CSE 222a, Spring 2017

Midterm 2 quiz for May 17, 2017

Name: Sample solutions

PID:

NOTE: Questions are on both sides of the paper. For short answers, only the response in the provided box will be graded. If you are using your computer, disable wifi/cellular connection.

Allowed resources:

- Papers and notes (printed or on your computer) in PDF or word processor format (*.pdf, *.txt, *.tex, *.rtf, etc). Local copies only.

Resources that are not allowed:

- Documents stored in Google Docs, Microsoft 365, anything online
- Any communication with anyone (IM, GChat, G+ hangouts, shared docs)
- Any web browser open
- Exchanging or sharing notes during the exam.
- Smartphones are not allowed.

This midterm is graded out of 100 points, and counts towards 15% of your final grade.

Name: _____

1. Programmable data paths (25pts)

1.1 (10 pts) Click has been successfully deployed as part of a number of commercial products, whereas the ANTS/Capsules approach to active networks has not. List two reasons why this might be:

A few reasons include: (1) ANTS requires that the client be involved in injecting code into the network, (2) ANTS' programming model is restricted, and can only act on capsules, rather than arbitrary traffic, (3) ANTS has a limited ability to keep and maintain state across packets, and cannot keep state across flows, making it not a good platform for implementing firewalls or intrusion detection systems, (4) ANTS runtimes were not generally tuned for performance, making them quite slow.

1.2 (10 pts) Although P4 isn't as expressive as Click (it can only rely on match/action tables compared to Click's Turing-complete programming environment), it is being deployed in new switches. What is an example of functionality that fits well into P4's programming model (5 pt), and what is the advantage of P4 compared to Click (5 pt)?

A few of the advantages of P4 over Click are that (1) it can operate at the line rate of modern switches due to its restricted programming model, (2) it has been adopted by industry and is in-built into recent commercial switches, (3) its programming model fits the match/action architecture of switches, which gives developers the ability to add processing to the dataplane, (4) P4 is target independent, meaning that a variety of vendors, devices, etc can all be the target of a single P4 program.

1.3 (5 pts) Click's "FromDevice" element relies on a version of the Ethernet driver that uses polling, instead of interrupts, to detect incoming packet arrival events. Why is this (choose 1)?

[ ] (a) To reduce the load on CPUs from this element

[ ] (b) To keep packet payload data consistent between cores

[XX] (c) To prevent a livelock event in which the overhead of interrupt handler invocations reduces overall throughput

[ ] (d) None of these

Name: _____

2. Scale-out network designs (40 pts)

2.1 (10 pts) ARP (Address Resolution Protocol) is a protocol used to map IP addresses to MAC addresses. In Portland, the standard ARP protocol doesn't work in the same way. Describe how ARP works in Portland, specifically describing the roles of the Fabric Manager and the Edge Switch in its implementation:

In Portland, VMs know their own AMACs. To communicate with another VM, they need to map that remote VM's IP address to a MAC address, in this case a PMAC. They issue an ARP, which is intercepted by the edge switch. If the edge switch has that mapping, it is returned to the VM. If not, then the edge switch contacts the fabric manager to retrieve the IP->PMAC mapping. It then returns that to the VM. If a virtual machine moves to another location, then the fabric manager will send a 'gratuitous' ARP message that installs the new IP->PMAC mapping in the edge switch.

2.2 (10 pts) In Data Center Networks, manual configuration has to be avoided at all costs. The Location Discovery Protocol (LDP) is an automatic mechanism for the switches to configure themselves. Briefly explain how this is implemented, including the role of the Location Discovery Message (LDM).

Initially when a switch comes online, it doesn't know its position in the network. It sends LDM messages out all its ports, and it listens for LDMs from its neighbors. If a switch does not hear LDMs from a majority of its ports, then it knows that it is an edge switch. It can then advertise that identity to its neighbors. A switch that receives LDMs from an edge switch knows that it is an aggregation switch, and can advertise that fact. Finally, switches that receive LDMs on all ports from aggregation switches know they are core switches.

2.3 (3 pts) IGP protocols like IS-IS or OSPF, when used in a FatTree network, will load balance traffic across multiple source-destination paths:

[ ] True

[XX] False

Name: _____

2.4 (7 pts) In the Jupiter paper, later generations of interconnects increasingly rely on fiber-optic based optical connections between switches in the datacenter, instead of copper cables. Why is this?

A few reasons include that: (1) copper cables have length restrictions, and are limited to only a few meters of length; fibers don't have that limit, (2) it is possible to use 'wavelength division multiplexing' to put multiple colors of light into the same fiber, which enables the construction of superlinks that have more bandwidth than a typical copper cable, and (3) fibers are much smaller than copper cables, reducing bulk.

2.5 (10 pts) In the Al Fares et al. paper, packets are forwarded based on a 2-level lookup implementation using specially chosen IP addresses. In Portland, packets are forwarded based on PMACs. Describe how both of these approaches result in good utilization of the many potential paths between each source and each destination.

Al Fares paper: in this paper, the first level of the lookup table does longest-prefix matching to direct packets to the right downward-facing port if the packet is in the destination pod. Note that on the downward direction, the path packets take is deterministic and static (assuming no failed links or switches). On the upward direction, the packet is sent to the 2nd level of the lookup table, which uses the suffix (last few bits) to act as a source of entropy, such that destination VMs/hosts differing in just a single bit take separate paths. Since the last couple bits of the IP address refer to different VMs, that means that two VMs in the same host take separate paths through the network.

Portland paper: The Portland paper relies on lower-level mechanisms to implement its forwarding (e.g., a FatTree), and as such the paper itself doesn't explicitly mention how multi-pathing is achieved. For this reason, we gave everyone 5 points on this sub-part of this problem.

Name: _____

3. Congestion control (35pts)

3.1. Define incast, queue buildup, and buffer pressure (10 pts), then describe how these can hurt the performance of an application (10 pts), and describe how DCTCP mitigates each of them (15 pts):

Incast is a short-term phenomenon that results in packet loss in a switch. Incast occurs when a batch of packets arrive to a switch, all destined to a single switch output port, and the batch is larger than the internal buffering devoted (or capable of being devoted) to that output port. As a result, packets are lost, resulting in failures of the application, or requiring the application to retransmit packets, increasing latency and/or decreasing throughput. Note that incast cannot be prevented with TCP, since TCP takes at least one RTT to detect and respond to congestion, and Incast develops at sub-RTT timeframes.

Queue buildup is when a large, bandwidth-oriented flows who have very large congestion windows 'use up' a large amount of buffers in a switch. Packets from low-latency flows that also want to use that output port end up having to wait behind that queue, increasing overall latency.

Buffer pressure arises in shared buffer switches, where the switch allocated more of the shared buffer to ports that are heavily used over ports that are less used. This means that a bandwidth-heavy flow that is using up a lot of buffering on one port can end up reducing the buffering available to a different port, which could lead to lower bandwidth and/or increased likelihood of incast.

DCTCP operates by (1) having switches mark ECN bits immediately when the queue length exceeds a threshold, and (2) having the senders respond quickly and strongly to ECN signals, reducing their sending rates very aggressively. This helps lessen, but not entirely prevent, incast by reducing overall buffer utilization in the system. In a DCTCP deployment, system-wide buffers are kept minimal by the aggressive reducing in sending rates. Having small buffers lets any bursts in traffic use as much of the available buffer as possible. Second, DCTCP prevents queue build-up by having bandwidth-sensitive flows reduce their sending rate when buffers get full, keeping buffer utilization low. Finally, this same effect (lowering buffer utilization of bandwidth-sensitive flows) prevents buffer pressure.

Name: _____

3.1 (con't, if needed)