BIG DATA MANAGEMENT

CE/CZ4123

OVERVIEW 2^ND HALF

Siqiang Luo

Assistant Professor

DISTRIBUTED SYSTEMS FOR BIG DATA: CHALLENGES

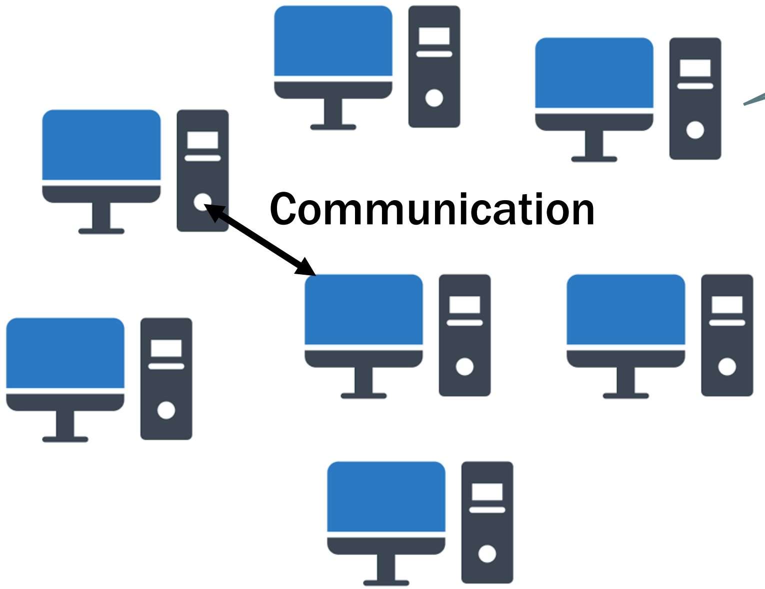
- ❑ How to organize the machines?
  - ❑ Fully-Distributed Mode
  - ❑ Master-Slave Mode
  - ❑ Fault-Tolerant

- ❑ How to store data across machines?
  - ❑ Data Partition
  - ❑ Data Replication

- ❑ How to compute using multiple machines?

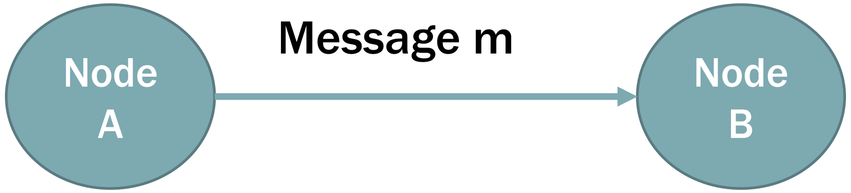
FULLY DISTRIBUTED MODELS

- Each machine has an IP address
- A knows machine B’s IP address: A can send messages to B
- Two machines can communicate with each other via IP address
  - i.e., sending messages between machines



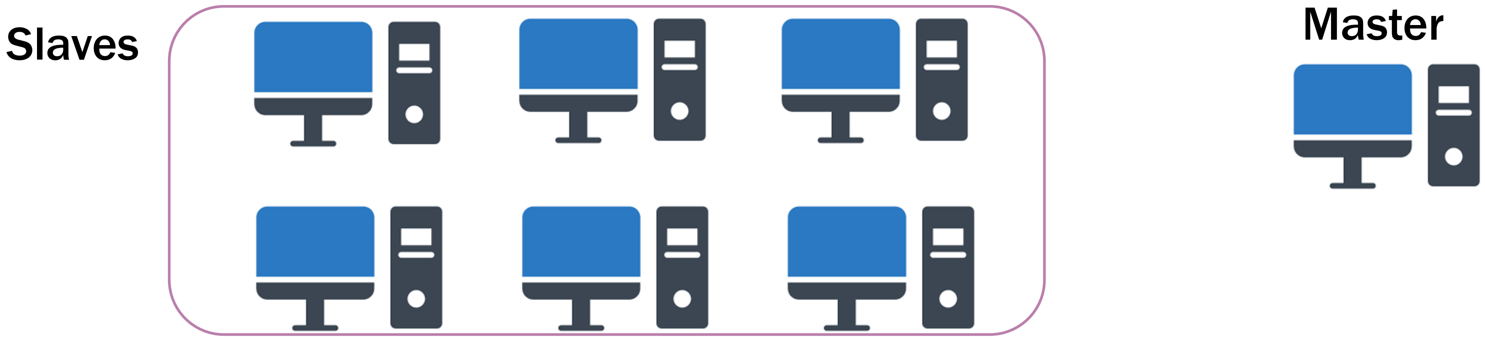
“node” or “machine”

Abstraction of communication



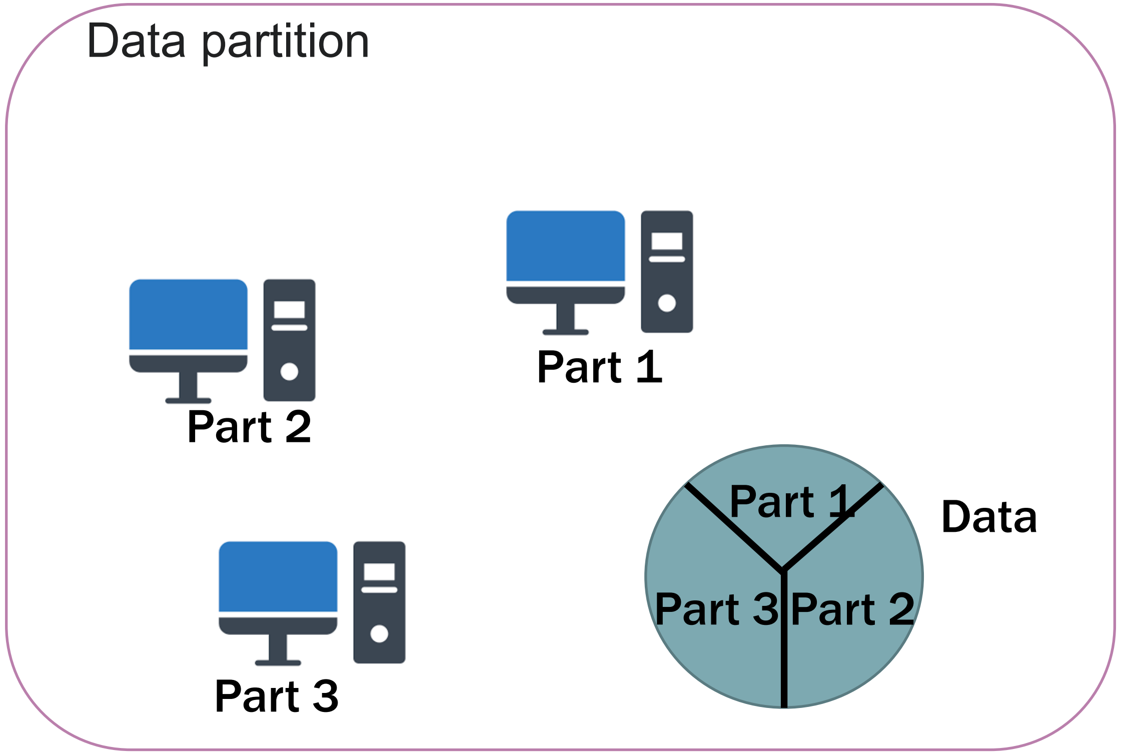
MASTER-SLAVE MODEL

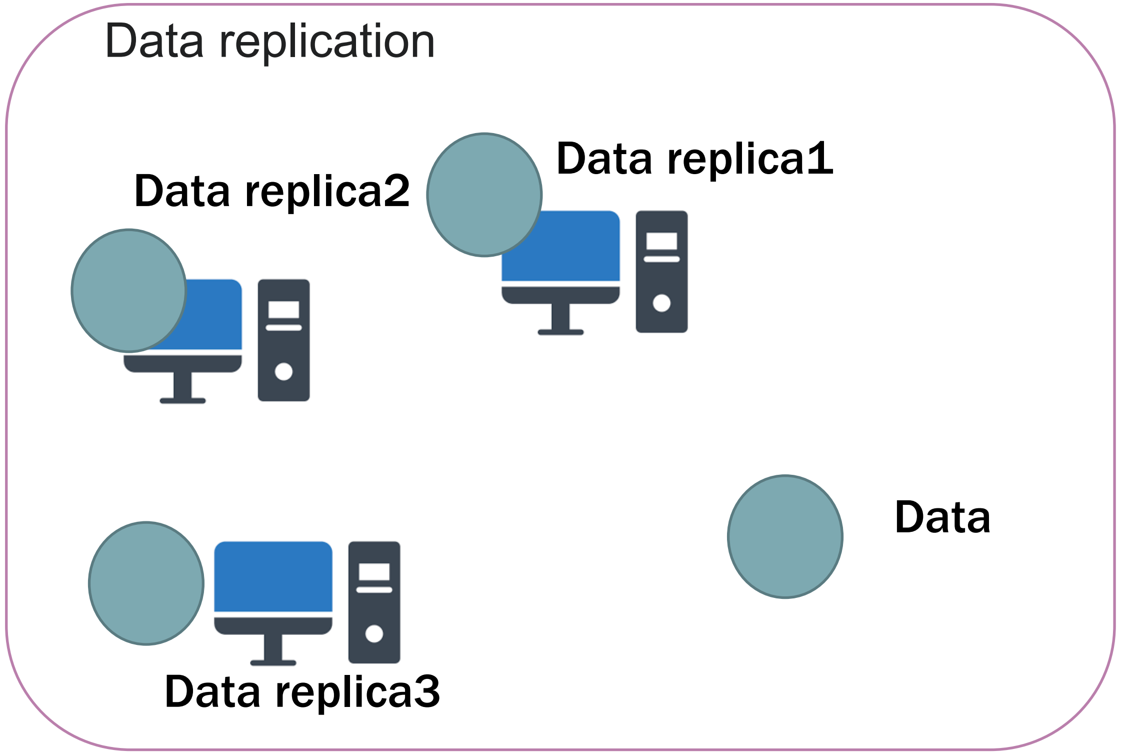
- Each machine has an IP address
- There is a machine called master, and the other machines are called slaves.
- Master is the coordinator, being responsible to
  - distribute tasks to the slaves, and
  - receive the results from the slaves



DATA PARTITION AND DATA REPLICATION

- ❑ Data partition: partition the data into different machines
- ❑ Data replication: each data item can be replicated to multiple copies.





MAPREDUCE

- ❑ Understand the basic model of MapReduce
  - ❑ Map function
  - ❑ Reduce function
  - ❑ Job
- ❑ Understand the execution workflow of MapReduce
  - ❑ Within a job, reduce function receives the pairs with the same intermediate key
- ❑ Know how to design algorithms (pseudo-code)
  - ❑ Wordcount
  - ❑ Table Join
  - ❑ Shortest distance
  - ❑ PageRank

NOSQL

- ❑ Property of NoSQL
  - ❑ Flexible schema (schemaless)
  - ❑ Easier to scale
  - ❑ Partially supports query language
  - ❑ Queries are less flexible, but can have higher performance
- ❑ Types of NoSQL Systems
  - ❑ Key-Value Stores
  - ❑ Wide-Column Database
  - ❑ Document Database
  - ❑ Graph Database

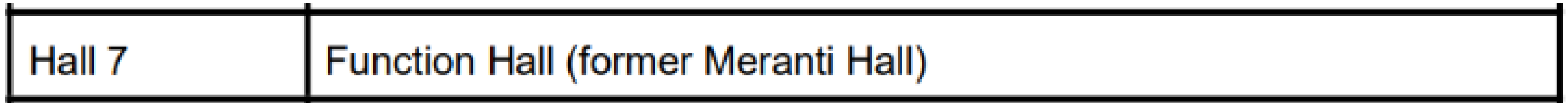
KEY-VALUE STORES

- ❑ LSM-tree
  - ❑ Get
  - ❑ Put
  - ❑ Delete
  - ❑ Fence Pointers
  - ❑ Bloom filters
  - ❑ FPR
  - ❑ I/O cost analysis
  - ❑ Tiering LSM-tree
  - ❑ Range Filter (Not in the scope of final exam)

FINAL EXAM TIME AND VENUE

❑ May 8 1pm-3pm (Come Early!)

❑ Hall 7



Hall 7	Function Hall (former Meranti Hall)

❑ Closed-Book

❑ Covers whole semester lectures (including those before quiz)

❑ Instructions to Examination Candidates
https://entuedu.sharepoint.com/sites/Student/dept/sasd/oas/Shared%20Documents/ExamAndAssessment/Exam/Instructions_to_candidates_physical_examinations_on-campus.pdf

The End
Thank you!