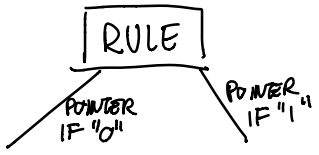
with $\rho \leq 0.25$ (i.e., $\lambda \leq 250$ Mbps).

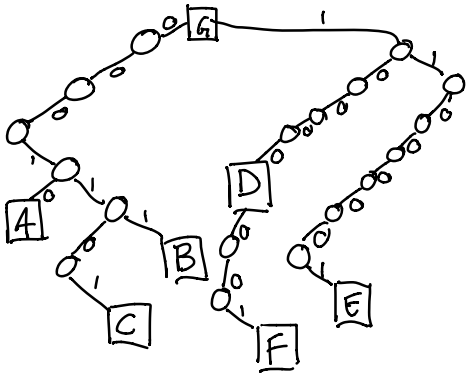
5. The block event in the Clos network corresponds to the event that all the paths from an idle input to an idle output are busy. Thus, in the data center network, this event corresponds to fact that all paths from an idle ToR switch to another idle ToR switch are "busy", i.e. congested by other traffic flows. Thus, P_b can be seen as the level of congestion in the data center. And $(1 - P_b)$ can be seen as the probability that a new flow from an idle ToR switch to another idle ToR switch will experience very small latency, due to the lack of congested links along the path.

   The Lee method is an approximated method and implies strong assumption on the data center system, limiting its applicability: (1) the traffic is uniformly distributed between any pair of ToR switches, (2) each server is source/destination of at most one traffic flow, (3) the routing algorithm distributes randomly the traffic across the topology and all the paths between two ToR traverse always the spine layer, (4) the congestion state among links is independent for all the data center links.

Problem C

The corresponding binary trie is shown in figure:





The corresponding Patricia trie is shown in figure: