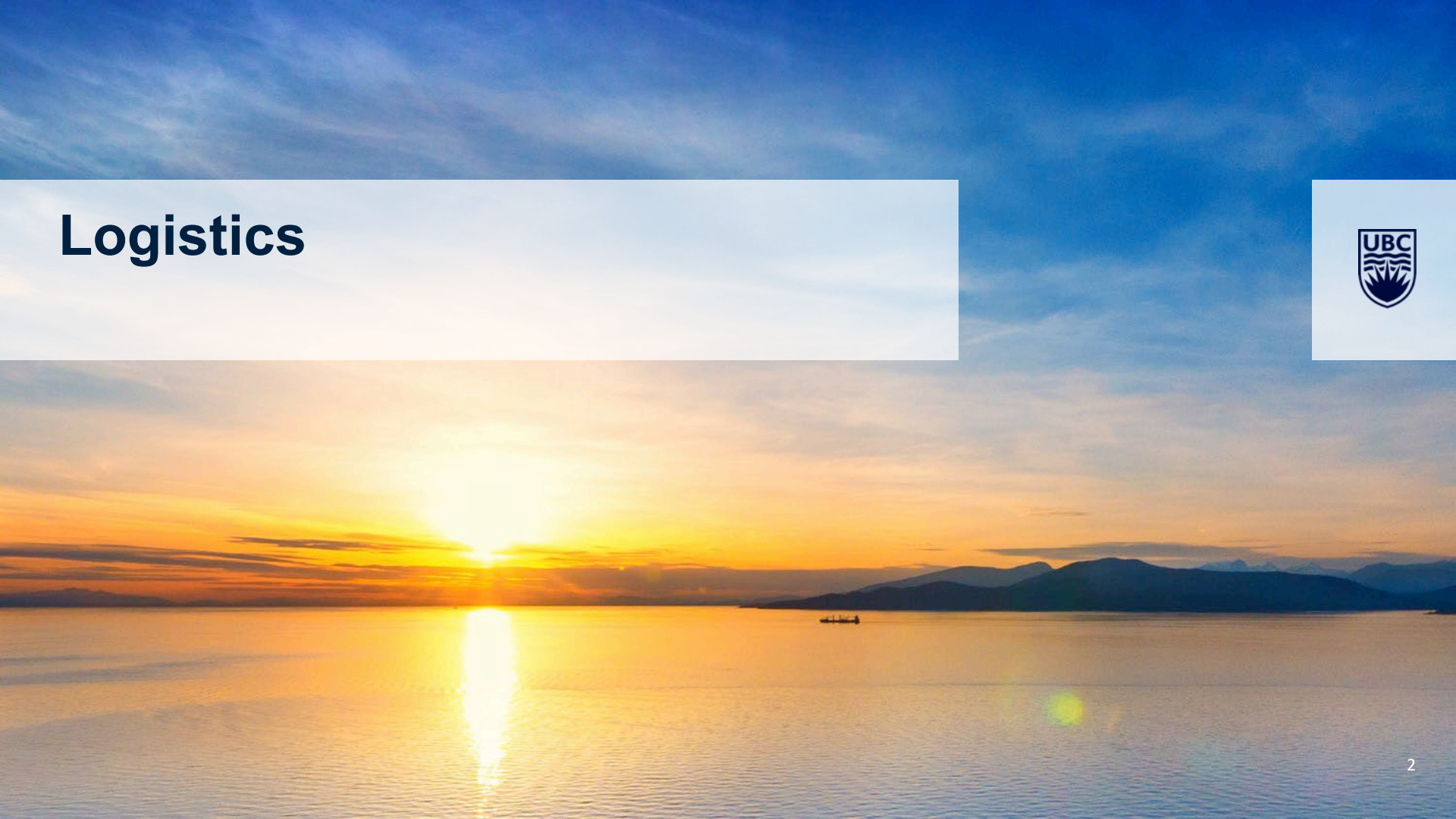
CPSC 416 Distributed Systems

Winter 2022 Term 2 (March 21, 2023)

Tony Mason (fsgeek@cs.ubc.ca), Lecturer

Logistics



Deadlines

Project 4 Released. Late Due: April 13, 2023.

Project 5 Released Due: April 13, 2023. **No extensions**.

All project work is due April 13, 2023. Late projects are scaled to 75% of the on-time max.

Final Exam: April 20, 2023, DMP 310, 08:30-11:00. Format TBA.

Deadlines

Alternate Path 1 & 2: Review in progress

- Piazza private threads need TLC
  - **Weekly updates due each Monday @ 23:59 PT**
- Final reports due no later than Thursday April 13, 2023 @ 23:59 PT
- Optional 10 min presentation April 13, 2023, up to 10 minutes.

Instructor Office Hours:

- Zoom Office Hours (Tuesday) @ 13:00-14:00
- Discord (Casual) Office Hours (Thursday) @ 14:00-15:00

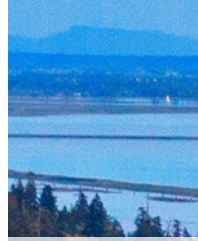
TA Office Hours:

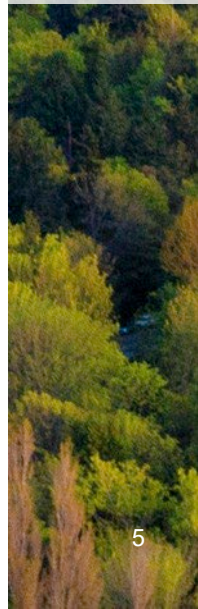
- Eric: Friday 9-11 am (in-person and Zoom)
- Japraj: Wednesday 3-5 pm (Zoom)
- Yennis: Thursday 2-4 (Zoom), Friday 2-4 (in-person)

Readings

Required:

Recommended:



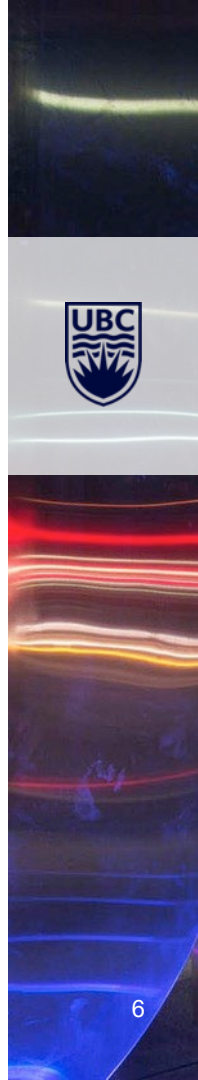


Questions?

Questions about the class?

Questions about the previous lecture?

Funny stories to share?



Today’s Failure

Gitlab.com Failure Example

Date: January 31, 2023 17:20 UTC

Event: Copied database from production to staging

19:00 UTC

Event: Increased Database load

Suspected cause: spam

Actual Issue(s):

- Users cannot post comments on issues/merge requests
- Background delete of Gitlab employee + data (due to accidental abuse flag)

Gitlab Failure

23:00 UTC

Event: Secondary replication process “falls behind”

Problem: **Primary** has already garbage collected log segments needed by secondary

Solution: Manually resynchronize primary and backup

- **Delete** the secondary backup
- **Copy** the primary to the secondary

Things get worse:

- Copy routine fails to start
- Copy routine blocks waiting for data from primary to secondary: no feedback
- Try to clean up the database directory on secondary
  - Oh no! Accidentally deleted it on the **primary**.

Gitlab Failure

Things got worse:

- They couldn’t find their backups in S3
- Turns out their backup process was using an **old** version of the backup tool.
  - It won’t backup a newer version of the database
  - Nobody noticed
  - Automatic cleaning of old backups had deleted **everything**.
- Azure disk snapshots **were not enabled for the database volumes**

Resolution:

- Restored from LVM snapshot
- Time to restore: around 18 hours

Gitlab Failure

Takeaways:

- Restore time is the **worst** time to figure out your backups didn't work
  - "Why was the backup procedure not tested on a regular basis? - Because there was no ownership, as a result nobody was responsible for testing this procedure."
- Redundancy is your friend
- Redundancy is *not* your accounting department's friend.

There is a complete write-up, including the DOS attack that led to the increased database load: [Postmortem of database outage of January 31 | GitLab]

Lesson Goals

Peer-to-Peer and Mobility

Tools for building distributed applications

Chord peer-to-peer system

Overlay networks for mobility

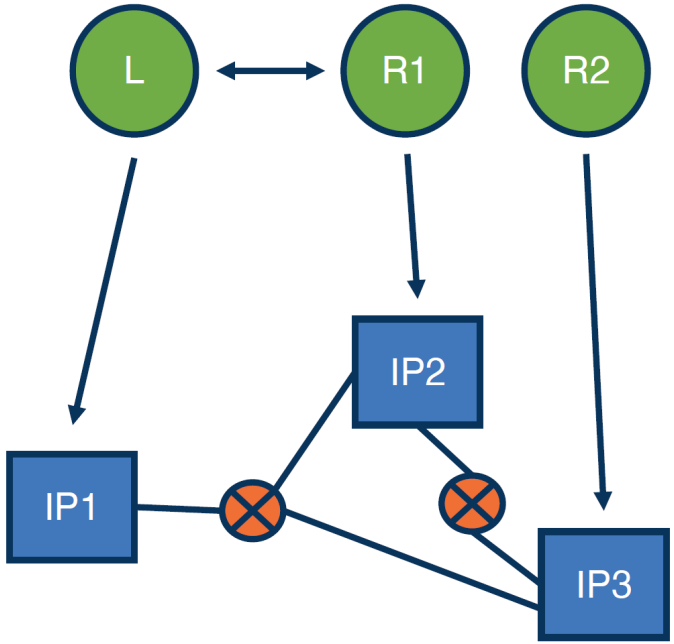
Network Abstraction

Application/service-level namespace

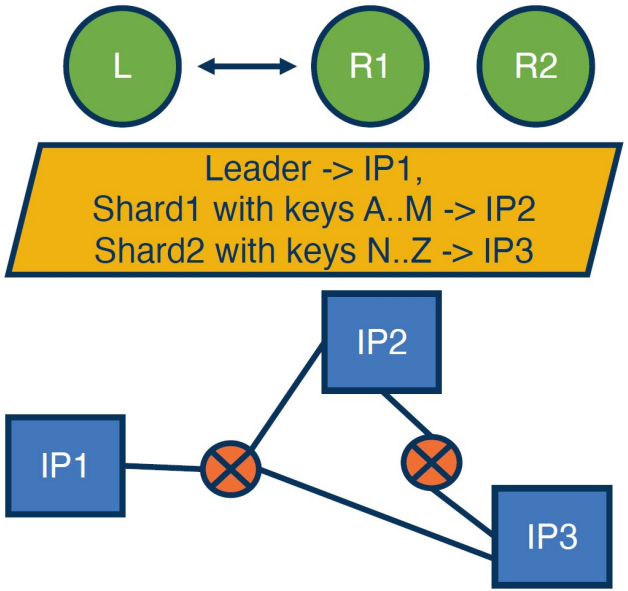
- Process names
- File names
- Object keys
- …

Network level

- IP addresses
- Network paths through switches & routers



Network abstraction



Metadata service

- Determines Overlay Network
- Part of control plane operation

Update on change

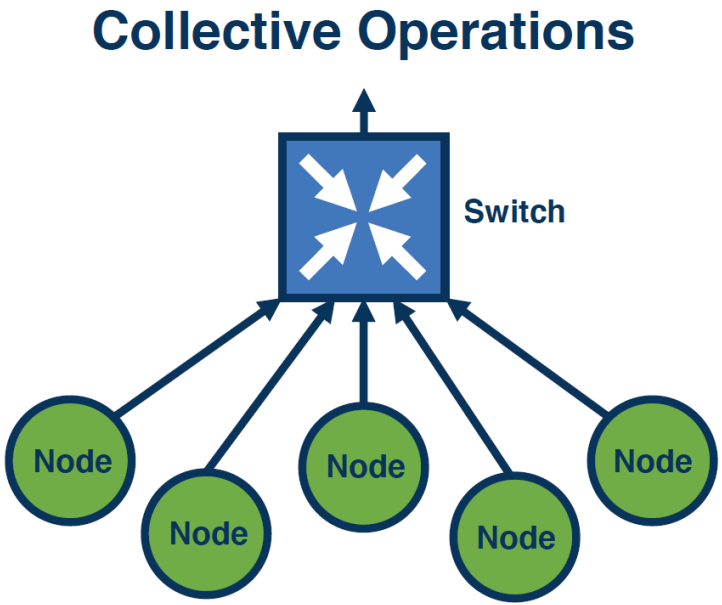
- Scale
- Geo-distribution
- Failures
- Multiple administrative domains

Interconnect Support

- Broadcast, multicast
- Gather/all-reduce
- Barrier
- Atomics (e.g., CAS)
- Timing
- RDMA (Remote Direct Memory Access)
- Direct cache injection (DDIO)

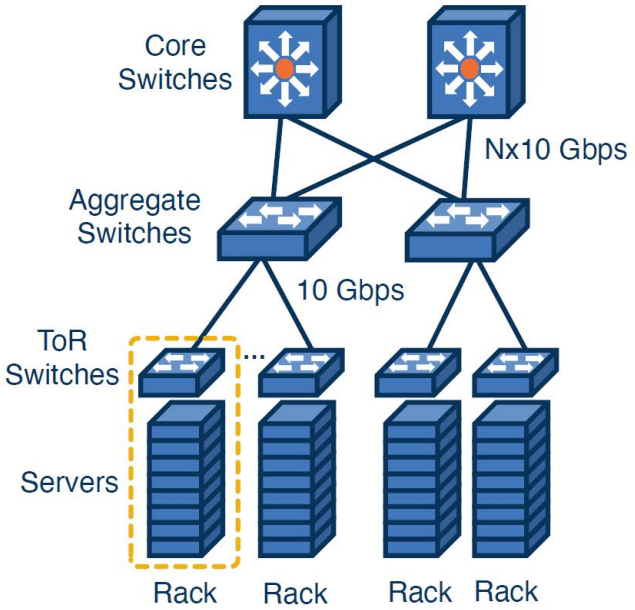
Hardware Scalable Implementations

- Separate dedicated networks
- Combining Tree Algorithms



Peer to Peer Systems

Datacenter Infrastructure



Wide-Area Distributed Infrastructure

Peer to Peer Systems

BitTorrent™

Tor

bitcoin

ethereum

Peer-to-peer Connectivity

How do you find the right peer?

Centralized Registry:

- Single round trip time (RTT) to find the peer IP
- Requires a centralized trusted authority

Example: Napster

Peer-to-peer connectivity

How to find the right peer?

Flood or Gossip based protocols

- No single point of failure
- No bound on lookup time

Examples:

- Gnutella
- Bitcoin

Peer-to-peer connectivity

How to find the right peer?

Provide Structured Routing Trees: Distributed Hash Table (DHT)

- Decentralized index
- Probabilistic bounded lookup time

Examples:

- Chord
- Kademlia
- Amazon DynamoDB

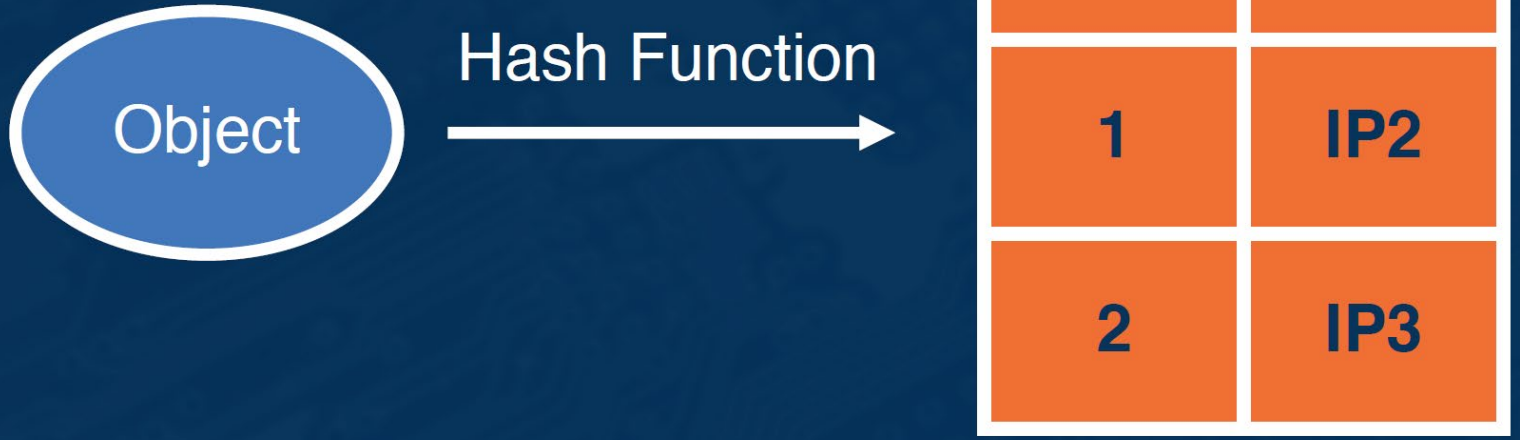
Distributed Hash Table

Hash Function:

- Maps a thing to a unique number within a range
- Key namespace to number namespace
  - File names
  - Song names

Uniform hash function use:

- Same mapping

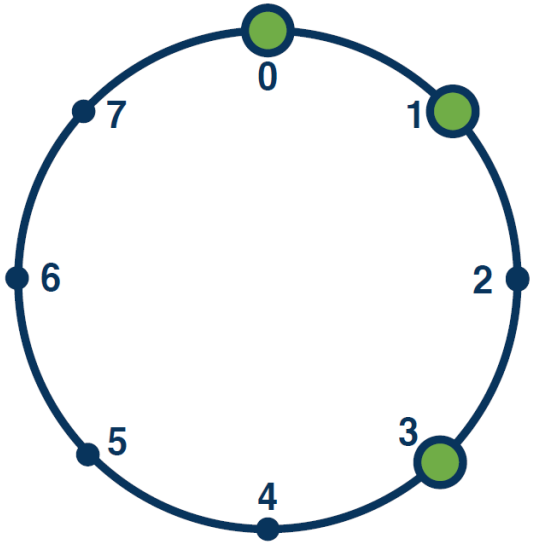

Object
Hash Function
0
IP1
1
IP2
2
IP3

Chord Distributed Hash Table Ring

Use cryptographic secure hash algorithm (SHA)

- Maps keys to a fixed length numeric value
- Maps IP addresses to a fixed length numeric value

Ring is N nodes {0,…N-1}

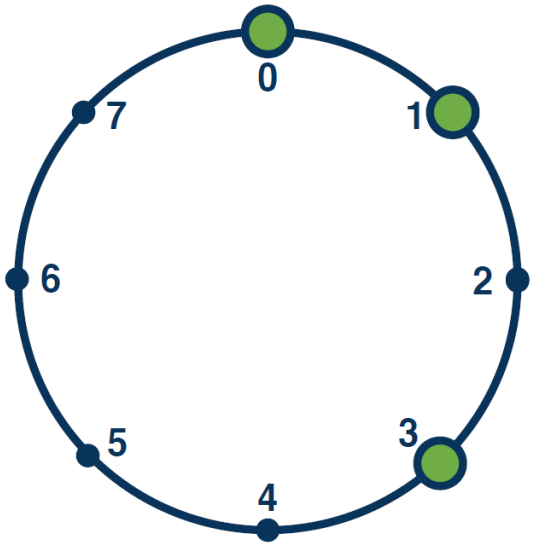


Insert Operation

SHA(key) = value

If node exists at value: update

Else: update successor node



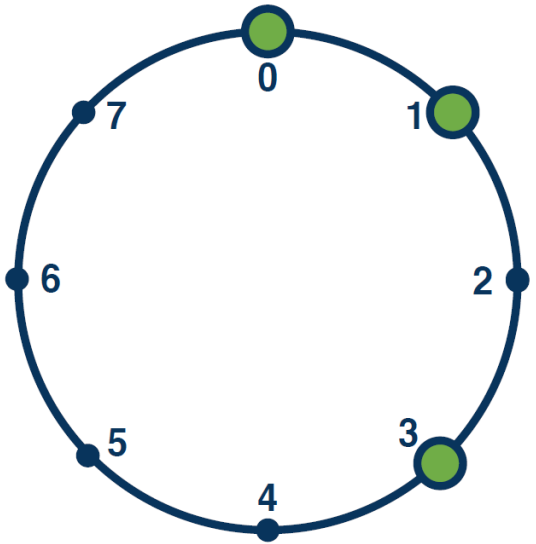
Lookup Operation

SHA(key) = value

If node exists at value: lookup

Else: lookup at successor node

Question: Can we improve over O(N)?



Finger Tables

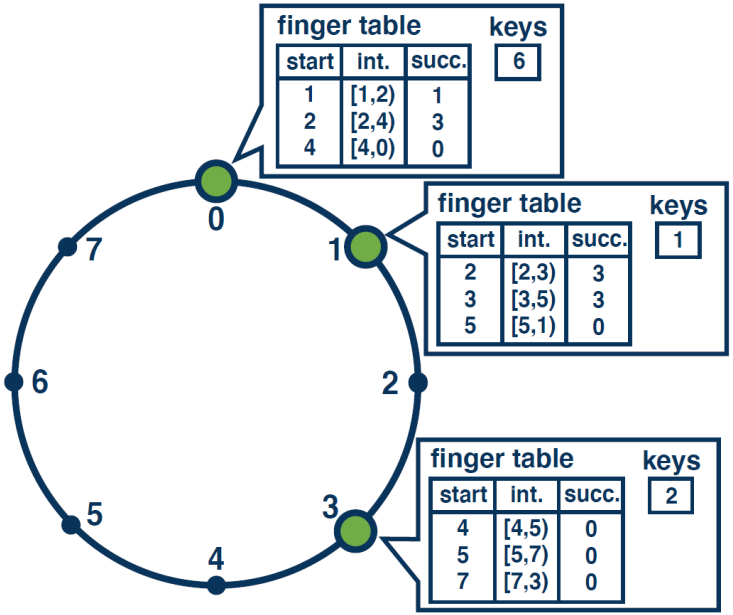
Finger Tables

Node ID for progressively longer ranges

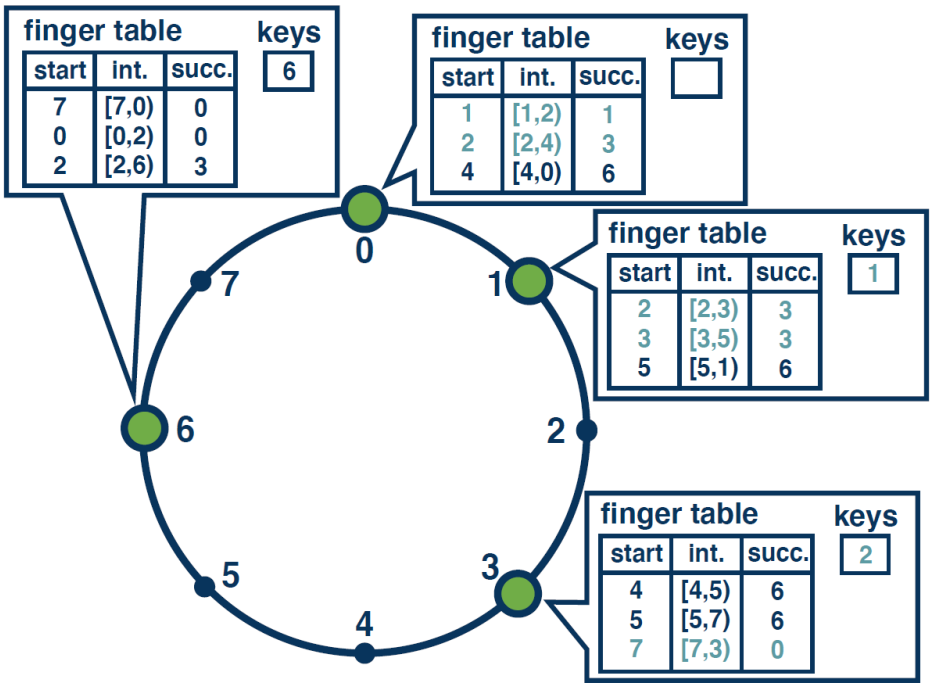
Finger table:

- At each node n
  - i-th finger entry starts at [n +2i]
  - For range of 2i elements

Lookup is O(log(N))



Chord: Managing the Ring



Nodes joining and departing

Redistributed data

Update finger tables

Improve performance with additional metadata

Probabilistic system performance guarantees

Hierarchical Designs

Cost of communications versus cost of overlay maintenance

Nodes with different properties:

- Point-to-point communication
- Stability, failure probability, mobility
- Number and type of nodes
- Communications patterns, locality

Hybrid approaches

- Large-scale datacenters
- Wide area
- Mobile networks

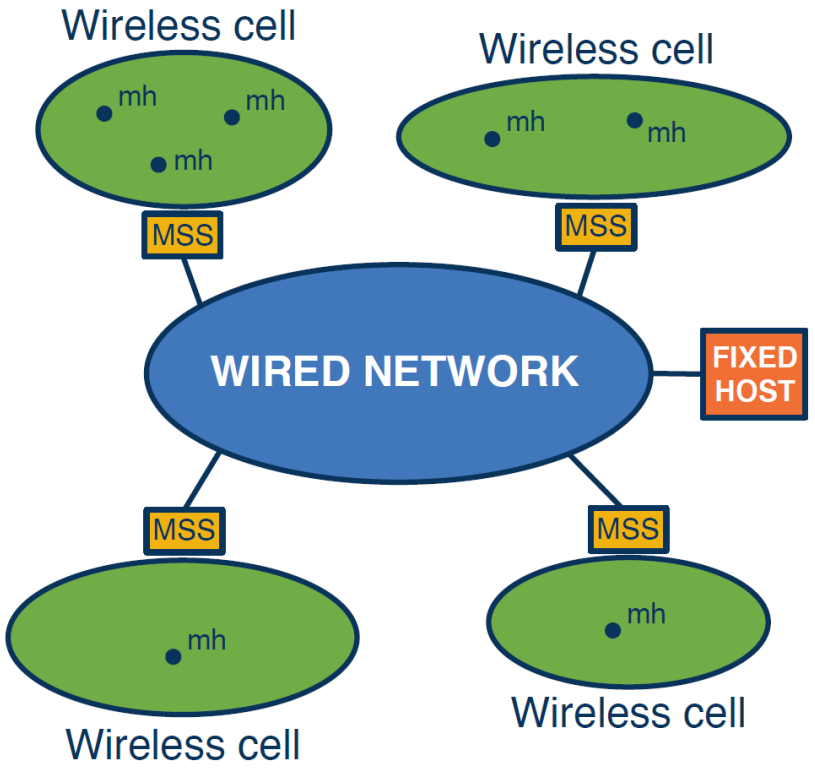
Mobile Network Model

Mobile Support Stations (MSS)

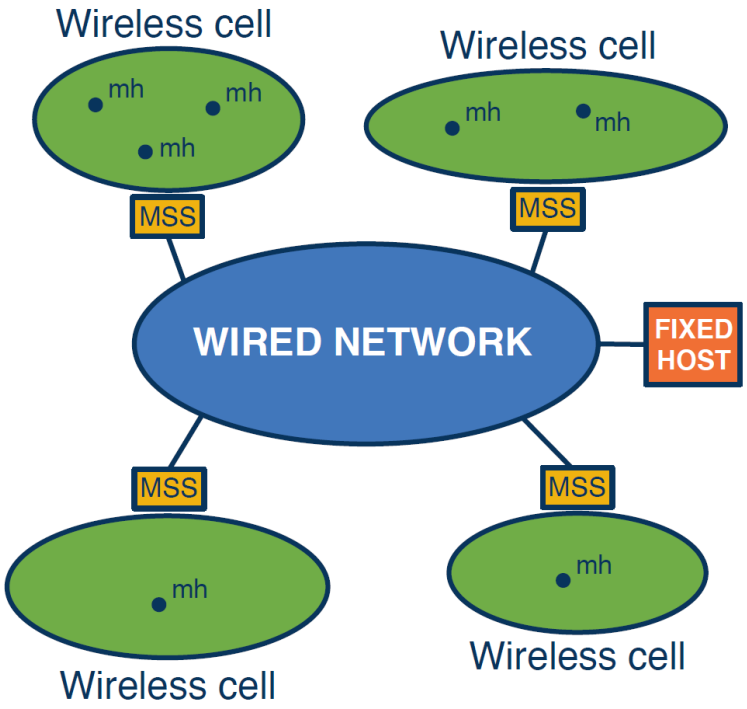
- Stationary
- High-speed wired network
- No power availability concerns

Mobile Hosts (MH)

- Associated with an MSS
- Mobile
- Lower speed mobile network
- Battery power concerns



Mobile Network Model



Goal:

- Fast lookup of MH
- Low overhead update of overlay state
  - Communications overhead
  - Battery/energy/compute overhead

Heterogenous nodes have *different* concerns

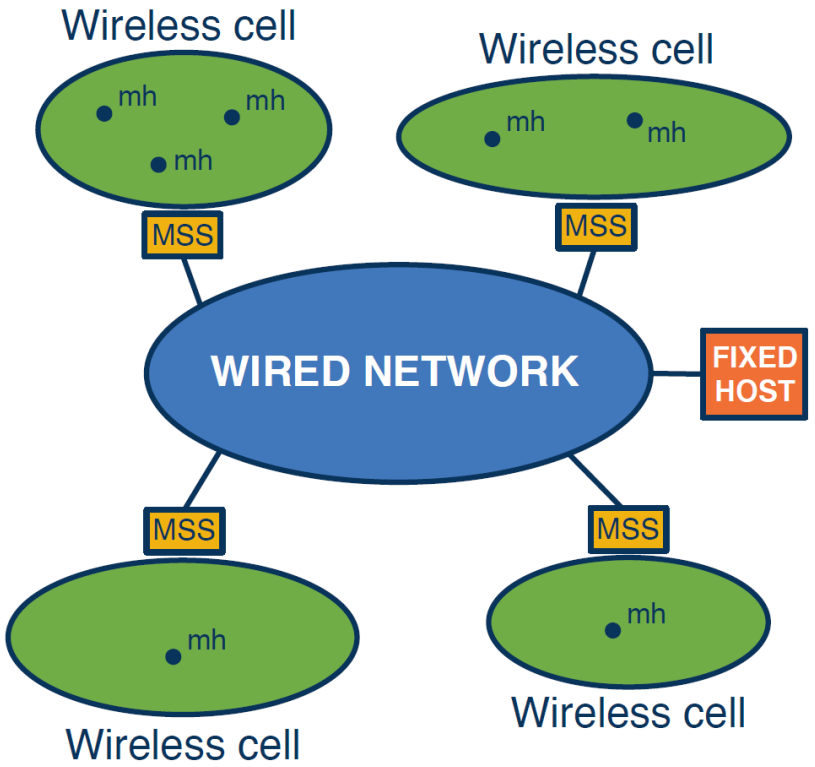
Different Algorithms

Metrics

- Search (lookup) cost
- Insert (add/remove) cost
- Mobility support update impact

Analysis Considerations

- $C_{wireless} >> C_{fixed}$
- $N_{mh} >> N_{mss}$



Communications Cost

$Cost_{communications} = 2 * Cost_{wireless} + Cost_{search}$

Algorithm 1:

- Logical ring of all mobile hosts
- $Cost_{search} \sim O(N_{mh}, Cost_{wireless})$

Algorithm 2:

- Two-tier hierarchical design
- Mobile Support Stations (MSS) in logical ring
- Each MSS knows about Mobile Hosts in its cell
- $Cost_{search} \sim O(N_{mss}, Cost_{fixed})$

Algorithm 2 is a clear winner here

Mobility Support Cost

Algorithm 1:

- Original MSS search for new MSS *on demand*
- No update on move, only when needs to reach MH

$Cost_{update}$ ~ O($Cost_{fixed\_search}$)

Algorithm 2:

- New MSS informs original MSS each time a new MH joins
- Update needed each time MH moves

$Cost_{update}$ ~ O(#moves * $Cost_{fixed}$)

Lesson Review

Network overlays are useful to properly route distributed systems messages

Peer-to-peer systems benefit from distributed name management (DHT)

Hierarchical design, heterogenous systems, and mobility require design for purpose

Questions?

How to use this template

Please note: This template has a variety of slides for your use. To select what slide you would like, click on the drop down menu beside “new slide” button in the top left corner, and pick the corresponding slide. To insert text, simply double click on the text box and start typing. Please be aware that copying and pasting text may change how the font looks. It is better to type directly onto the slide. Also note that larger fonts (size 14+) work better for presentations than smaller sizes. This template uses the font Arial, as PowerPoint users will experience technical difficulties if using UBC’s official fonts. If desired, images can replaced by going into the “Master” view and applying your own image. Please ensure you have the rights to an image before using it.

The following slides are here for visual reference only. Please delete or edit as needed for your own presentation. If you have any questions about how to use this template, please contact UBC Communications and Marketing at comm.marketing@ubc.ca

Insert title here

Insert subtitle here

Name, position

Insert title here

Insert subtitle here

Name, position

Insert title here

Insert subtitle here

Name, position

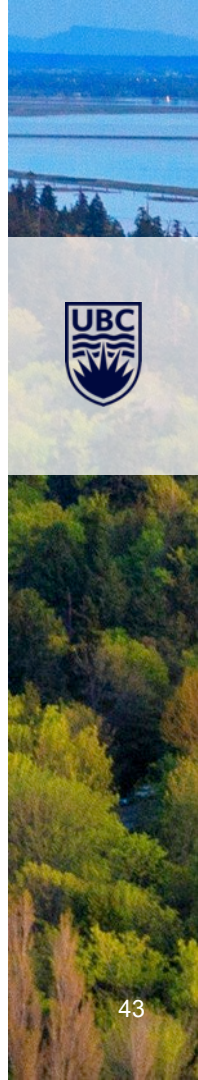
Insert title here

Insert subtitle here

Name, position

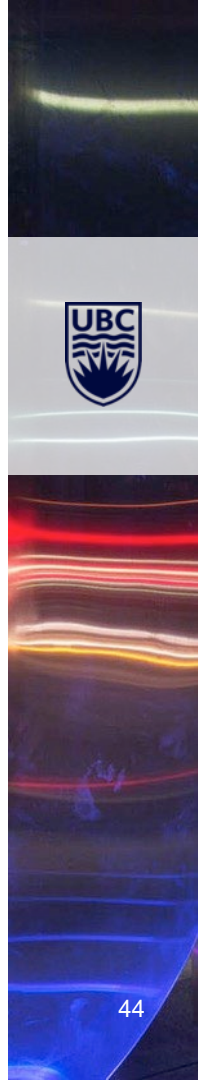
Page title

- **Bullet point** list
- **Bullet point** list
- **Bullet point** list
- **Bullet point** list



Page title

- **Bullet point** list
- **Bullet point** list
- **Bullet point** list
- **Bullet point** list



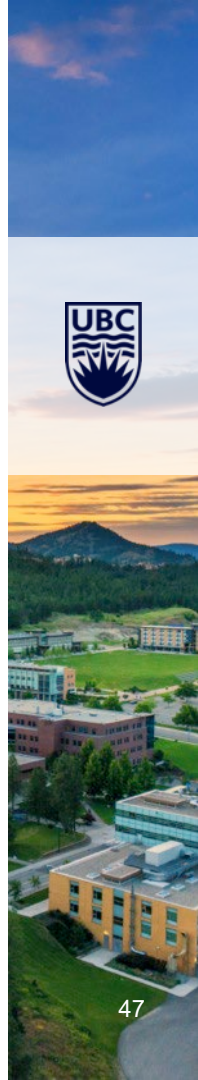
Insert chapter title

Page title

- Bullet point list
- Bullet point list
- Bullet point list
- Bullet point list

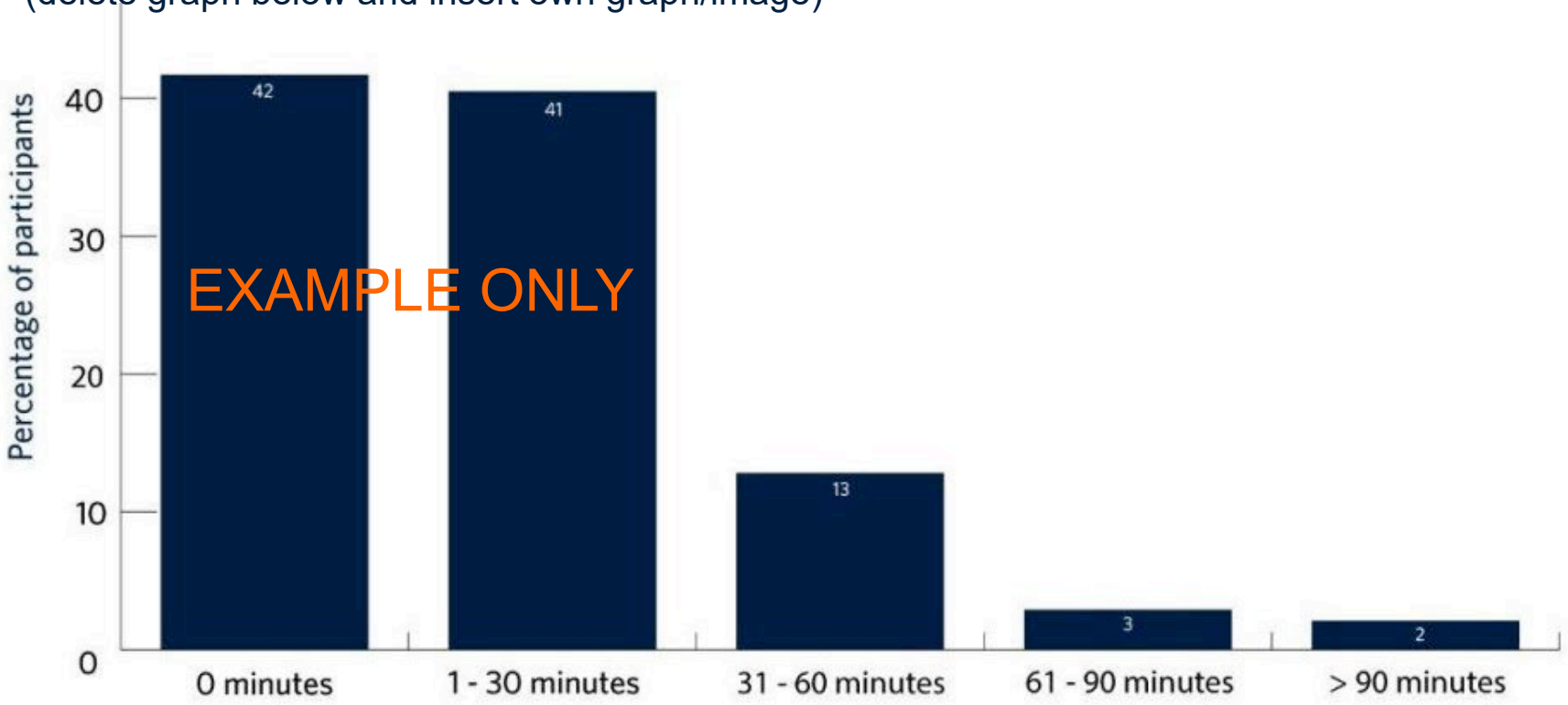
Page title

- **Bullet point** list
- **Bullet point** list
- **Bullet point** list
- **Bullet point** list



Insert title

(delete graph below and insert own graph/image)




UBC
THE UNIVERSITY OF BRITISH COLUMBIA


THE UNIVERSITY OF BRITISH COLUMBIA
THE UNIVERSITY OF BRITISH