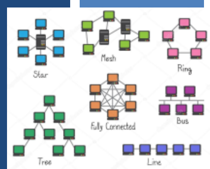
TCSS 558:
APPLIED DISTRIBUTED COMPUTING

Chapter 6 – Coordination – V
Final Exam Practice

Wes J. Lloyd
School of Engineering
& Technology (SET)
University of Washington - Tacoma



1

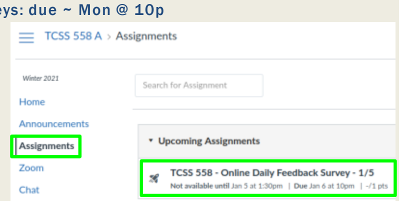
OBJECTIVES – 3/9

- Questions from 3/7
- Assignment 2: Replicated Key Value Store
- Class Activity 4 – Total Ordered Multicasting
- Class Activity 5 – Causality and Vector Clocks
- Chapter 6: Coordination
  - Chapter 6.3: Distributed Mutual Exclusion
  - Chapter 6.4: Election Algorithms

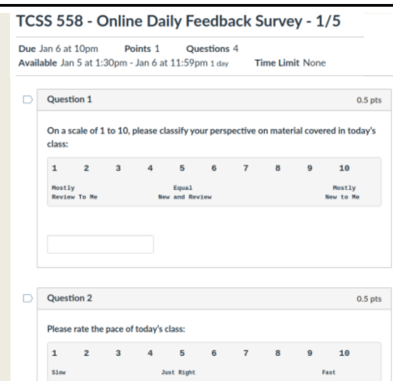
2

ONLINE DAILY FEEDBACK SURVEY

- Daily Feedback Quiz in Canvas – Available After Each Class
- Extra credit available for completing surveys ***ON TIME***
- Tuesday surveys: due by ~ Wed @ 10p
- Thursday surveys: due ~ Mon @ 10p



3



4

MATERIAL / PACE

- Please classify your perspective on material covered in today's class (29 respondents):
- 1-mostly review, 5-equal new/review, 10-mostly new
- **Average – 6.38** (↓ - *previous 6.47*)

- Please rate the pace of today's class:
- 1-slow, 5-just right, 10-fast
- **Average – 5.90** (↑ - *previous 5.83*)

5

FEEDBACK FROM 3/7

- ***For the centralized TCP membership server, if IP address is the key, if we run more than one server on the same IP address like localhost, then the key will be overwritten. Should we assume for this assignment that only one TCP server will be run on a single IP?***
- To enable each node to have a unique key in the membership directory when all nodes are hosted on the same computer, you can simply combine the **IP+PORT** together as the unique key, and then use a generic value or placeholder as the value.
  - Essentially, can then just ignore the value
  - The membership list will be the list of keys

6

FEEDBACK - 2

- ***Should we cache the TCP client connection from the leader node to the other nodes, or is it okay to close and create a new connection for every request?***
- For each "put" or "del" request, the transaction leader can open new connections to each node, and then close them when finished.
- Caching connections (leaving them open between transactions) can provide an optimization, but this creates an issue when there is a long gap before new transactions.
- TCP connections can timeout. The Java socket library allows a timeout value to be specified. If no value is specified, the default OS timeout is used. In Ubuntu the default is visible in the file: "**cat /proc/sys/net/ipv4/tcp_keepalive_time**"
- The default timeout appears to be 2 hours.
- It is out of scope to implement connection caching for assignment 2. It is sufficient to open & close connections for each transaction.

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.7

7

OBJECTIVES – 3/9

- Questions from 3/7
- **Assignment 2: Replicated Key Value Store**
- Class Activity 4 – Total Ordered Multicasting
- Class Activity 5 – Causality and Vector Clocks
- Chapter 6: Coordination
  - Chapter 6.3: Distributed Mutual Exclusion
  - Chapter 6.4: Election Algorithms
- Practice Final Exam Questions

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.8

8

SHORT-HAND-CODES FOR MEMBERSHIP TRACKING APPROACHES

- Include readme.txt or doc file with instructions in submission
- Must document membership tracking method

>> please indicate which types to test <<

ID	Description
F	Static file membership tracking – file is not reread
FD	Static file membership tracking DYNAMIC - file is periodically reread to refresh membership list
T	TCP membership tracking – servers are configured to refer to central membership server
U	UDP membership tracking - automatically discovers nodes with no configuration

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.9

9

ASSIGNMENT 2

- **Friday March 17th *(late Sat March 18th)***
- Goal: Replicated Key Value Store
- Team signup posted on Canvas under 'People'
- Builds off of Assignment 1 GenericNode
- Focus on TCP client/server w/ replication
- How to track membership for data replication?
  - Can implement multiple types of membership tracking for extra credit

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.10

10

OBJECTIVES – 3/9

- Questions from 3/7
- Assignment 2: Replicated Key Value Store
- **Class Activity 4 – Total Ordered Multicasting**
- **Class Activity 5 – Causality and Vector Clocks**
- Chapter 6: Coordination
  - Chapter 6.3: Distributed Mutual Exclusion
  - Chapter 6.4: Election Algorithms
- Practice Final Exam Questions

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.11

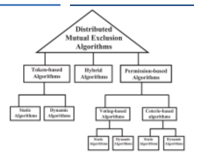
11

OBJECTIVES – 3/9

- Questions from 3/7
- Assignment 2: Replicated Key Value Store
- Class Activity 4 – Total Ordered Multicasting
- Class Activity 5 – Causality and Vector Clocks
- Chapter 6: Coordination
  - **Chapter 6.3: Distributed Mutual Exclusion**
  - Chapter 6.4: Election Algorithms
- Practice Final Exam Questions

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.12

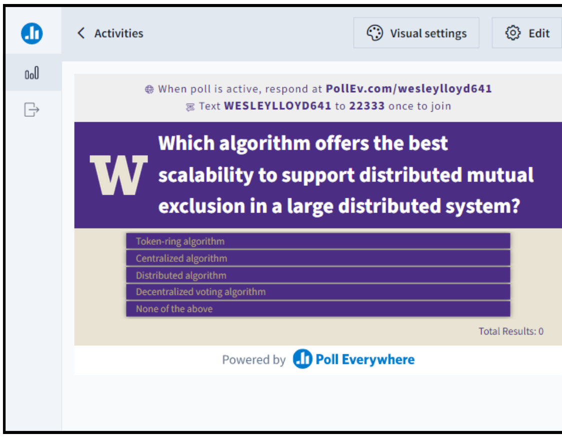
12



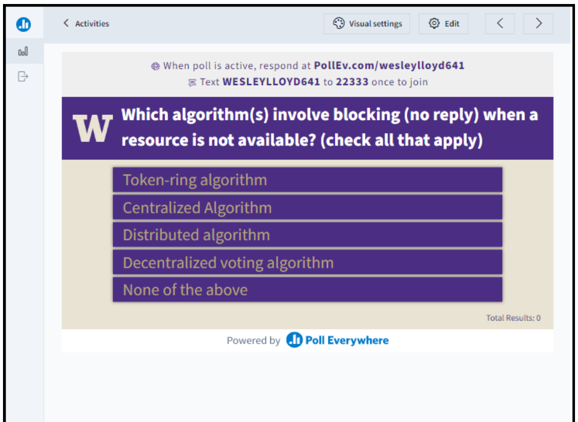
CH. 6.3: DISTRIBUTED MUTUAL EXCLUSION

L19.13

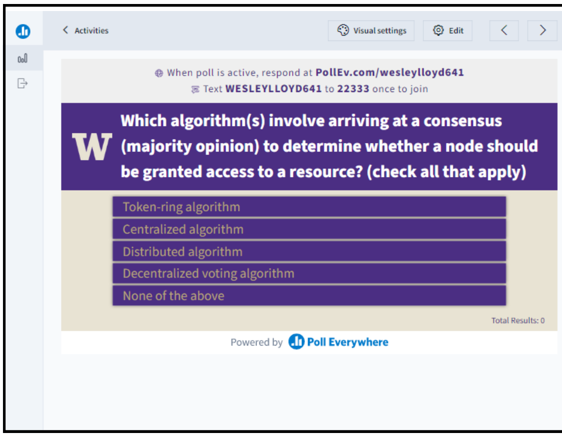
13



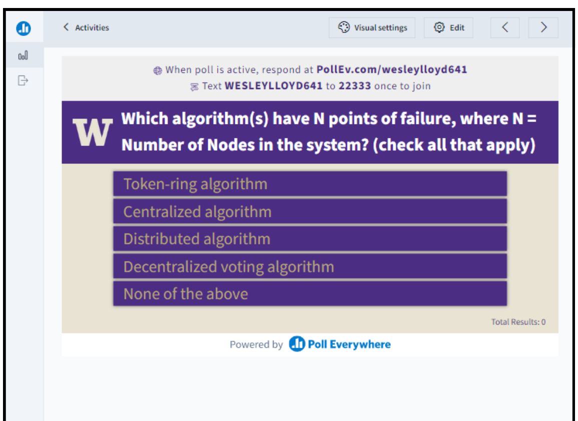
14



15



16



17

DISTRIBUTED MUTUAL EXCLUSION ALGORITHMS REVIEW

- Which algorithm offers the best scalability to support distributed mutual exclusion in a large distributed system?
- (A) Token-ring algorithm
- (B) Centralized algorithm
- (C) Distributed algorithm
- (D) Decentralized voting algorithm

March 9, 2023	TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma	L19.18

18

DISTRIBUTED MUTUAL EXCLUSION ALGORITHMS REVIEW - 2

- Which algorithm(s) involve blocking (no reply) when a resource is not available?
- (A) Token-ring algorithm
- (B) Centralized algorithm
- (C) Distributed algorithm
- (D) Decentralized voting algorithm

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.19

19

DISTRIBUTED MUTUAL EXCLUSION ALGORITHMS REVIEW - 3

- Which algorithm(s) involve arriving at a consensus (majority opinion) to determine whether a node should be granted access to a resource?
- (A) Token-ring algorithm
- (B) Centralized algorithm
- (C) Distributed algorithm
- (D) Decentralized voting algorithm

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.20

20

DISTRIBUTED MUTUAL EXCLUSION ALGORITHMS REVIEW - 4

- Which algorithm(s) have N points of failure, where N = Number of Nodes in the system?
- (A) Token-ring algorithm
- (B) Centralized algorithm
- (C) Distributed algorithm
- (D) Decentralized voting algorithm

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.21

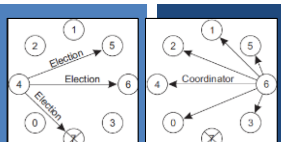
21

OBJECTIVES – 3/9

- Questions from 3/7
- Assignment 2: Replicated Key Value Store
- Class Activity 4 – Total Ordered Multicasting
- Class Activity 5 – Causality and Vector Clocks
- Chapter 6: Coordination
  - Chapter 6.3: Distributed Mutual Exclusion
  - **Chapter 6.4: Election Algorithms**
- Practice Final Exam Questions

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.22

22



CH. 6.4: ELECTION ALGORITHMS

L19.23

23

ELECTION ALGORITHMS

- Many distributed systems require one process to act as a coordinator, initiator, or provide some special role
- Generally any node (or process) can take on the role
  - In some situations there are special requirements
  - Resource requirements: compute power, network capacity
  - Data: access to certain data/information
- Assumption:
  - Every node has access to a "node directory"
  - Process/node ID, IP address, port, etc.
  - Node directory may not know "current" node availability
- Goal of election: at conclusion all nodes agree on a coordinator or "leader"

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.24

24

ELECTION ALGORITHMS

- Consider a distributed system with N processes (*or nodes*)
- Every process has an identifier id(P)
- Election algorithms attempt to locate the highest numbered process to designate as coordinator

- **Algorithms:**
- Bully algorithm
- Ring algorithm
- Elections in wireless environments
- Elections in large-scale systems

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.25

25

BULLY ALGORITHM

- When ***any*** process notices the coordinator is no longer responding to requests, it initiates an election
- Process P_k initiates an election as follows:
  1. P_k sends an ELECTION message to all processes with higher process IDs (P_{k+1}, P_{k+2}, … P_{N-1})
  2. If no one responds, P_k wins the election and becomes coordinator
  3. If a "higher-up" process answers (P_{k+n}), it will take over and run the election. P_k will quit sending ELECTION messages.
- When the higher numbered process receives an ELECTION message from a lower-numbered colleague, it responds with "OK", indicating it's alive, and it takes over the election.

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.26

26

BULLY ALGORITHM - 2

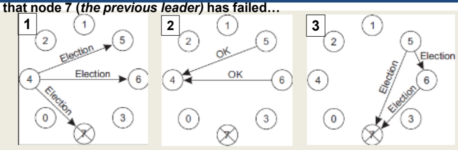
- The higher numbered process then holds an election with ***only*** higher numbered processes (nodes).
- Eventually ***all*** processes give up except one, and the remaining process becomes the new coordinator.
- The coordinator announces victory by sending all processes a message stating it is starting as the coordinator.
- If a higher numbered node that was previously down comes back up, it holds an election, and ultimately takes over the coordinator role.
- The process with the *"biggest"* ID in town always wins.
- Hence the name, **bully algorithm**

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.27

27

BULLY ALGORITHM - 3

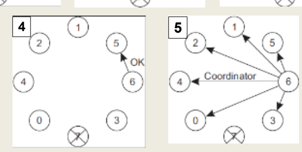
Note that node 7 (*the previous leader*) has failed…



[1] Process 4 starts an election

[2] Process 5 and 6 respond

[3] Process 5 and 6 each hold an election



[4] Process 6 tells Process 5 to stop

[5] Process 6 wins and tells everyone

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.28

28

BULLY ALGORITHM - 4

- **Requirement:** Every node knows who is participating in the distributed system
  - Each node has a group membership directory
- First process to notice the leader is offline launches a new election
- GOAL: Find the highest number node that is running
  - Loop over the nodes until the highest numbered node is found
  - May require multiple election rounds
- Highest numbered node is always the ***"BULLY"***

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.29

29

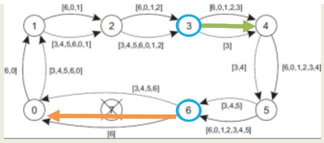
RING ALGORITHM

- Election algorithm based on a network of nodes in logical ring
- *Does not use a token*
- Any process (P_k) starts the election by noticing the coordinator is not functioning
  1. P_k builds an **election message**, and sends to its successor in the ring
     - If successor is down, successor is skipped
     - Skips continue until a running process is found
  2. When the **election message** is passed around, each node adds its ID to a *separate* **active node list**
  3. When **election message** returns to P_k, P_k recognizes its own identifier in the **active node list**. Message is changed to COORDINATOR and "**elected(P_k)**" message is circulated.
     - Second message announces P_k is the NEW coordinator

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.30

30

RING: MULTIPLE ELECTION EXAMPLE



- **PROBLEM:** Two nodes start election at the same time: P_3 and P_6
- P_3 sends **ELECT(P_3)** message, P_6 sends **ELECT(P_6)** message
  - P_3 and P_6 both circulate ELECTION messages at the same time
- Also circulated with ELECT message is an **active node list**
- Each node adds itself to the **active node list**
- Each node votes for the highest numbered candidate
- P_6 wins the election because it's the candidate with the **highest ID**

31

ELECTIONS WITH WIRELESS NETWORKS

- Assumptions made by traditional election algorithms not realistic for wireless environments:
  - >>> Message passing is reliable
  - >>> Topology of the network does not change
- A few protocols have been developed for elections in ad hoc wireless networks
- Vasudevan et al. [2004] solution handles failing nodes and partitioning networks.
  - Best leader can be elected, rather than just a random one

32

VASUDEVAN ET AL. WIRELESS ELECTION

1. Any node (***source***) (P) starts the **election** by sending an ELECTION message to immediate neighbors (any nodes in range)
2. Receiving node (Q) designates sender (P) as parent
3. (Q) Spreads election message to neighbors, ***but not to parent***
4. Node (R), receives message, designates (Q) as parent, and spreads ELECTION message to neighbors, ***but not to parent***
5. Neighbors that have already selected a parent immediately respond to R.
   - If ***all*** neighbors already have a parent, R is a leaf-node and will report back to Q quickly.
   - When reporting back to Q, R includes metadata regarding battery life and resource capacity
6. Q eventually acknowledges the ELECTION message sent by P, and also indicates the most eligible node (based on battery & resource capacity)

33

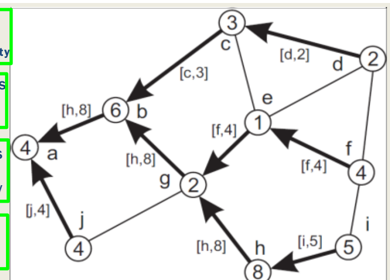
WIRELESS ELECTION - 2 SOURCE NODE: [A]

Node [A] initiates election: find the highest capacity

Election messages propagated to all nodes

Each node reports to its parent node with best capacity

Node A then facilitates Node H becoming leader



34

WIRELESS ELECTION - 3

- When multiple elections are initiated, nodes only join one
- Source node tags its ELECTION message with unique identifier, to uniquely identify the election.
- With minor adjustments protocol can operate when the network partitions, and when nodes join and leave

35

ELECTIONS FOR LARGE-SCALE SYSTEMS

- Large systems often require several nodes to serve as coordinators/leaders
- These nodes are considered ***"super peers"***
- ***Super peers*** must meet operational requirements:

1. Network latency from normal nodes to ***super peers*** must be low
2. ***Super peers*** should be evenly distributed across the overlay network (ensures proper load balancing, availability)
3. Must maintain set ratio of ***super peers*** to normal nodes
4. ***Super peers*** must not serve ***too many*** *normal nodes*

36

ELECTIONS FOR DHT BASED SYSTEMS

- DHT-based systems use a bit-string to identify nodes
- **Basic Idea**: Reserve fraction of ID space for super peers
- Reserve first **$\log_2(N)$** bits for super-peer IDs
- m=number of bits of the identifier
- k=# of nodes each node is responsible for (Chord system)
- **Example:**
- For a system with m=8 bit identifier, and k=3 keys per node
- Required number of super peers is $2^{(k - m)} \cdot N$, where N is the number of nodes
  - In this case N=32
  - **Only 1 super peer is required for every 32 nodes**

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.37

37

SUPER PEERS IN AN M-DIMENSIONAL SPACE

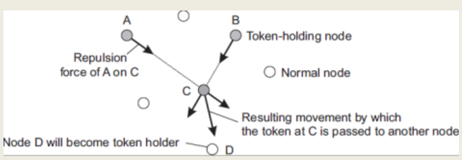
- Given an overlay network, the idea is to position superpeers throughout the network so they are evenly disbursed
- **Use tokens:**
- Give N tokens to N randomly chosen nodes
- No node can hold more than (1) token
- Tokens are "repelling force". Other tokens move away
- All tokens exert the same repelling force
- This automates token distribution across an overlay network

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.38

38

OVERLAY TOKEN DISTRIBUTION

- Gossping protocol is used to disseminate token location and force information across the network
- If forces acting on a node with a token exceed a **threshold**, token is moved away
- When nodes hold token for awhile they become superpeers



March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.39

39

WE WILL RETURN AT 2:40 PM

40

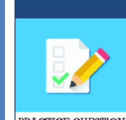
OBJECTIVES – 3/9

- Questions from 3/7
- Assignment 2: Replicated Key Value Store
- Class Activity 4 – Total Ordered Multicasting
- Class Activity 5 – Causality and Vector Clocks
- Chapter 6: Coordination
  - Chapter 6.3: Distributed Mutual Exclusion
  - Chapter 6.4: Election Algorithms
- **Practice Final Exam Questions**

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.41

41

PRACTICE QUESTIONS



42

QUESTION 1: MULTI-TIERED ARCHITECTURE

- Consider a multi-tiered software architecture consisting of three distinct components: a web application server, a relational database server, and a log server. Describe the differences between a vertical distribution and a horizontal distribution of these components (Lecture 6)?
  - *Address the implications of these distributions for **scalability***

  - *Web application server:*
  - *Relational database server:*
  - *Log server:*

QUESTION 2: CENTRALIZED SERVER ARCHITECTURE

- Consider a traditional centralized server architecture where many client nodes communicate with a single server node.
- Consider the four design goals of distributed systems from Ch. 1: Resource sharing, Distribution Transparency, Openness (*interoperability, portability, extensibility*), and Scalability.
- Describe challenges with ensuring these design goals when adopting a centralized server architecture.
- >> Consider citing an example if helpful.

Resource sharing:

Distribution transparency:

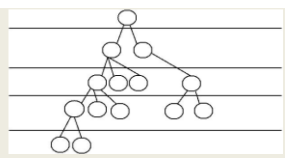
Openness:

Scalability:

QUESTION 3: ARCHITECTURE DIFFERENCES

- Describe two communication differences between a traditional connection-oriented client/server architecture, and a publish/subscribe architecture where clients and servers communicate by exchanging tuples in a shared data space.

QUESTION 4: UNSTRUCTURED PEER-TO-PEER NETWORK

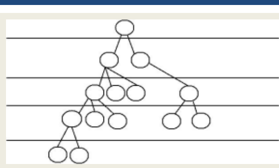


- Fourteen nodes communicate using an unstructured peer-to-peer network using random walks. The head node pictured at the top of the graph for this network receives a client request to retrieve a data element. The data element is only stored at one node. The precise node is unknown. Starting at the head node using message flooding without a specified time-to-live (TTL), how many messages are sent to locate the data item?

  assume that messages are not resent to originating nodes

QUESTION 4 (2): UNSTRUCTURED PEER-TO-PEER NETWORK

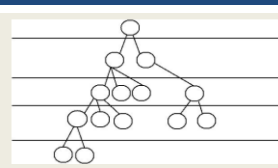
2-parts



- Using a random walk beginning at the head node at the top of the graph where only one walk per level is performed without a specified time-to-live (TTL), how many nodes will be visited? ***assume that messages are not resent to originating nodes***
- Given this number of node visitations, and considering that the data element is not replicated in the network as it exists at only one node, what is the probability (in %) that the data element will be found? **Specify a min to max range.**

QUESTION 4 (3): UNSTRUCTURED PEER-TO-PEER NETWORK

2-parts

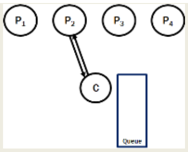


- If we perform two parallel walks per level without a TTL, what is the worst-case probability (%) of finding the data element?

  at each node, two paths are followed ---
- For this scenario, what is the best-case probability (%) of finding the data element?

  at each node, two paths are followed ---

QUESTION 5: DISTRIBUTED MUTUAL EXCLUSION



- List one advantage, and one disadvantage for centralized distributed mutual exclusion:
- Advantage: Disadvantage:

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.49

49

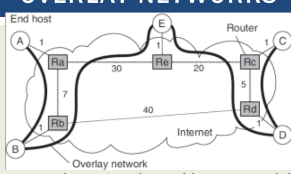
QUESTION 6: TIME MANIA

- Approaches to synchronizing time across all of the nodes of a distributed system focus on ensuring either one or both of the following: **accuracy** and/or **precision**
- For each time tracking approach below, identify whether it provides accuracy, precision, or both for coordinating time across the nodes in a distributed system.
- NTP:
- Berkeley:

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.50

50

QUESTION 7: OVERLAY NETWORKS

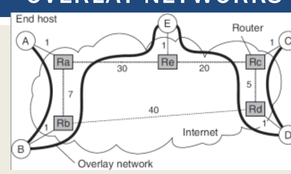


- In the figure, an overlay network provides connectivity among the nodes: A, B, C, D, and E.
- The overlay network is implemented using "underlying" networks. In this case, the underlying network consists of a series of routers: Ra, Rb, Rc, Rd, and Re. Network "Weights" are assigned to each of the links between the routers indicating approximate communication delay. For example, the communication delay between Ra and Rb is 7 units, whereas the communication delay between node A and Ra is just 1 unit.

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.51

51

QUESTION 7: OVERLAY NETWORKS



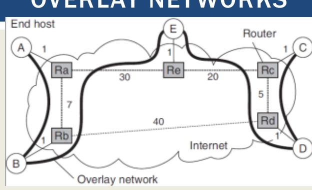
- When nodes communicate using the overlay network, they must route messages via (by way of) the "overlay" links. In the diagram above, there are overlay links between: A → B, B → E, E → D, and D → C.

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.52

52

QUESTION 7: OVERLAY NETWORKS

2-parts

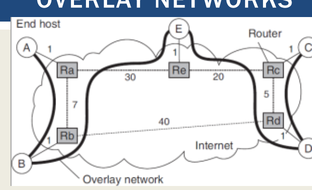


- (A) What is the network delay when routing a message using the overlay network from node D to B? _______________units
- (B) What is the network delay when sending this same message from node D to B via the most efficient path using the underlying network? _______________ units

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.53

53

QUESTION 7: OVERLAY NETWORKS

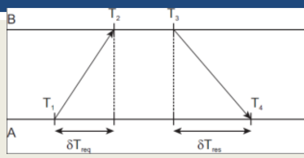


- (C) Network "Stretch" is the ratio of the overlay network delay to the underlying network delay. For this example, what is the network stretch? _______________units

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.54

54

QUESTION 8: SYNCHRONIZATION



- In the Network Time Protocol, node A is a client that communicates with node B, which is an NTP server. The communication propagation delay is estimated with the formula:

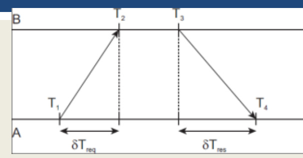
$$\theta = \frac{|T_2 - T_1| + |T_3 - T_4|}{2}$$

- (a) What key assumption is made about the propagation delay between A and B?

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.55

55

QUESTION 8: SYNCHRONIZATION



- (b) When NTP is used to synchronize clocks of client computers, when client clocks are ahead of the NTP server due to clock skew, why do clients never set their local clock(s) backwards to match the time of the NTP server?

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.56

56

QUESTIONS

March 9, 2023 | TCSS558: Applied Distributed Computing [Winter 2023] School of Engineering and Technology, University of Washington - Tacoma | L19.57

57