Apache Cassandra tour (lab exercise)

COSC430—Advanced Databases

David Eyers

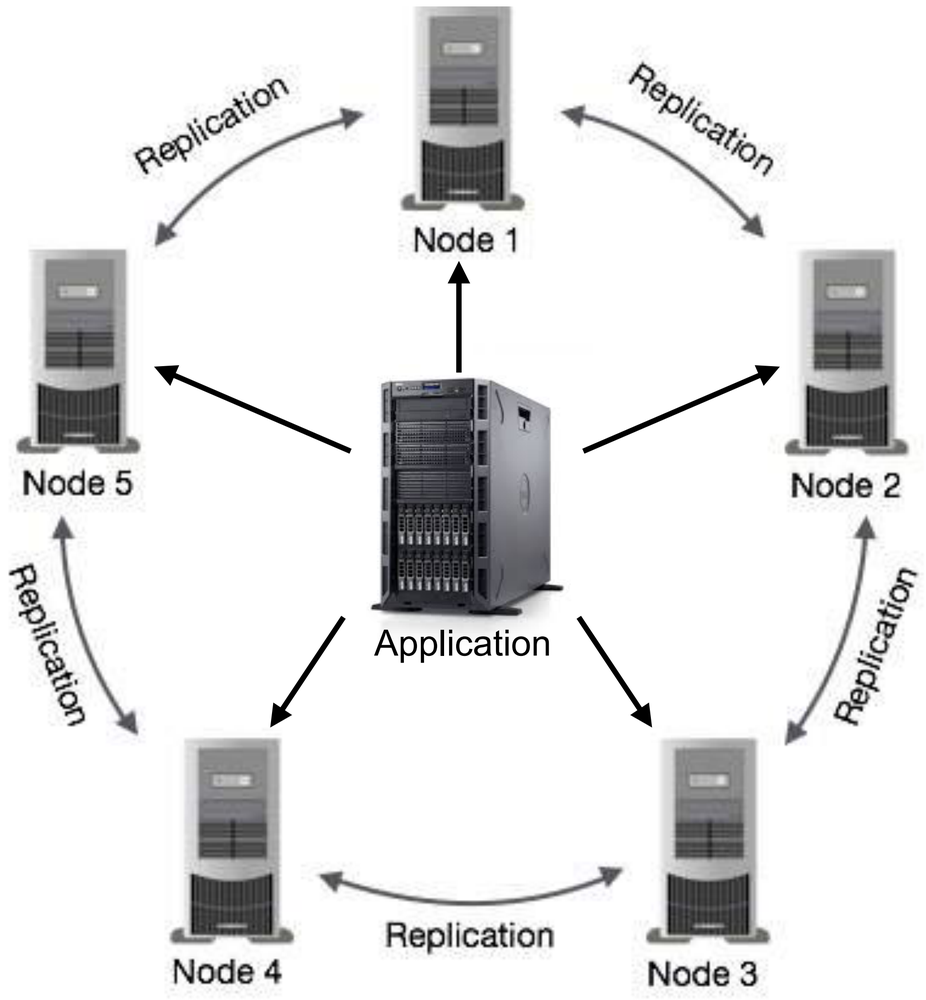
Learning objectives

- You should be able to
  - understand the architecture of Cassandra and its replication strategies
  - explain how a distributed database works using Cassandra as an example
  - understand the installation and configuration of Cassandra
  - understand why Cassandra can provide high availability with no single point failure
- There is no assessment for this lab

What is Apache Cassandra?

- Apache Cassandra is a free and open-source **distributed NoSQL** DBMS signed to handle vast amounts of data across **large clusters of commodity servers**, providing **high availability** with no single point of failure

Cassandra uses peer-to-peer architecture



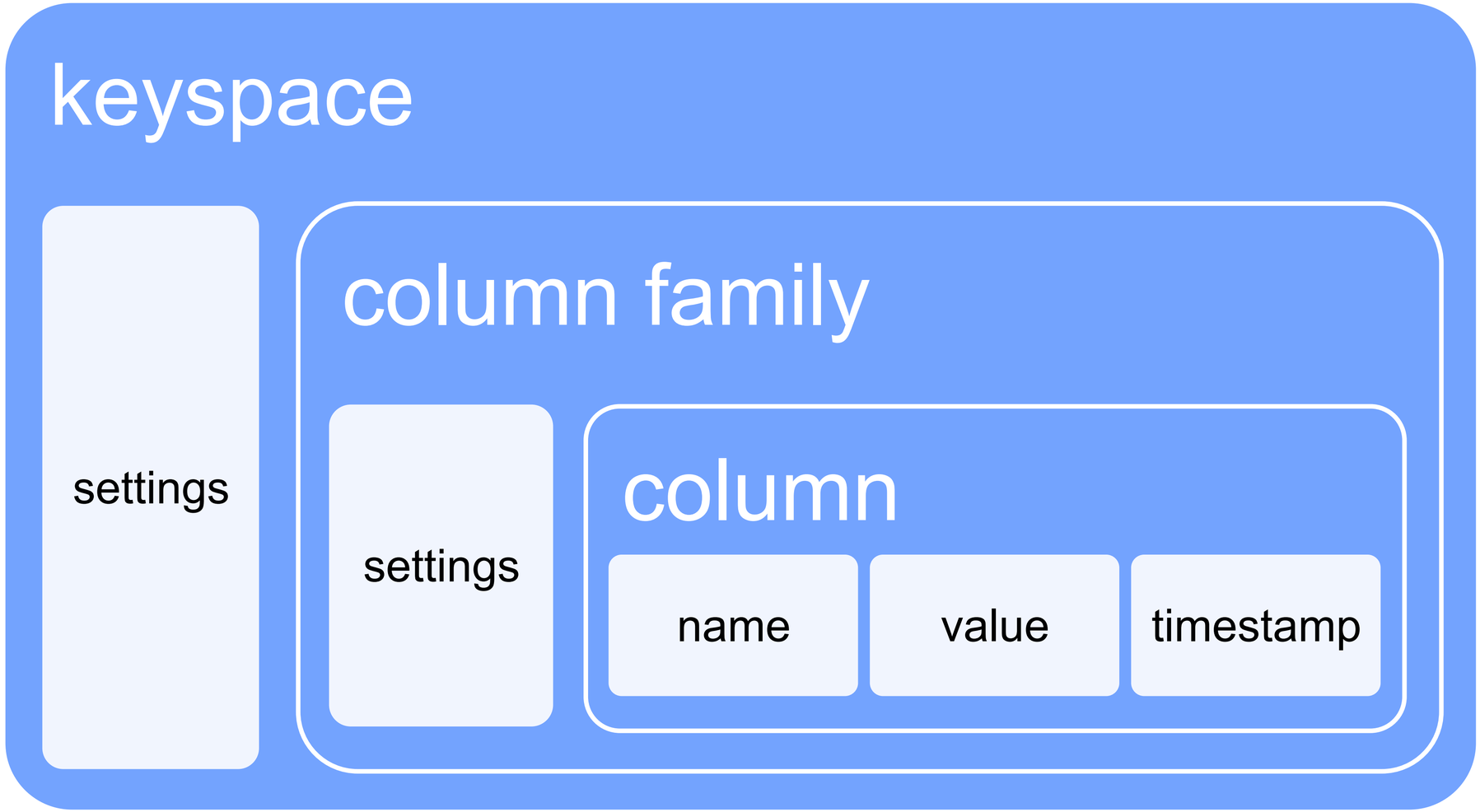
Elements in Cassandra:

- Cluster
  - Data center(s)
    - Rack(s)
      - Server(s)
        - Node(s)
- Uses a **gossip protocol** for communication between nodes
- Cassandra Query Language (CQL)—many similarities to SQL

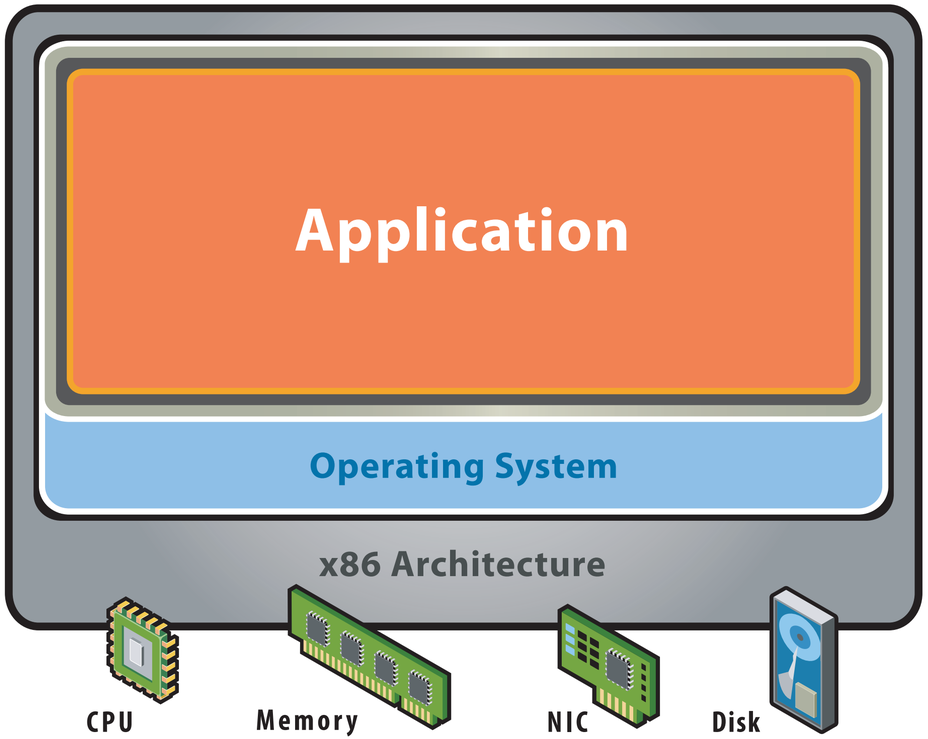
Data replication

- Nodes are **logically** structured in a ring topology
- Each data item replicated at N (replication factor) nodes
- Two replication strategies:
  - **SimpleStrategy**
    - use only for a single data centre and one rack
    - replicas are placed on the next node clockwise in the ring without considering topology (i.e., rack or datacenter location)
  - **NetworkTopologyStrategy**
    - cluster can be deployed across multiple data centres
    - attempts to place replicas on distinct racks because nodes in the same rack (or similar physical grouping) often fail at the same time

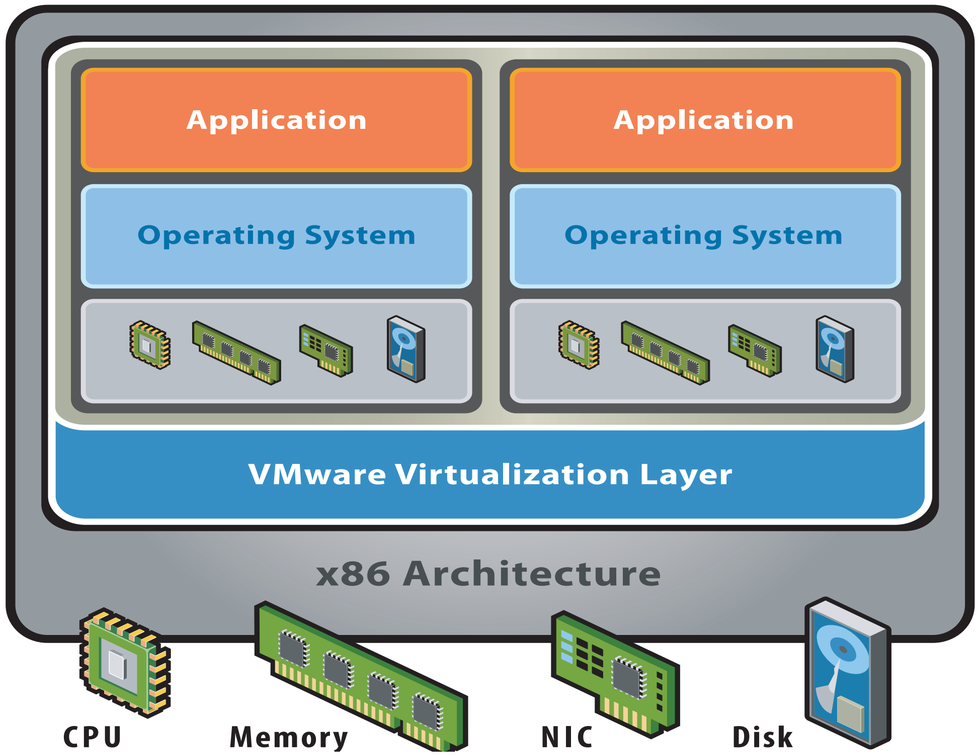
Apache Cassandra's data model



Virtualisation: abstracting over resources



- Single OS per machine
- Software and hardware tightly coupled
- Underutilised resources
- Inflexible and costly infrastructure



- Hardware independent of OS and applications
- Virtual machines to any system
- OS and application as a single unit into virtual

Docker and Vagrant

- Docker
  - Provides OS-level virtualisation, also known as containerisation
  - Package an application and its dependencies in a virtual container that can be installed and run on any Linux server
  - Lightweight—a single server or virtual machine can run a large number of containers simultaneously
- Vagrant
  - An open-source software platform for managing virtual software development environments
  - Vagrant sits as a layer over the top of virtualisation software

Apache Cassandra lab exercise

- You can view a formatted version of the Markdown file containing the instructions at the following URL: https://altitude.otago.ac.nz/cosc430/cassandra-intro/-/blob/master/README.md

- In the past a PDF version of the instructions was provided, however some people's PDF viewers were copy/pasting commands with extra spaces, so I have removed the PDF version