Problem Set 2

Instructor: V. Arun
CS 653, Fall 2010
October 23, 2010

1. **(Short answer questions, 20 points)** In the short answer questions below, "explain" means explain briefly in a couple sentences.

   (a) Can shortest path routing (with a single shortest path between each pair of nodes) achieve optimal traffic engineering (TE)? Explain. Can shortest path routing with equal cost multipath (ECMP) splitting achieve optimal TE?

   (b) Explain three mechanisms that ASes use to engineer inbound traffic.

   (c) MPLS is a mechanism that enables optimal traffic engineering for a given traffic matrix. True or false? Explain.

   (d) Oblivious routing circumvents the traffic matrix estimation problem. Explain.

   (e) Gallager's distributed min-delay routing circumvents the traffic matrix estimation problem. True or false? Explain.

   (f) Explain two reasons why the traffic engineering problem is a more complex network management problem than the abstract optimization problem characterized as "Input: Traffic matrix, Output: Set of link weights".

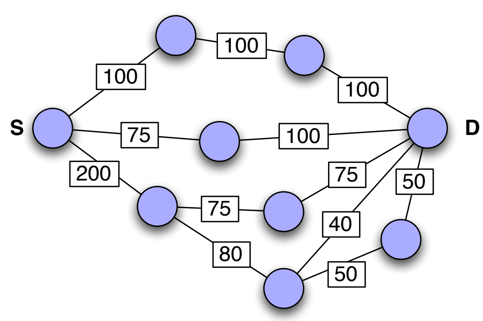

Figure 1: Topology with links annotated with their capacities.

2. **(OSPF, 12 points)** Consider the topology shown in Figure 1. What is the maximum amount of traffic demand from s to d that can be supported by

   (a) OSPF with weights set to the inverse capacity.

   (b) Hop-count OSPF, i.e., with uniform weights of 1.

   (c) An OSPF-like link-state protocol that selects routes based on bottleneck capacity along the route.

(d) OSPF with weights set to the inverse capacity that considers two paths to be of equal cost if they are within a threshold of $\epsilon = 1/100\text{Mbps}$.

3. **(More OSPF, 10 points)** Consider a cost function $\phi_i = 1 + 200u_i^2$, where u_i is the load on link i divided by the capacity of link i. Let $\phi = \sum_i \phi_i$ over all links i. Give an example of a topology and flow demands where OSPF with equal cost splitting is at least 100 times worse than optimal.

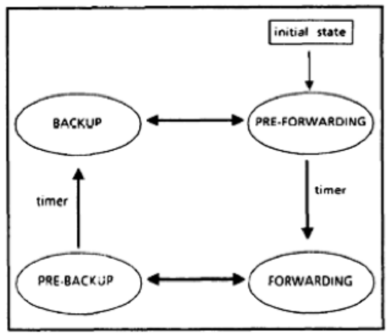

Figure 2: State machine for Perlman's spanning tree bridging algorithm.

4. **(Spanning tree bridging protocol) (15 points)** In the spanning tree bridging algorithm discussed in class [Perlman85], let PRE_BACKUP_DELAY and PRE_FORWARD_DELAY denote the time spent respectively in the PRE_BACKUP and PRE_FORWARDING states as shown in Figure 2 below. Denote the expiry time of messages in the system by MAX_AGE.
   (a) Explain why PRE_BACKUP_DELAY must be greater than 2*MAX_AGE.
   (b) Explain why PRE_FORWARD_DELAY must be greater than 3*MAX_AGE.
   (c) Give three reasons why the spanning tree protocol is unsuitable for a large data center network.

5. **Data center networking (10 + 9 + 8 + 3 points)** Consider a data center network with $N = 16,384$ servers. For all questions below, assume that each server is trying to send data to every other server as fast as possible.
   (a) Consider the SEATTLE architecture discussed in class. Assume that the network is constructed using a link state core of $M = 32$ switches with a switch-to-switch line speed of $C = 10$ Gbps.
      i. For how many servers does each switch resolve routes or addresses?
      ii. How many forwarding table entries does each switch maintain when route caching is not used?
      iii. How many forwarding table entries does each switch maintain when route caching is used assuming all-to-all communication? How does this compare to a more traditional switched architecture?
      iv. What is the average bandwidth of each flow in the network when route caching is not used?
      v. What is the average bandwidth of each flow in the network when route caching is used?

(b) Consider the VL2 architecture discussed in class. Assume that the aggregation switches have $D_A = 32$ ports each and the intermediate switches have $D_I = 16$ ports each.

   i. Assuming that $M = 32$ aggregation switches are used, how many intermediate switches are required?
   ii. What is the expected bandwidth of bottlenecked flows in the network?
   iii. Explain how a packet sent from an Internet client to a server inside the data center gets routed to that server. In particular, are MAC (48-bit Ethernet) addresses used in this process? If so, explain how. If not, explain why not.

(c) Consider the BCube architecture discussed in class.

   i. How many switches are required?
   ii. What is the maximum path length?
   iii. How many disjoint paths exist between each pair of nodes?
   iv. What is the average bandwidth of each flow?

(d) Which of the above architectures use multiple paths to route a switch-to-switch flow?