

Distributed Systems

Christian J. Rudder

January 2025

Contents

Contents	1
1 Introduction	6
1.1 High-level Computer Architecture Overview	6
1.1.1 System Review	6
1.1.2 CPU and Memory Orchestration Review	7
1.1.3 Motivation for Distributed Systems	12
1.2 Understanding RPCs & Synchronization with Go	13
1.2.1 Establishing a Client-Server Connection	13
1.2.2 Asynchronous Function Calls	17
1.2.3 Synchronization: Data Races & Deadlocks	24
1.2.4 References & Pointers in Go	27
1.2.5 Waiting for Goroutines to Finish	29
1.2.6 Sending Messages Between Goroutines	30
1.2.7 Task, Data, and Pipeline Parallelism	34
1.2.8 Arrays & Slices in Go	36
1.2.9 Repeating Tasks: Tick and Ticker in Go	38
1.2.10 Conditionally Reading from Channels: Select in Go	39
1.3 Time, Clocks, and Logical Ordering	40
1.3.1 Accuracy of Time: Atomic Clocks & NTP	40
1.3.2 Logical Clocks: Lamport & Vector Clocks	42
1.4 Implementing RPCs with Go	48
1.4.1 Typing in Go	48
1.4.2 Go's RPC Package	50
2 Working with Distributed Systems	53
2.1 Saving System State: Snapshots	53
2.2 Replication: Synchronizing State	58
2.3 Raft Protocol: Consensus Replication	64

2.3.1	Procedure Outline: Heartbeats, Elections, & Log Replication	65
2.3.2	Safety: Restricting Leader Election	70
2.3.3	AppendEntries, State, and RequestVote RPC Schema	71
2.3.4	Cluster Reconfiguration (Adding, Removing, and Replacing Servers)	75
2.3.5	Log Compaction & Snapshotting	76
2.3.6	InstallSnapshot RPC Schema	78
2.3.7	Raft Algorithm Paper	79
2.4	Failure Models	80
2.4.1	Defining Failures	80
2.4.2	Failures Model Hierarchy	82
2.5	Consistency Models	83
2.5.1	Introduction	83
2.5.2	Strong Consistency Models: Linearizability & Sequential Consistency	84
2.5.3	Handling Shared Data via Mutex: Release & Lazy-release Consistency	89
2.5.4	Weak Consistency Models: Causal & Eventual Consistency	90
	Bibliography	93

This page is left intentionally blank.

Big thanks to **Professor Ioannis Liagouris**
and **Dr. Anna Arpacı-Dusseau** for teaching CS351: Distributed Systems
at Boston University [1].

All illustration contain original assets.

*Disclaimer: These notes are my personal understanding and interpretation of the course material.
They are not officially endorsed by the instructor or the university. Please use them as a
supplementary resource and refer to the official course materials for accurate information.*

Prerequisites

This text assumes the reader has a basic understanding of computer science and programming. It will also assume they are somewhat familiar with computer architecture and operating systems at a high level. The text will review these concepts briefly for completeness, but it will not try to teach them from scratch or provide a full understanding of these topics.

The main focus will be on distributed systems, and will touch on:

- **Concurrency and Parallelism**
 - Concurrency, Parallelism, Threads
- **Consistency and Fault Tolerance**
 - Consistency, Fault-tolerance, Atomicity
- **Distributed Systems and Coordination**
 - Asynchrony, Coordination, Logical Time, Snapshots
- **Consensus Algorithms**
 - Raft, Paxos, Consensus
- **Replication and Data Management**
 - Replication, Sharding, Cluster
- **Protocols and Computing Models**
 - RPC, 2PC, Broadcast
- **Technologies and Tools**
 - MapReduce, Spanner, Dynamo, GFS, TLA+, Golang

Introduction

1.1 High-level Computer Architecture Overview

1.1.1 System Review

To understand distributed systems, we must first review the architecture of a single computer.

Definition 1.1: Turing Machine

Conceptualized by Alan Turing in 1936, a Turing machine is a mathematical model of computation that defines an abstract machine that manipulates symbols on a strip of tape according to a table of rules. Despite its simplicity, the machine can simulate the logic of any computer algorithm.

Definition 1.2: Von Neumann Architecture

The Von Neumann architecture, also known as the Princeton architecture, is a design architecture for an electronic digital computer with these components:

- **A processing unit** that contains an arithmetic logic unit and a control unit.
- **A memory unit** that stores data and instructions.
- **Input and output mechanisms.**

Fast forward, modern computers have the following components:

Definition 1.3: Modern Computer Components

- **CPU:** Central Processing Unit. The brain of the computer that performs instructions.
- **Memory:** Stores data and instructions.
- **Storage:** Hard drives, SSDs, etc.
- **Network Interface:** Connects the computer to the network.
- **Input/Output Devices (I/O):** Keyboard, mouse, monitor, etc.
- **Motherboard:** The central printed circuit board that interconnects all of the computer's components, including the CPU, storage devices, and I/O devices.

Before diving deeper into the inner workings of a single computer, let's define a distributed system:

Definition 1.4: Distributed System

A distributed system is a system whose components are located on different networked computers, which communicate and coordinate their actions by passing messages to one another. The components interact with one another in order to achieve a common goal.

In the words of Andrew S. Tanenbaum,

"A set of nodes, connected by a network, which appear to its users as a single coherent system."

or in the words of Leslie Lamport,

A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable.

Tip: **Andrew S. Tanenbaum** is a computer scientist and professor emeritus at the Vrije Universiteit Amsterdam in the Netherlands who is best known for his books on computer science. **Leslie Lamport** is an American computer scientist known for his work in distributed systems and as the initial developer of the document preparation system **L^AT_EX**.

1.1.2 CPU and Memory Orchestration Review

Now at a high-lever, we discuss how the a system interacts with all its components to perform tasks.

Definition 1.5: CPU (Central Processing Unit)

The CPU is made of the following components:

- **ALU (Arithmetic Logic Unit):** Performs arithmetic and logical operations.
- **Control Unit:** Manages the execution of instructions.
- **Registers:** Small, fast storage locations in the CPU that temporarily hold data and instructions.

Definition 1.6: Memory Segments

A program's memory is typically divided into several segments:

- **Text Segment:** Contains the executable code.
- **Data Segment:** Stores global and static variables.
- **System Stack:** A memory region that manages temporary data related to function calls in a first-in-last-out manner.
- **System Heap:** A memory region that dynamically allocates references to data from the stack memory.

Definition 1.7: Instruction Execution Cycle

The instruction execution cycle, also known as the *fetch-decode-execute* cycle, is the process by which the CPU processes instructions. In each cycle:

1. **Fetch:** The CPU retrieves an instruction from memory using the *instruction pointer* (or program counter).
2. **Decode:** The instruction is interpreted to determine what action is required.
3. **Execute:** The CPU performs the instruction's operation, which may involve arithmetic calculations, memory accesses, or I/O operations.

The CPU performs instructions via the following steps:

Definition 1.8: CPU Registers

Registers are small, high-speed storage locations within the CPU that temporarily hold data, instructions, and control information. Key registers include:

- **Instruction Pointer (Program Counter):** Holds **addresses**, which are the locations of the next instruction to fetch.
- **Stack Pointer:** Points to the top of the current stack in memory.
- **General-Purpose Registers:** Used for arithmetic operations and temporary data storage.

Definition 1.9: RAM and Volatile Memory

RAM (Random Access Memory) is a type of volatile memory used to store data and instructions that are actively used by the CPU. Since it is volatile, its contents are lost when the computer is powered off.

Definition 1.10: Physical Storage and I/O Devices

Physical storage refers to non-volatile memory devices such as hard drives and SSDs, which retain data without power. Many of these devices are accessed via input/output (I/O) operations and are thus considered part of the system's I/O mechanism.

Definition 1.11: Virtual Memory and Address Translation

Virtual memory is a memory management technique that provides an abstraction of a large, contiguous memory space. It works by mapping virtual addresses used by programs to physical addresses in RAM via structures such as page tables, which are managed by the Memory Management Unit (MMU).

Definition 1.12: CPU Cores

A CPU core is a physical processing unit within a central processing unit (CPU) responsible for executing instructions and performing computations. Modern CPUs often contain multiple cores, enabling them to handle multiple tasks at the same time.

Definition 1.13: Task, Job, and Process

- A **Task** is a single unit of work in various states (waiting, running, completed).
- A **Job** is a high-level operation comprising multiple tasks.
- A **Process** is an executing instance of a program that manages system resources instructing the CPU to execute tasks.

Definition 1.14: Threads: Concurrency & Parallelism

A **thread** is a unit of logic (a segment of code) to be executed by a CPU core. A **core can only run one thread at a time**. The core itself is called the **hardware-thread**, while our units of logic are called **software-threads** or **OS-threads**.

The OS system scheduler manages the hardware-threads, and assigns software-threads to them. Switching between software-threads on a hardware-thread is called **context switching**. Context switching is expensive, as it requires saving the current state of the software-thread and loading the state of the new software-thread. Though it provides the illusion of tasks running simultaneously, called **concurrency**.

With multiple cores come **multi-threading**, where multiple threads run on other cores simultaneously. This true simultaneity is called **parallelism**.

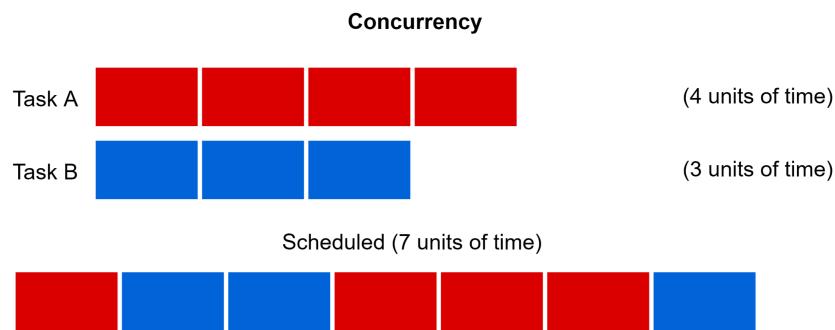


Figure 1.1: Concurrency: Multiple software-threads running on a single hardware-thread.

In reality, many tasks perform I/O operations, which don't concern the CPU:

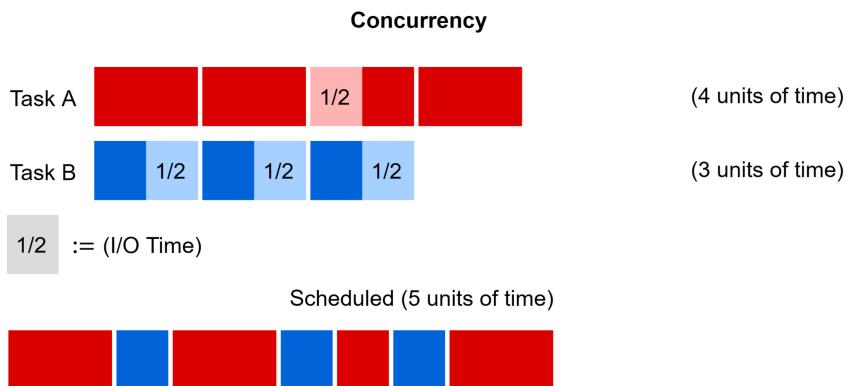


Figure 1.2: Concurrency with I/O: Multiple software-threads running on a single hardware-thread.

Now since the CPU isn't idle on I/O operations, the overall time between tasks is cut significantly.

Definition 1.15: Kernel

The kernel is the central component of the operating system. It manages hardware resources—including the CPU, memory, and I/O devices—and provides core services such as process management, memory management, and device control.

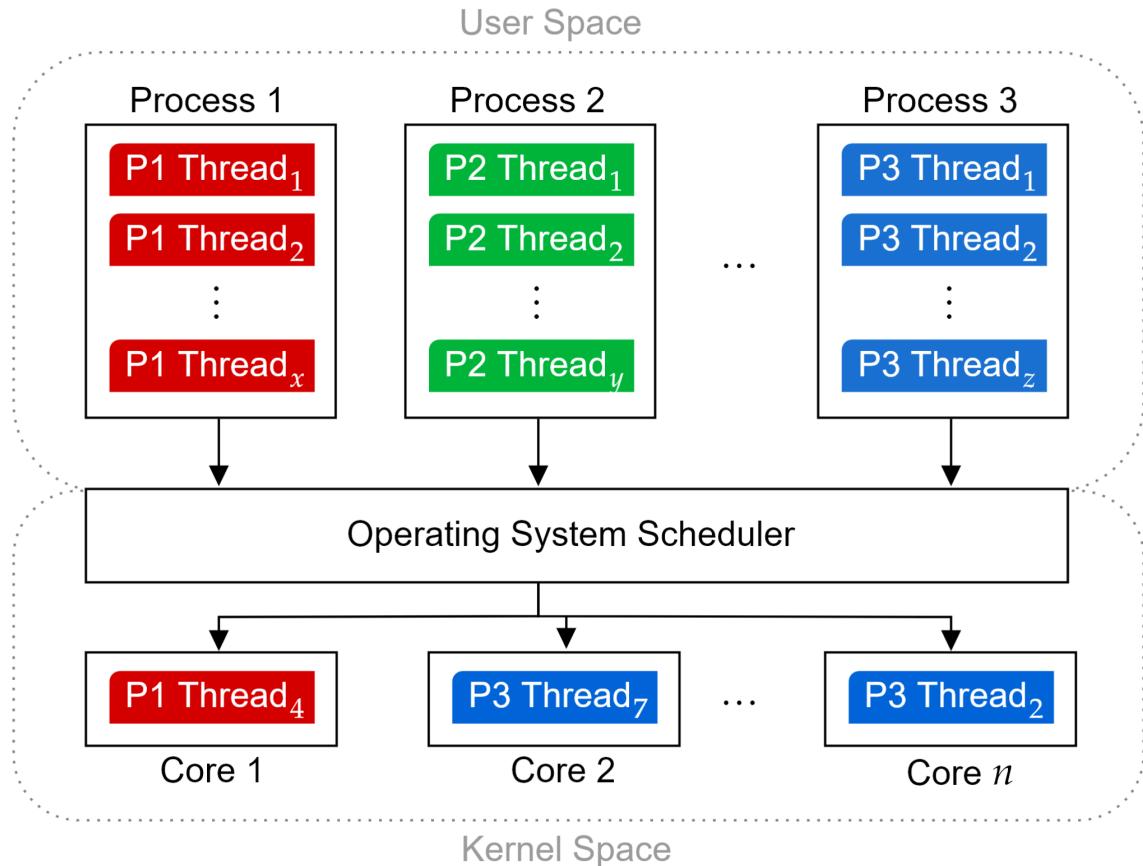


Figure 1.3: Process threads scheduled by the OS kernel and processed by the CPU.

In summary, what we need to know are these key points:

- The CPU executes instructions via processing units called cores/hardware-threads.
- The OS schedules software-threads from processes to run on hardware-threads.
- A core can only run one software-thread at a time, but can switch between them.
- Context switching on a single core is called concurrency, while utilizing multiple cores in unison (multi-threading) is called parallelism.

1.1.3 Motivation for Distributed Systems

Distributed systems cover a vast and diverse range of applications, including:

- **Offloading Computation:** Perhaps a system A offloads a heavy computation to system B .
- **Fault Tolerance:** If a critical system A fails, an almost identical system B can take over.
- **Load Balancing:** Say a system A is overwhelmed with requests, it can distribute the load to system B , acting as one system, from the requests point of view.

In today's market, there are numerous applications of distributed systems, such as: Cloud Computing, Social Networks, E-commerce, Streaming Services, Search Engines, Renting Computation (AI training), etc.

Let's begin to define the problem space. Say there are two individuals Alice and Bob, who wish to communicate:



Figure 1.4: Alice sends a letter m_1 overseas to Bob.

How does Alice know that her message m_1 was received by Bob? Bob would have to send a message back to Alice, acknowledging the receipt of m_1 .

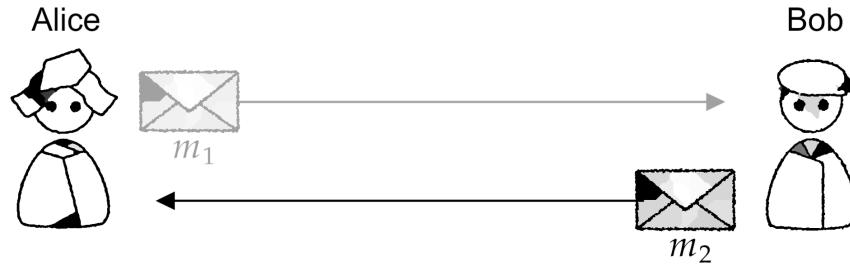


Figure 1.5: Bob sends an acknowledgment letter back to Alice.

Though problems can arise, what if Alice's letter gets lost in the mail, what if Bob receives multiple letters from Alice, how does Bob know which letter is the most recent? These are the fundamental problems of distributed systems.

1.2 Understanding RPCs & Synchronization with Go

This section will cover the concept of Remote Procedure Calls (RPCs) and how they are used in distributed systems and use the Go programming language to demonstrate such.

1.2.1 Establishing a Client-Server Connection

Definition 2.1: client-server model

The client-server model is a distributed application structure that partitions tasks or workloads between the providers of a resource or service, **called servers**, and service requesters, **called clients**.

Often clients and servers communicate over a computer network on separate hardware, but both client and server may reside in the same system.

Definition 2.2: Remote Procedure Call (RPC)

A Remote Procedure Call (RPC) is a protocol that allows a **client** computer request the execution of functions on a separate **server** computer.

RPC's abstract the network communication between the client and server enabling developers to write programs that may run on different machines, but appear to run locally.

Definition 2.3: RPC Call Stack

The RPC call stack facilitates communication between two systems via four layers:

1. **Application Layer:** The highest layer where the client application initiates a function call. On the server side, this layer corresponds to the service handling the request.
2. **Stub:** A client-side stub acts as a proxy for the remote function, **marshaling arguments** (converting them into a transmittable format) and forwarding them to the RPC library. On the server side, a corresponding stub, **the dispatcher**, receives the request, unmarshals the data, and passes it to the actual function.
3. **RPC Library:** The RPC runtime that manages communication between the client and server, ensuring request formatting, serialization, and deserialization.
4. **OS & Networking Layer:** The lowest layer, responsible for transmitting RPC request and response messages over the network using underlying transport protocols.

The request message travels from the client's application layer down through the stack and across the network to the server. The server processes the request in reverse, executing the function and returning the result to the client.

To illustrate the RPC call stack, observe the following diagram:

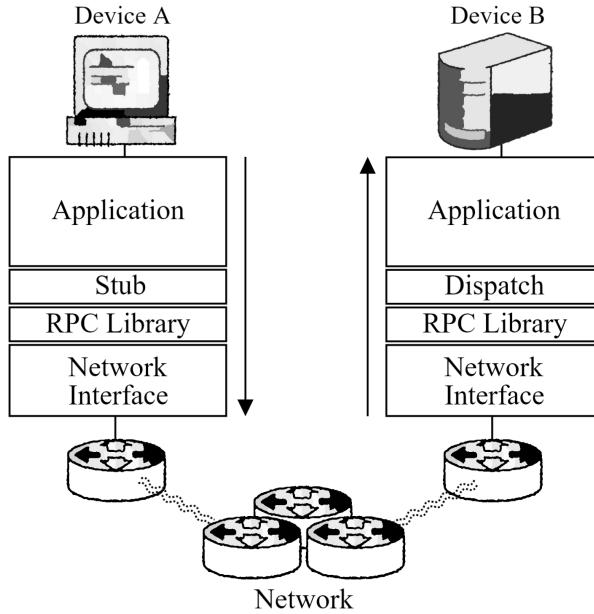


Figure 1.6: Client system *A* making a request to Server system *B* over RPC. This is reciprocated by *B* to return the result.

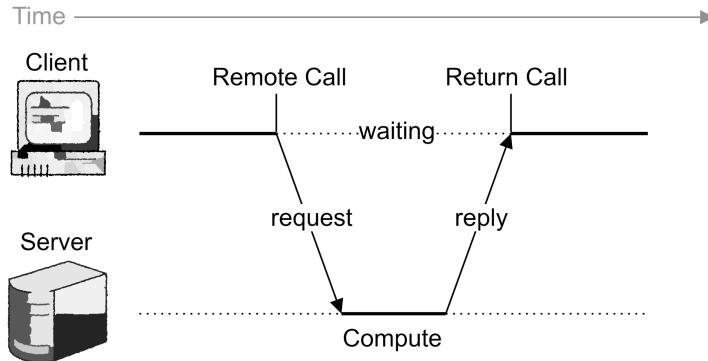


Figure 1.7: RPC call stack over time. Once the client makes the call it waits for the server to process the request and return the result. The programmer need not worry beyond sending the request and receiving the response. The RPC deals with all the heavy work of facilitating the communication.

Now to discuss what marshaling and unmarshaling are:

Definition 2.4: Marshaling and Unmarshaling

Marshaling handles data format conversions, converting the object into a byte stream (binary data). This conversion is known as **serialization**. This is done as the network can only transmit bytes

Unmarshaling is the process of converting the byte stream into the original, object called **deserialization**. This allows the server to process the request.

To illustrate serialization and deserialization, consider the following diagram:

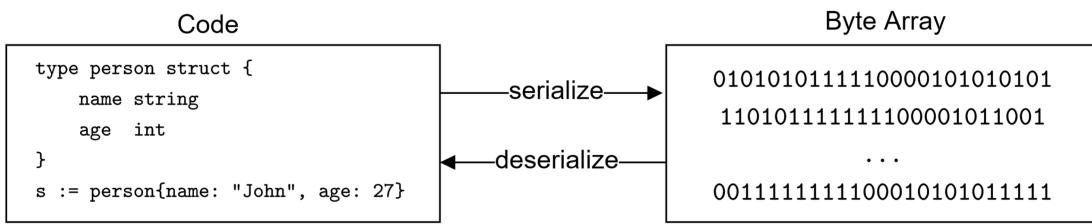


Figure 1.8: Serialization and Deserialization of data.

There is one cardinal rule to remember when dealing with RPCs:

Theorem 2.1: Network Reliability

The network is always unreliable.

That is to say, the network can drop packets, delay messages, or deliver them out of order. Anything that can go wrong will go wrong.

To handle network unreliability, we'll first consider two failure models:

Definition 2.5: At-least-once & At-most-once

- **At-least-once (client):** The client repeats requests until a response is heard. Reads alone are fine, but writes cause race conditions.
- **At-most-once (server):** The server handles duplicate requests from clients via some ID system, enabling cached responses (requires client to send IDs).

For our communication to work *reliably* we need At-least-once and At-most-once with unlimited tries coupled by a fault-tolerant implementation. This brings us to the **GO RGC library**.

Definition 2.6: Go RPC Library

The Go RPC library provides a simple way to implement RPCs in the programming language Go. This gives us:

- At-most-once model with respect to a single client-server
- Built on top of single **TCP connection** (Transport Layer Protocol). This protocol ensures reliable communication between client and server.
- Returns error if reply is not received, e.g., connection broken (TCP timeout)

Now to discuss briefly how a basic TCP connection is made:

Definition 2.7: Establishing a TCP Connection (SYN ACK)

First a three-way handshake is a method used in a TCP/IP network to create a connection between a local host/client and server. It is a three-step method that requires both the client and server to exchange **SYN (synchronize)** and **ACK (acknowledgment)**.

1. The client sends a SYN packet to the server requesting to synchronize sequence numbers.
2. The server responds with a SYN-ACK packet, acknowledging the request and sending its own SYN request.
3. The client responds with an ACK packet, acknowledging the server's SYN request.

After the three-way handshake, the connection is established and the client and server can communicate exchanging SYN and ACK data-packets. To end the connection another three-way handshake takes place, where instead SYN, **FIN (finish)** is used.

Such a system preserves the order of packets sent and received, making it a **FIFO (First In First Out)** system.

We now combine at-least-once and at-most-once to create the following model:

Definition 2.8: Exactly-Once Model

Utilized both **at-least-once** and **at-most-once** (client and server) models. Hence, client send requests indefinitely and the server handles duplicates via using IDs provided by the client. Duplicate requests receive a response of the cached response.

Below we illustrate a simple TCP connection:

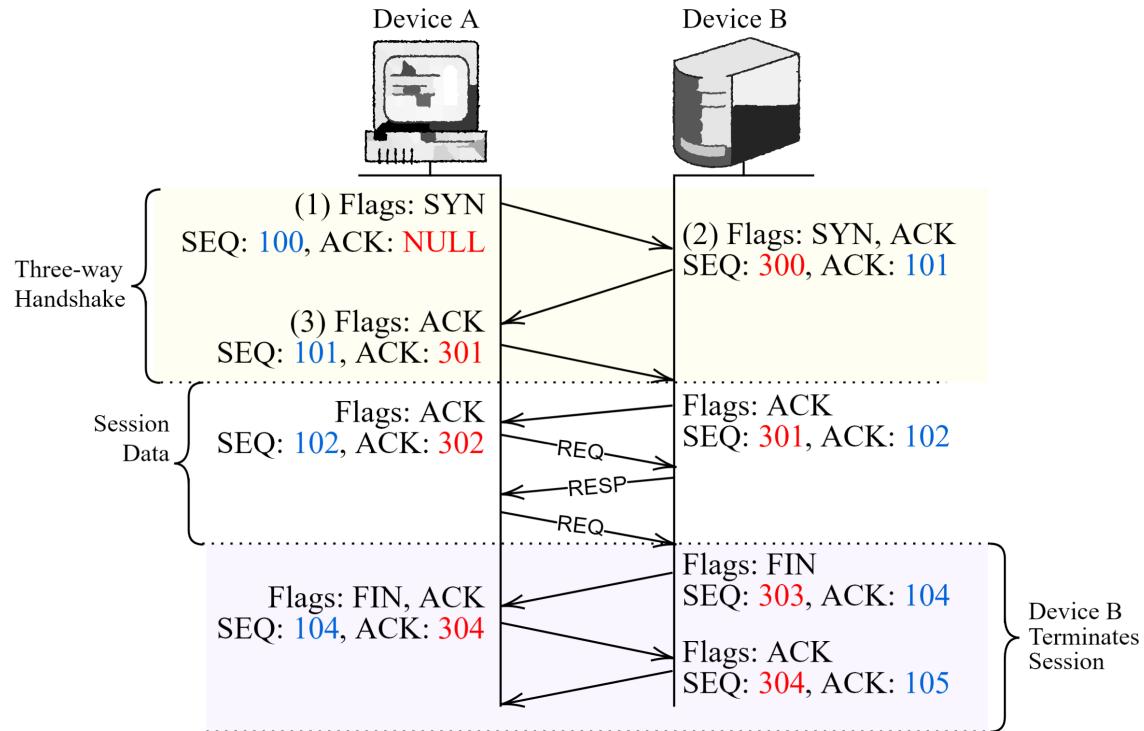


Figure 1.9: TCP Handshake, data transfer, and session termination. Here the client (Device A) begins a three-way handshake with the server (Device B) to establish a connection. Both start with arbitrary sequence numbers for security purposes. With each packet received the two devices increment their sequence numbers accordingly.

Tip: If there still resides curiosity for the networking aspect of RPCs, consider reading our other notes: <https://github.com/Concise-Works/Cyber-Security/blob/main/main.pdf>

1.2.2 Asynchronous Function Calls

Let's begin to discuss how functions can run simultaneously using Go's **goroutines**:

Definition 2.9: Asynchronous Function Calls

An **asynchronous function call** is a function that executes independently of the main program flow, enabling tasks to run concurrently or in parallel.

To illustrate the difference between synchronous and asynchronous function calls, consider the following diagram:

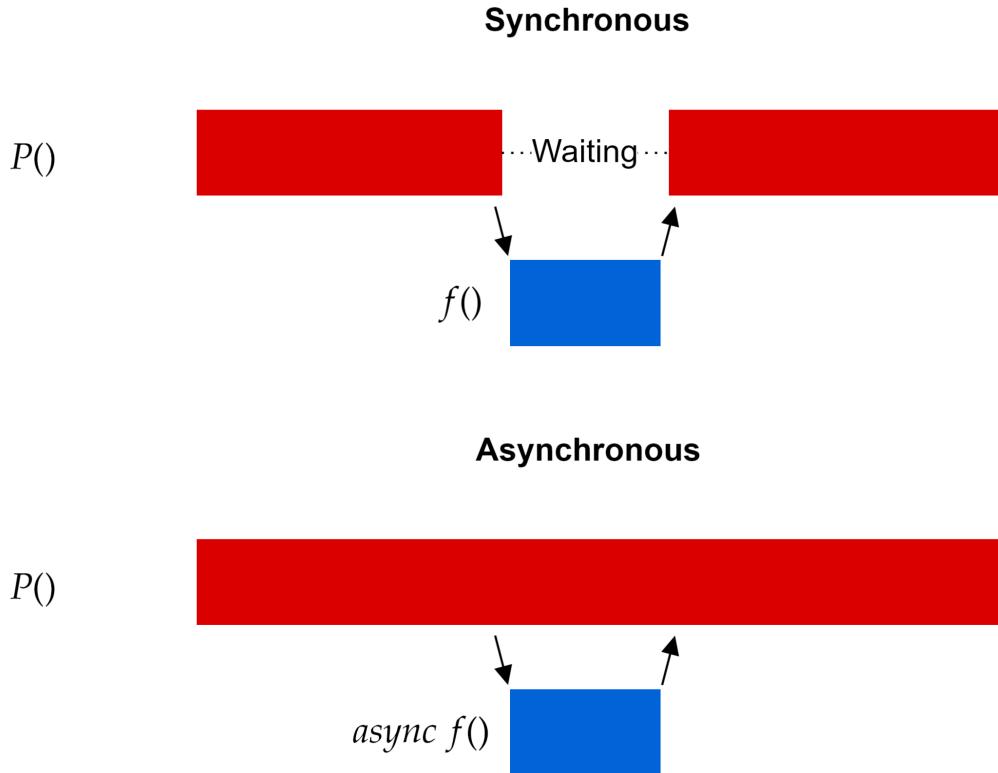


Figure 1.10: Synchronous vs. Asynchronous Function Calls.

- Function *P()* (above) is the main function for which our program is running. It makes a synchronous call to function *f()*, which blocks the main program flow until *f()* completes.
- In contrast, function *P()* (below) instead makes an asynchronous call to function *f()*, allowing the main program to continue executing while *f()* runs independently.

Definition 2.10: Asynchronous Function Calls in Go: Goroutines

A **goroutine** is a **lightweight** (lower memory overhead and scheduling cost compared to traditional OS threads) concurrent execution thread in Go. Goroutines enable functions to run asynchronously. Unlike traditional operating system threads, goroutines are managed by Go's runtime.

A goroutine is created using the `go` keyword before a function call, signaling to the Go runtime to run the function asynchronously from the main program flow.

The below details the Go runtime scheduler. **Note:** that overtime the Go runtime's algorithm may change to improve performance. This isn't key to understanding the content of this text, but is provided for completeness sake. Here—at the time of writing—is how it works at a high-level:

Definition 2.11: Go Runtime Scheduler

The Go runtime scheduler is responsible for managing goroutines via three conceptual entities:
The Go Scheduler: G, M, P

- **G (Goroutine):** A goroutine that holds the code to be executed.
- **M (Thread):** An OS thread that executes Go code via system calls or remains idle.
- **P (Processor):** Represents resources needed to execute code. The number of processors is determined by `GOMAXPROCS`.

If there are multiple goroutines, threads, and available processors, the scheduler matches them as follows:

- Many **Gs** (goroutines) are mapped to available **Ms** (OS threads), which execute them using **Ps** (processors) as execution resources.

Queues in the Scheduler

- **Global Run Queue (GRQ):** Holds all new goroutines that are yet to execute.
- **Local Run Queue (LRQ):** Holds goroutines that are assigned to a specific **P**.

For example, let the processors in the scheduler be defined as $P = \{P_1, P_2, \dots, P_n\}$ where $n = \text{GOMAXPROCS}$. Then the scheduler follows the following steps:

1. If P_1 has no more goroutines to execute, it follows these steps:
 - a) Check **GRQ** for a **G** (goroutine) roughly *1/61th of the time*.
 - b) If nothing is found, check **LRQ** again.
 - c) If nothing is found, attempt to **steal** work from other **Ps**.
 - d) If nothing is found, check **GRQ** one last time.
 - e) Finally, **poll the network** (i.e., check for incoming network work).

The next page includes a diagram of the Go runtime scheduler above.

To illustrate the Go runtime scheduler on a high-level, consider the following diagram in contrast to Figure 1.3:

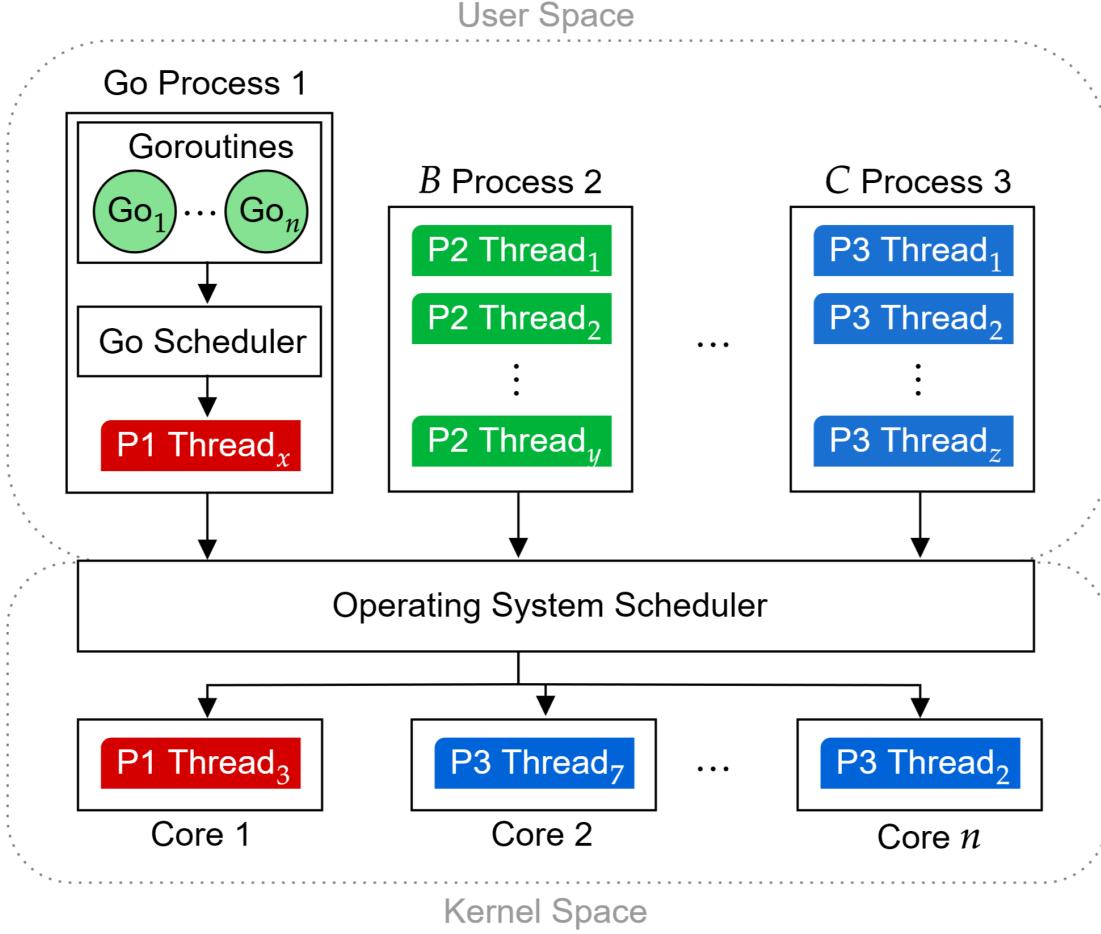
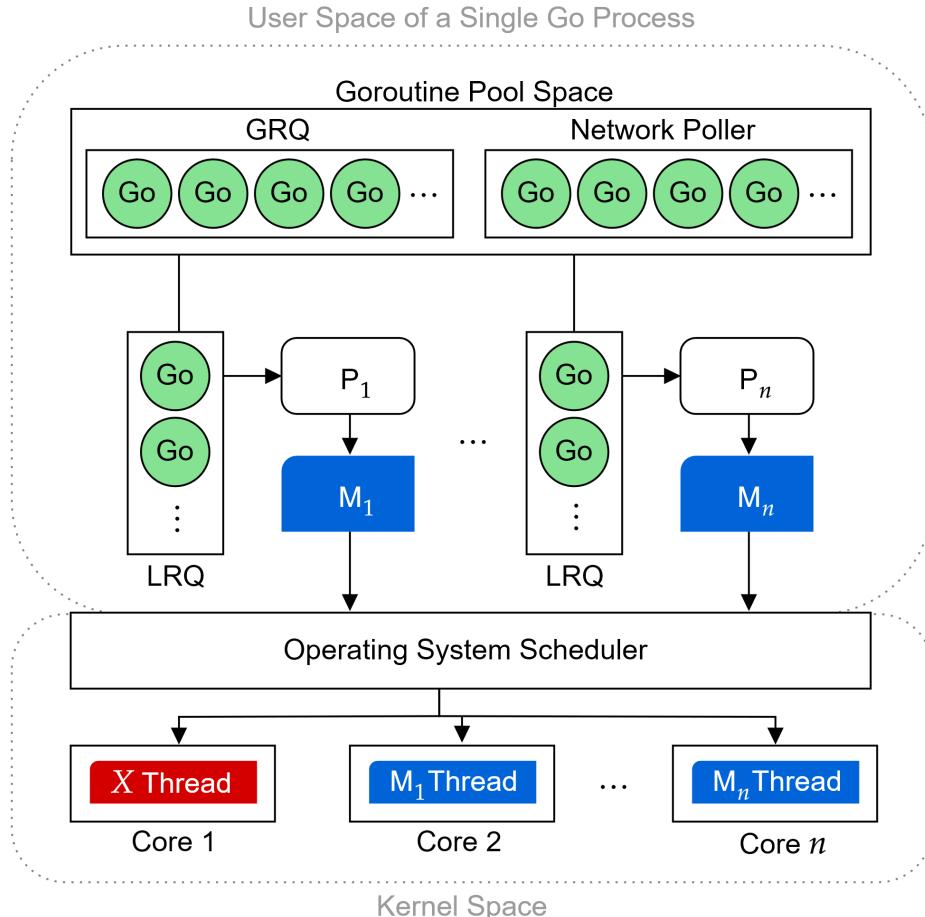


Figure 1.11: Go runtime scheduler with G, M, P entities.

Where the system has many different processes B, C and so on, we focus on the process that contains an instance of the Go runtime scheduler. Within this process each goroutine is scheduled by the Go runtime scheduler. The Go scheduler determines the number of threads needed and assigns goroutines to them. The process then presents these threads to the OS Scheduler which assigns them to the available cores.

In particular, **The OS has the final decision on which threads run on which cores.** The Go runtime scheduler only manages what threads to present. This is still helpful as the Go runtime can context switch the threads between goroutines before the handoff. Hence, the name **lightweight threads**, as context switching for the OS is expensive.

In particular we zoom in on the Go runtime scheduler of a single process:



GRQ := Global Run Queue

LRQ := Local Run Queue

X := Arbitrary Thread Belonging to Another Process

Figure 1.12: Go runtime scheduler within a single process.

Here our “Goroutine Pool Space” represents the latent goroutines waiting to be executed. We populate the LRQs of each P and their assigned M thread. The Ms are presented to the OS Scheduler. Moreover, the X is some arbitrary thread belonging to another process on the system. For emphasis

Theorem 2.2: Go Runtime Scheduler vs. OS Scheduler

The Go runtime only manages threads to present; The OS schedules threads to cores.

In all, the asynchronous nature of goroutines may create undefined behavior if not handled properly.

Example 2.1: Count to n using Goroutines

Consider the following Go program that counts to n using a goroutine:

Listing 1.1: Goroutine Example: Count to n

```
package main // Required for Go programs to run as executables
import (
    "fmt"
    "time"
) // Import required packages : fmt for printing and time for sleep

// 'i:=1' is short for 'var i int = 0'
func countUp(n int) {
    for i := 1; i <= n; i++ {
        // Anonymous function declared as a goroutine
        go func() {
            fmt.Println("Goroutine:", i)
        }()
    }
}
// Main entry point of the program
func main() {
    countUp(5)
}
```

However, this won't print anything as the main function exits before the independent goroutine can finish. A simple fix we'll do for now is put a sleep to wait for the goroutine to finish.

Listing 1.2: Adding a Sleep to Wait for Goroutine

```
func main() {
    countUp(5) // contains a goroutine
    time.Sleep(2 * time.Second) // Wait for goroutine to finish
    fmt.Println("Main function exits")
}
```

Ensuring the main function waits for the goroutine to finish. ■

Theorem 2.3: Main Goroutine Thread

the main function of a goroutine is too a kind of goroutine. We may refer to it as the **main goroutine** or **main thread**. In particular, the main goroutine is the first to run and may finish before any other goroutine.

Though since calls happen independent of each other means they happen simultaneously.

Example 2.2: Count to n using Goroutines Corollary

Continuing off from the previous example (2.1), we'll get an output such as:

Listing 1.3: Output of Goroutine Example

```
...
func countUp(n int) {
    for i := 1; i <= n; i++ {
        go func() {
            fmt.Println("Goroutine:", i)
        }()
    }
}

func main() {
    countUp(5) // Runs countUp concurrently
    time.Sleep(2 * time.Second) // Wait for goroutine to finish
    fmt.Println("Main function exits")
}

/* Output:
Goroutine: 4
Goroutine: 3
Goroutine: 5
Goroutine: 2
Goroutine: 1
Main function exits
*/
```

The goroutine spawns multiple threads for each print of counter i . Therefore the order at which they execute is up to the Go runtime scheduler. ■

Theorem 2.4: Goroutines and Multithreading

Goroutines will attempt to run on multiple threads to achieve parallelism. However, if there isn't enough cores available, threads will run concurrently on the same core.

To declare how many cores can use, `runtime.GOMAXPROCS` from the `runtime` package can be used.

Listing 1.4: Setting the Number of Cores for Goroutines

```
import "runtime"
runtime.GOMAXPROCS(n) // n = number of cores to use
```

By default, Go will use the number of cores available on the machine.

Try these examples out in Go to get a feel for how goroutines work.

Definition 2.12: Installing and Running Go Programs

First, install Go from the official website: <https://go.dev/doc/install>. The Go file extension is `.go`:

- **To run a Go program:** Use the command `go run <filename>.go`.
- **To build a Go program:** Use the command `go build <filename>.go` to create an executable. Then run the program in a terminal via `./<filename>`.

Tip: This text will teach the necessary components as we go along. However, if one wishes to learn on their own a little first, consider the following resource: <https://gobyexample.com/>. Though this text does assume prior programming knowledge and should be follow-able without the resource.

1.2.3 Synchronization: Data Races & Deadlocks

Asynchronous functions introduces a problem: If two threads access the same memory location at the same time, we face corruption of data as they try to write over each other:

Definition 2.13: Data Race

A **data race** occurs when multiple threads or goroutines access the same memory location concurrently, and at least one of the accesses is a write operation, without proper synchronization. This leads to undefined behavior, including inconsistent data and unpredictable program execution.

To avoid data races we implement the following strategy:

Definition 2.14: Mutex (Mutual Exclusion)

A **mutex** (short for *mutual exclusion*) is a synchronization primitive that prevents multiple threads from simultaneously accessing shared resources. This allows a single thread to place a **lock** on the resource, ensuring exclusive access until the lock is released.

Go has their own mutex implementation:

Definition 2.15: Go Mutex

In Go, the `sync.Mutex` type provides a way to control access to shared data. A `Mutex` has two main methods:

- `Lock()` : declares that the current goroutine from which it resides has exclusive access to the resource.
- `Unlock()` : Releases the mutex, allowing other goroutines to access the resource.

Example 2.3: Increasing a Counter Variable with Goroutines

Consider the following example where a function `incCounter()` increments a shared counter variable:

Listing 1.5: Incrementing a Counter Variable

```
...
var counter int // declaring global counter variable

func incCounter() {
    counter = counter + 1
}

func main() {
    // forloop spawning an instance of incCounter() in a goroutine
    for i := 0; i < 1000; i++ {
        go func() {
            incCounter()
        }()
    }
    time.Sleep(5 * time.Second)
    fmt.Println("Counter:", counter)
}
/* Output: Counter: 982 */
```

By the end of the forloop, the counter will most often not be 1000. This is due to counter having a different state in each goroutine. To fix this, we'll use a mutex. So it is very possible that the first 2 goroutines look like this:

- Goroutine 1: `counter = 0 + 1`
- Goroutine 2: `counter = 0 + 1`

Where all three goroutines see the counter as 0, increment it all setting it to 1. ■

Now to fix the previous example (2.3) using a mutex:

Definition 2.16: Increasing a Counter Variable with a Mutex

To ensure a global variable counter is incremented correctly, we'll use a mutex:

Listing 1.6: Using a Mutex to Increment a Counter Variable

```
... // imported the "sync" package for the mutex
var counter int
var mu sync.Mutex // declaring a mutex

func incCounter() {
    mu.Lock() // Lock the mutex
    counter = counter + 1
    mu.Unlock() // Unlock the mutex
}

func main() {
    for i := 0; i < 1000; i++ {
        go func() {
            incCounter()
        }()
    }
    time.Sleep(5 * time.Second)
    fmt.Println("Counter:", counter)
}

/* Output:
Counter: 1000
*/
```

By using a mutex, we ensure that only one goroutine can access the shared counter variable at a time. **Important Note:** This does not ensure the order in which the goroutines run.

Though with mutexes may come another problem, what if a goroutine never releases the lock?

Definition 2.17: Deadlock

A **deadlock** occurs when two or more asynchronous processes are waiting for each other to release a resource, preventing all processes from progressing. This results in a program that hangs indefinitely.

In a large project a logical mistake in a sea of processes can lead to a deadlock.

Example 2.4: Deadlock Scenario

Say we have functions `task1()` and `task2()` that each require a mutex lock:

Listing 1.7: Deadlock Scenario

```
... // dots represent some passage of code

go func task1() {
    lockA.Lock() ... lockB.Lock()
    ...
    lockB.Unlock() ... lockA.Unlock()
}

go func task2() {
    lockB.Lock() ... lockA.Lock()
    ...
    lockA.Unlock() ... lockB.Unlock()
}

...
```

Depending on how the scheduler runs, these two tasks will lock each other out, halting the program indefinitely. ■

1.2.4 References & Pointers in Go

In Go, problems may arise from how Go deals with scoped variables:

Definition 2.18: Reference vs. Value Types

In Go, variables can be either **reference types** or **value types**:

- **Reference Types:** Point to a memory location where the actual data is stored. Changes to the reference type will affect all variables pointing to the same memory location.
- **Value Types:** Store the actual data in memory. Changes to a value type will not affect other variables.

Definition 2.19: Closures and Reference Types

In Go, if a variable isn't explicitly pass to a function, but is rather accessible from the function's scope, it is considered a **closure**. This closure is a reference to the variable, not the data itself.

Example 2.5: Closures and Goroutines

Let `data` be some channel and `do_something()` be some function that returns a value:

```
...
batch := 0
for i := 0; i < k; i++ {
    go func() {
        data <- do_something(batch)
    }()
}
batch++
...
```

Here, the `go func` closure will reference the `batch` variable, not the value. Hence the main program flow (the main thread) might increment `batch` before the goroutine runs, leading to undefined behavior. To fix this, we pass the variable as an argument to the goroutine:

```
...
batch := 0
for i := 0; i < k; i++ {
    go func(batch int) {
        data <- do_something(batch)
    }(batch)
}
batch++
...
```

Now the goroutine will receive the value of `batch` at the time of the loop iteration. ■

Many data-structures in Go pass by value. Pointers ensure we are updating the original object:

Definition 2.20: Passing Pointers in Go

Pointers in Go pass the memory address of a variable via the `&` operator. To access the value stored at the memory address, use the `*` operator. E.g.,

```
var x int = 5
var y *int = &x // y stores the memory address of x
fmt.Println(*y) // Prints the value stored at the memory address
```

1.2.5 Waiting for Goroutines to Finish

Previously we used `time.Sleep()` to wait for goroutines to finish; However, Go provides a solution to this problem:

Definition 2.21: Wait Groups

A **wait group** is a synchronization primitive in Go that allows the main program to wait for a collection of goroutines. A wait group is a counter spawned with `sync.WaitGroup` and has three main methods:

- `Add(n int)` : Increments the wait group counter by `n`.
- `Done()` : Decrements the wait group counter by 1.
- `Wait()` : Blocks the main program until the wait group counter reaches 0.

Example 2.6: Using Wait Groups

Let's consider the following example where we use a wait group to wait for goroutines to finish:

```
...
var wg sync.WaitGroup // declaring a wait group
var mu sync.Mutex
for i := 0; i < 1000; i++ {
    wg.Add(1) // Increment the wait group counter
    go func() {
        mu.Lock()
        incCounter()
        mu.Unlock()
        wg.Done() // Decrement the wait group counter
    }()
}
wg.Wait() // Wait for wg counter to reach 0
```

An aside on a handy feature of Go:

Definition 2.22: Deferred Function Calls

In Go, the `defer` keyword **defers a function call** to run at the end of the innermost scoped function. Deferred functions are often used to ensure cleanup tasks are executed, such as closing files or releasing resources.

Example 2.7: Deferred Function Calls

Consider the previous example (2.6) with a deferred function call of `wg.Done()`:

```
...
for i := 0; i < 1000; i++ {
    wg.Add(1)
    go func() {
        defer wg.Done() // Deferred function call
        mu.Lock()
        incCounter()
        mu.Unlock()
    }()
}
...
```

The `wg.Done()` is deferred until the goroutine completes. ■

1.2.6 Sending Messages Between Goroutines

Now, say there are tasks *A* and *B*, for which *B* depends on the completion of *A*. Since *A* and *B* both run independently, we need a way for *B* to wait for a signal from *A*. This is where **channels** come in:

Definition 2.23: Channels

A **channel** is a typed conduit through which goroutines communicate. Channels allow goroutines to send and receive data. Channels are created using the `make()` function with the `chan` keyword. Channels must be closed after use to prevent memory leaks with `close()`.

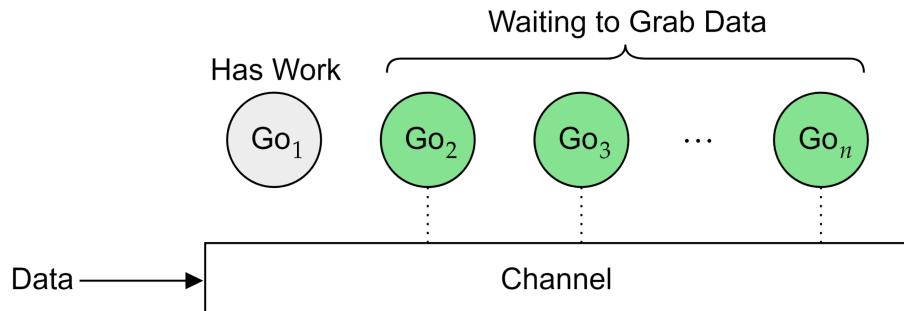


Figure 1.13: A collection of Goroutines competing for the next channel resource.

Note, despite the diagrams order of goroutines, the order in which they run is up to the Go runtime scheduler.

Example 2.8: Synchronizing Incoming Data Processing with Channels

Consider the following example of downloading and processing data concurrently:

Listing 1.8: Using Channels to Synchronize Downloading and Processing

```
package main
import (
    "fmt"
    "sync"
    "time"
)

// Download function simulates downloading data and sends a signal when
// done
func download(i int, ch chan int) {
    fmt.Printf("Downloading: Resource_%d...\n", i)
    time.Sleep(5 * time.Second) // Simulating download time
    fmt.Printf("Download complete: Resource_%d\n", i)
    ch <- i // Send signal that download is complete
}

// Process function waits for a signal before processing
func process(ch chan int, wg *sync.WaitGroup) {
    defer wg.Done()
    i := <-ch // Wait for download to complete
    fmt.Printf("Processing: Resource_%d...\n", i)
    time.Sleep(1 * time.Second) // Simulating processing time
    fmt.Printf("Processing complete: Resource_%d\n", i)
}

func main() {
    n := 5
    ch := make(chan int) // Unbuffered channel
    var wg sync.WaitGroup
    wg.Add(n)
    // Spawn n goroutines to download and process resources
    // concurrently
    for i := 0; i < n; i++ {
        go download(i, ch)
        go process(ch, &wg)
    }

    wg.Wait()
    close(ch) // Close channel after all downloads are completed
    fmt.Println("Main function exits")
}
```

Theorem 2.5: Channel Types

In Go, channels can be either **unbuffered** or **buffered**:

- **Unbuffered Channels:** Require a sender and receiver to be ready to communicate. If the receiver is not ready, the sender will block until the receiver is ready.
- **Buffered Channels:** Allow a sender to send data to a channel without the receiver being ready. The channel will store the data until the receiver is ready. Full channels block the sender, and empty channels block the receiver.

Unbuffered channels undergo a **handshake** process:

Unbuffered Channels in Go

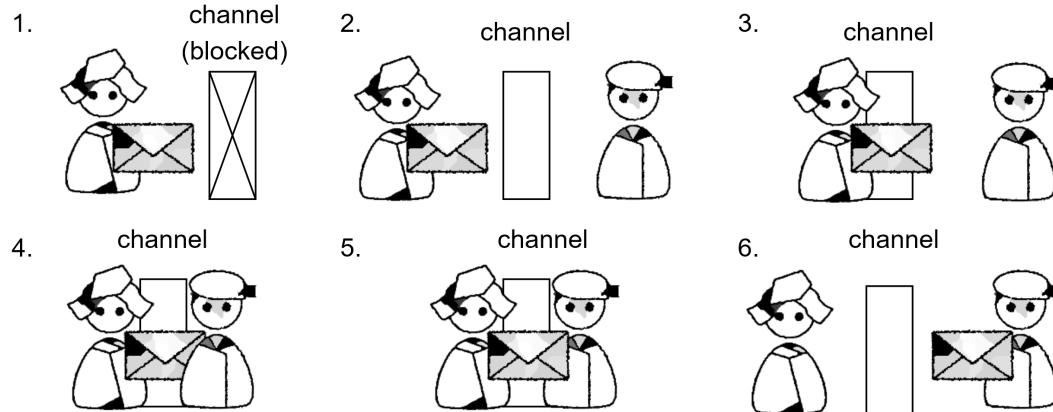


Figure 1.14: Handshake process of an unbuffered channel.

Say Alice (left) want to send a message to Bob (right) over a channel. (1) The channel is blocked until Bob is ready to receive. (2) The channel is no longer blocked read for the exchange. (3) Alice performs a **send** and is locked into the operation until Bob **receives** the message. (4-5) Bob enters the channel and receives the message; Both Alice and Bob are locked into the operation until the message exchange is complete. (6) they both are free to continue their operations.

Theorem 2.6: Unbeffered Blocking

Let A and B be two goroutines attempting to send messages over an unbuffered channel. If A enters the channel first, B will be blocked until A finishes the transaction.

The previous example (2.8) uses an unbuffered channel.

In contrast buffered channels allow for a more asynchronous approach:

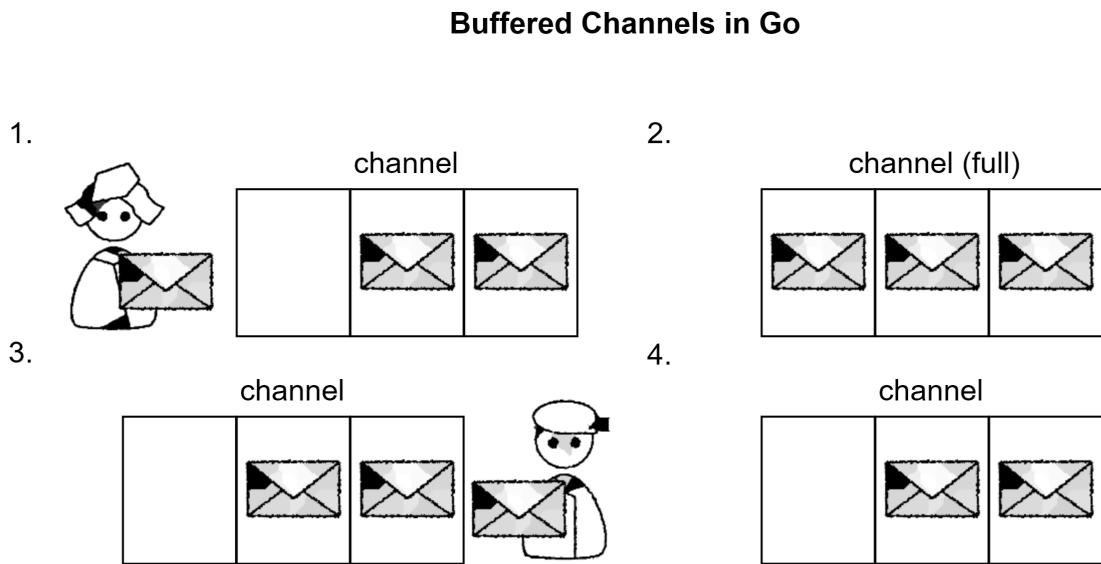


Figure 1.15: Buffered channel allowing for asynchronous communication.

(1) Here Alice (left) fills the channel from left to right with messages. (2) The channel is full and Alice is free to continue her operations. (3) Bob (right) takes the rightmost message from the channel. (4) Bob is free to continue his operations, leaving the channel.

To make the last example (2.8) use a buffered channel, we can modify the channel creation.

Example 2.9: Using Buffered Channels

Consider the previous example (2.8) with a buffered channel:

Listing 1.9: Using Buffered Channels to Synchronize Downloading and Processing

```

...
ch := make(chan int, n) // Buffered channel with a capacity of 5
...

```

Here, the channel has a buffer size of n, allowing up to n messages to be stored before the receiver is ready. ■

1.2.7 Task, Data, and Pipeline Parallelism

Task, data, and pipeline parallelism are three common forms of parallelism:

Definition 2.24: Task Parallelism

Task parallelism involves running multiple tasks simultaneously. Each task is independent and can run in parallel with other tasks, perhaps even on the same data.

In essence, we may think same data, different tasks:

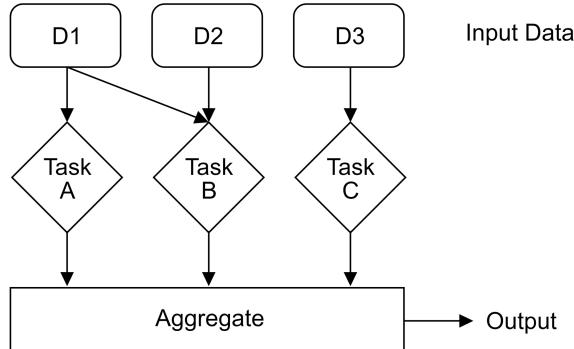


Figure 1.16: Task Parallelism Culminating into an Aggregate Result

Definition 2.25: Data Parallelism

Data parallelism involves running the same task on multiple data items. Each task is identical, but the data is different.

We may think same task, different data:

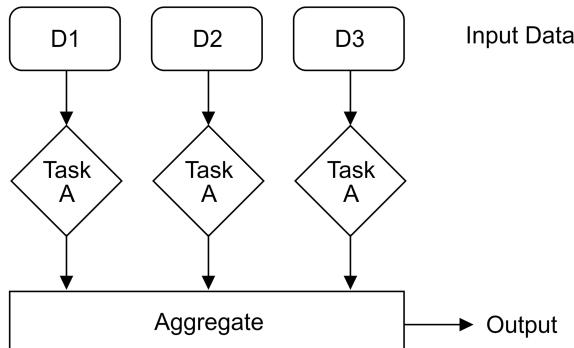


Figure 1.17: Data Parallelism Culminating into an Aggregate Result

To continue, we have:

Definition 2.26: Pipeline Parallelism

Pipeline parallelism involves breaking a task into multiple stages, each of which can be executed concurrently. The output of one stage is the input to the next stage.

For instance, consider the following pipeline:

- **Task A:** “Search for a flight.” (1 time unit)
- **Task B:** “Book a flight.” (1 time unit)

First consider the scenario where we only have one resource to work with, resulting in concurrency:

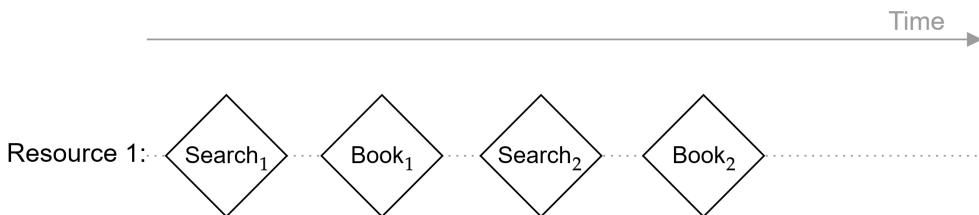


Figure 1.18: Searching and Booking flights concurrently

Now consider the scenario where we have two resources to work with, resulting in parallelism: In

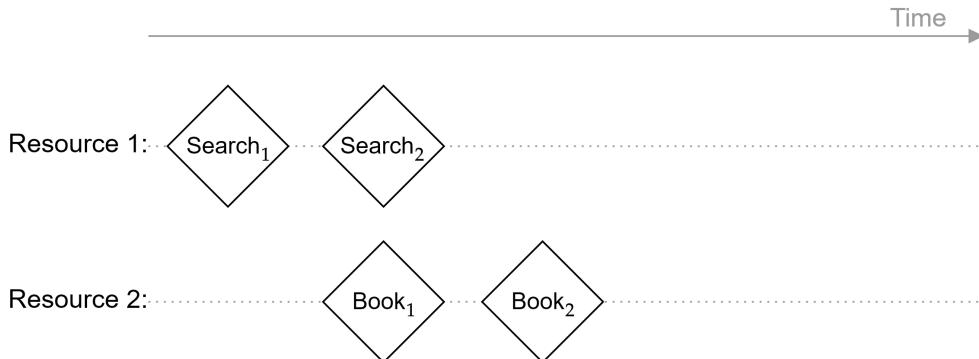


Figure 1.19: Searching and Booking flights in parallel

this case, once the first search is done, we can start booking the flight and search for the next flight in parallel.

1.2.8 Arrays & Slices in Go

Arrays in Go act like arrays in other languages, a fixed-size collection of items of the same type. Slices, on the other hand, allow us to work with a dynamically-sized sequence of elements.

Definition 2.27: Arrays in Go

An **array** in Go is a fixed-size collection of elements of the same data type. Arrays in Go are value types, meaning they are copied when assigned to a new variable.

Arrays are declared using the syntax:

```
var arr [size]Type
```

For example, an array of integers with 5 elements:

```
var numbers [5]int
```

Elements in an array can be accessed using zero-based indexing:

```
numbers[0] = 10 // Assign value
fmt.Println(numbers[0]) // Access value
```

Arrays cannot be resized, and their size must be known at compile time. For dynamic collections, slices are preferred.

Example 2.10: Doubling Items in an array

Consider the following example where we double each element in an array:

```
package main

import "fmt"

func main() {
    // Initialize an array
    numbers := [5]int{1, 2, 3, 4, 5}

    // Double each element in the array
    for i := 0; i < len(numbers); i++ {
        numbers[i] *= 2 // Shorthand for numbers[i] = numbers[i] * 2
    }

    // Print the modified array
    fmt.Println(numbers)
}
```

// Output: [2 4 6 8 10]

In contrast, slices:

Definition 2.28: Slices in Go

A **slice** is a dynamically-sized reference to a portion of a single underlying array. Slices are declared using square brackets without specifying a fixed size:

```
1  var numbers []int // A slice of integers
```

Slices are typically created using the `make` function or by slicing an existing array:

```
1  // Using make()
2  numbers := make([]int, 5) // Creates a slice with length 5
3
4  // Slicing an array
5  arr := [5]int{1, 2, 3, 4, 5}
6  slice := arr[1:4] // Slice from index 1 to 3 -> {2, 3, 4}
```

Slices maintain a reference to the original array, meaning modifications affect both:

```
1  arr := [5]int{1, 2, 3, 4, 5}
2  slice := arr[1:3]
3  slice[0] = 99 // Modifies arr[1] as well
4  fmt.Println(arr) // Output: [1 99 3 4 5]
```

The `append()` function modifies slices given there's enough capacity:

```
1  arr := [5]int{1, 2, 3, 4, 5}
2  slice := arr[:1] // Slice from index 0 to 0 -> {1}
3  slice = append(slice, 7, 8) // Adds elements to the slice
4  fmt.Println(slice) // Output: [1 7 8]
5  fmt.Println(arr) // Output: [1 7 8 4 5] (modified)
```

If `append()` exceeds the slice's capacity, a new array is allocated and referenced by the slice:

```
1  ... // Previous code
2  fmt.Println(cap(slice)) // Output: 5 (capacity of the slice)
3  slice = append(slice, 7, 8, 9, 10, 11) // Exceeds capacity, (6 total)
4  fmt.Println(slice) // Output: [1 7 8 9 10 11]
5  fmt.Println(arr) // Output: [1 2 3 4 5] (unchanged)
```

To copy an array to a slice, use the `copy()` function:

```
1  arr := [5]int{1, 2, 3, 4, 5}
2  slice := make([]int, len(arr))
3  copy(slice, arr) // Syntax: copy(destination, source)
4  slice[0] = 99
5  fmt.Println(slice) // Output: [99 2 3 4 5]
6  fmt.Println(arr) // Output: [1 2 3 4 5] (unchanged)
```

1.2.9 Repeating Tasks: Tick and Ticker in Go

In Go, the `time` package provides two types for repeating tasks at regular intervals:

Definition 2.29: `time.Tick` and `time.Ticker` in Go

The `time` package in Go provides two mechanisms for scheduling repeated tasks at fixed intervals:

- `time.Tick(duration)`: Returns a channel that sends the current time at regular intervals. It is a convenience function but cannot be stopped.
- `time.NewTicker(duration)`: Creates a `Ticker` object, which provides a `.Stop()` method to halt the ticker. Additionally the `.C` returns a channel from which the signal can be read. This channel is read only, hence a type of `<-chan time.Time`.

Example 2.11: Record a signal n times every second

Say we want to record a signal n times every second. We can use a `time.Ticker`:

```
package main
import "fmt"; import "time"; import "sync"

func Status(ch <-chan time.Time, wg *sync.WaitGroup) {
    defer wg.Done()
    <-ch // Wait for signal
    fmt.Println("Status: OK")
}

func main() {

    n := 5
    ticker := time.NewTicker(time.Second)
    tickerChan := ticker.C
    var wg sync.WaitGroup

    // Spawns n goroutines which waiting for the signal
    for i := 0; i < n; i++ {
        wg.Add(1)
        go Status(tickerChan, &wg)
    }
    wg.Wait()
    ticker.Stop()
}
```

This type of example can be extended to perform any task at a given interval. ■

1.2.10 Conditionally Reading from Channels: Select in Go

Now to discuss conditionally reading from multiple channels:

Definition 2.30: select in Go

The `select` statement in Go allows a goroutine to wait on multiple channels and perform an action as soon as one of them receives a value:

```

1  select {
2      case val := <-ch1:
3          // Received from ch1
4      case val := <-ch2:
5          // Received from ch2
6      default:
7          // Executes if no channels are ready
8  }
```

Note: `default` is optional. If multiple channels are ready, one is chosen at random.

Example 2.12: Conditionally Waiting for a Signal

Consider a situation where we conditionally wait on a channel for a signal:

```

... // Imported "math/rand"
func Status(rc chan string) {
    for { // Loops forever sending a message every second
        // Randomly select a status
        rc <- []string{"Great", "Ok", "Slow"}[rand.Intn(3)]
        time.Sleep(1 * time.Second)
    }
}

func main() {
    replyChannel := make(chan string) // Channel for receiving status
    go Status(replyChannel) // Asynchronous status generator
    for { // Loop forever waiting for such status
        select {
        case <-replyChannel:
            fmt.Println("Status:", <-replyChannel)
        default:
            fmt.Println("Waiting...")
            time.Sleep(351 * time.Millisecond)
        }
    }
}
```

1.3 Time, Clocks, and Logical Ordering

1.3.1 Accuracy of Time: Atomic Clocks & NTP

Time allows us to order and identify events. Say we ran `time.Now()` on two different machines:

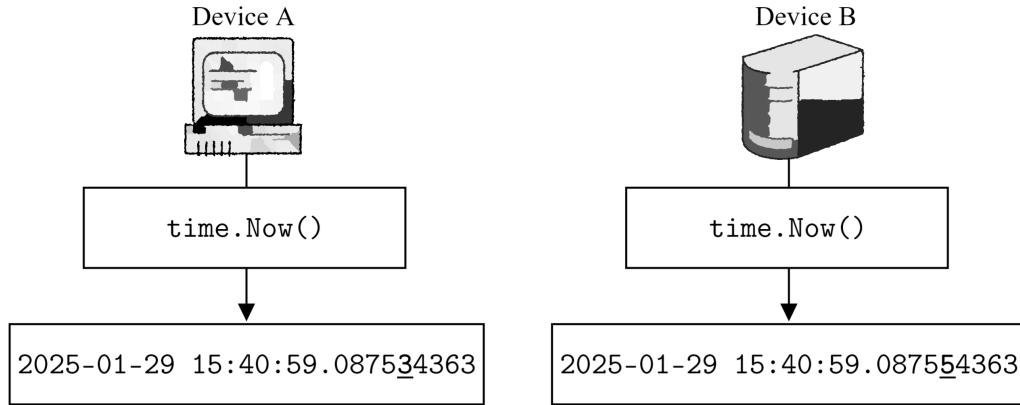


Figure 1.20: Using `time.Now()` on two different machines

Despite Device *A* appearing to be ahead of Device *B*, we cannot be certain via the following reasons:

Theorem 3.1: Clock Synchronization Impossibility

There are two key reasons why perfect clock synchronization is impossible:

- **Clock Skew:** There's a difference between every system clock (ideally 0), as they maintain their own local clock via a hardware oscillator incrementing a counter register.
- **Clock Drift:** Even if systems initialize with a reference time, their clocks will inevitably diverge due to variations in manufacturing, age, or environmental factors such as temperature. We measure the deviation by, $\frac{dC}{dt} = 1 + \rho$, where C is the clock time and t is the real time, and ρ (rho) is the drift rate (ideally 0).

We may formalize what we may consider synchronized clocks as follows:

Definition 3.1: Clock Synchronization Threshold

Let there be two clocks C_i and C_j . They are (δ) δ -synchronized if for all t time units:

$$|C_i(t) - C_j(t)| \leq \delta$$

E.g., C_i and C_j are δ -synchronized within 10ms if $|C_i(t) - C_j(t)| \leq 10ms$

In practice we to achieve semi-synchronized clocks, we developed the following protocol:

Definition 3.2: Network Time Protocol (NTP)

The NTP is a protocol synchronizes network clocks via a ground-truth time distribution system. The ground-truth time is are GPS satellite **atomic clocks**, which exhibit negligible drift over millions of years.

NTP employs a **round-trip time (RTT)** calculation to estimate the clock offset request latency. It also organizes synchronization strength into a **stratum hierarchy**, where lower-numbered strata indicate more accurate time sources:

- **Stratum 0:** Ground truth **atomic clocks/GPS receivers**.
- **Stratum 1:** NTP servers that directly synchronize with Stratum 0 reference clocks.
- **Stratum 2:** NTP servers synchronized to Stratum 1 servers.
- **Stratum 3 and beyond:** Weaker NTP servers synchronized to higher-stratum servers.
- **Stratum 16:** A system considered *unsynchronized* (e.g., a freshly booted system).

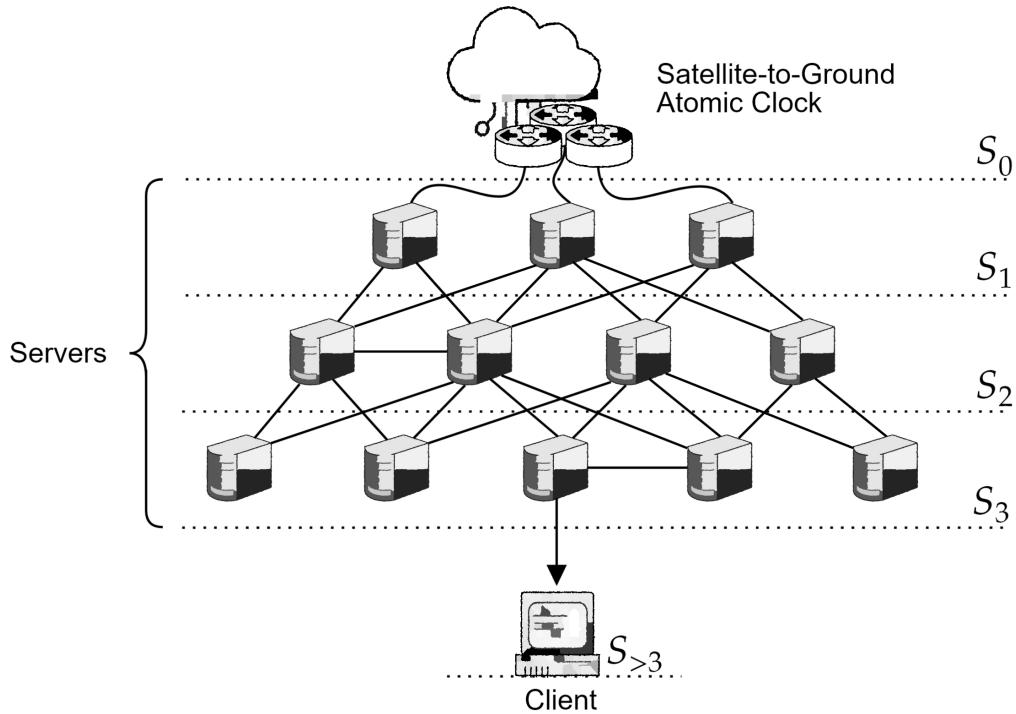


Figure 1.21: NTP Stratum Hierarchy From GPS Satellite Atomic Clock to Client.

1.3.2 Logical Clocks: Lamport & Vector Clocks

To get away from the limitations of physical clocks, we may use logical clocks to order events.

Definition 3.3: Logical Clocks

Let a and b be two events part of a totally ordered set of events. Let function $t(x)$ denote the time of event x . Then,

$$a \rightarrow b \implies t(a) < t(b)$$

Where $a \rightarrow b$ denotes that event a happens before b , which implies $t(a) < t(b)$.

We may become more formal about cause and effect relationships with the following definition:

Definition 3.4: Causal Order

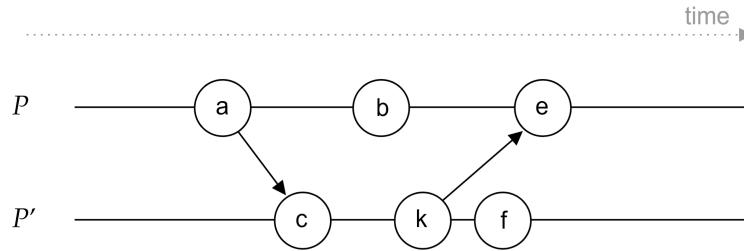
For r execution trace (sequence of events), the causal order relationship \rightarrow_r is defined as:

- If a happens before b in the same process, then $a \rightarrow_r b$.
- If a is a **sender** and b the **receiver**, then $a \rightarrow_r b$.
- **Transitive Property:** if $a \rightarrow_r b$ and $b \rightarrow_r c$, then $a \rightarrow_r c$.
- Events a and b are **concurrent** (denoted as $a \parallel b$) if:
 - ▶ $a \not\rightarrow_r b$ and $b \not\rightarrow_r a$, meaning neither event happened before the other.

Example 3.1: Causal Order Example

Determine the causal order relationship between events in processes P and P' :

- (a): $a ? b$; (b): $a ? k$; (c): $c ? b$; (d): $c ? e$



Answer on the next page. ■

Example (3.1) Answer: (a): $a \rightarrow_r b$; (b): $a \rightarrow_r k$; (c): $c \parallel b$; (d): $c \rightarrow_r e$.

Now we discuss a method that utilizes causal order, though assigns logical timestamps to events:

Definition 3.5: Lamport Clocks

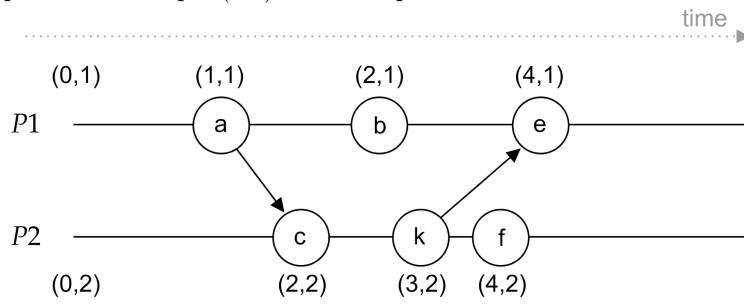
Named after Leslie Lamport, Lamport Clocks assign a logical timestamp to each event:

Let t_p store the logical time of process p . Then,

- **Initialization:** t_p is initialized to 0.
- **Timestamp Syntax:** Timestamps are tuples (t_p, p) , assigned to each e event.
- **Incrementing:** For each e in process p , increment t_p by 1 and assign the (t_p, p) to e .
- **Sending:** If p sends a message m to process q , the timestamp included is $((t_p + 1), p)$.
- **Receiving:** Upon receiving message m , process q sets $t_q = \max((t_p + 1), t_q)$.

Example 3.2: Lamport Clocks Example (3.1) Extended

Consider the previous Example (3.1) with Lamport Clocks:



In practice the only thing we have access to are these logical timestamps, which we must evaluate:

Theorem 3.2: Comparing Lamport Timestamps Causality

Given two events a and b with timestamps $t(a)$ and $t(b)$, with r trace, we only guarantee:

- If $a \rightarrow_r b$, then $t(a) < t(b)$.
- Let $a := (3, 1)$ and $b := (1, 2)$, then $t(a) > t(b)$.
- If $t(a) \geq t(b)$, then $a \not\rightarrow_r b$.
- ID ordering break ties: $(1, 1) < (1, 2)$.

We may now derive the following about concurrency:

Theorem 3.3: Non-causality

Two events a and b are concurrent ($a \parallel b$) under r trace if **both** conditions hold:

- $a \not\rightarrow_r b$ (a does not happen before b).
- $b \not\rightarrow_r a$ (b does not happen before a).

Lamport Clocks are useful for causal ordering, but they do not capture the full context of events:

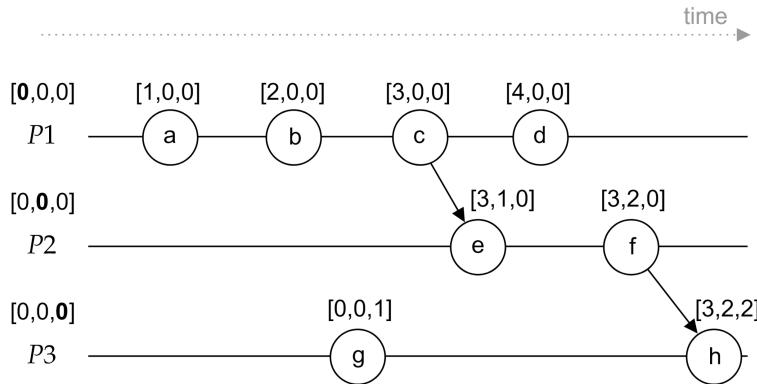
Definition 3.6: Vector Clocks

Let there be p_1, p_2, \dots, p_n processes each with a vector (array) v of size n . Index $v_i[i]$ stores the logical time of process p_i . Then, the following rules apply:

- **Initialization:** Each $\ell \in v_i$ of p_i is initialized to 0 (e.g., $[0, 0, \dots, 0]$).
- **Incrementing:** For each event e in process p_i , increment $v_i[i]$ by 1.
- **Sending:** When p_i sends a message m to p_j , include v_i in m (**no increment**).
- **Receiving:** Upon receiving message m , process p_j sets $v_j[j] = \max(v_j[j], v_i[j])$.

Example 3.3: Vector Clocks Example

Observe the following events and their corresponding vector clocks:



There too is a method of comparing vector clocks:

Theorem 3.4: Comparing Vector Clocks

Given two vectors v_p and v_q for processes p and q , we may derive the following:

- $v_p \leq v_q \iff \forall \ell : v_p[\ell] \leq v_q[\ell]$
- $v_p < v_q \iff \forall \ell : v_p[\ell] < v_q[\ell]$
- $v_p <> v_q$ (non-comparable) $\iff \forall \ell : \neg(v_p < v_q) \wedge \neg(v_p > v_q)$

i.e.,

- $v_p \leq v_q$: if and only if all corresponding indexes in v_p are less than or equal to v_q .
- $v_p < v_q$: if and only if all corresponding indexes in v_p are less than v_q .
- $v_p <> v_q$: if and only if there are at least two element pairs between v_p and v_q that are greater and less than each other. (e.g., $v_p = [1, 2]$ and $v_q = [2, 1]$, as $1 < 2$ and $2 > 1$).

And for causality we have:

Theorem 3.5: Vector Clocks Causality

Given two events a and b with vector clocks $v(a)$ and $v(b)$, we may derive the following:

- If $v(a) < v(b)$, then $a \rightarrow_r b$.
- If $a \rightarrow_r b$, then $v(a) < v(b)$.

Theorem 3.6: Total vs. Partial Ordering

Total Ordering means all timestamp pairs are comparable. Hence:

- **Lamport Clocks** define a **Total Ordering** as we can evaluate all pairs of timestamps.
- **Vector Clocks** define a **Partial Ordering** as there are situations such as $[1,0]$ and $[0,1]$, which are non-comparable.

Exercise 3.1: Given the following Lamport timestamps, determine the causal order relationships:

- (a) $(0, P) \ ? \ (0, Q)$
- (b) $(1, P) \ ? \ (0, Q)$
- (c) $(0, P) \ ? \ (1, Q)$

Exercise 3.2: Given the following Vector clocks, determine the causal order relationships:

- (a) $[0, 0] \ ? \ [0, 0]$
- (b) $[1, 0] \ ? \ [0, 0]$
- (c) $[0, 1] \ ? \ [1, 0]$

Answers on the next page.

Answer 3.1: Given the following Lamport timestamps, determine the causal order relationships:

- (a) $(0, P) ? (0, Q)$: $P \parallel Q$. To break the tie, the processes would need some order. If P is before Q in process order, then $P \rightarrow Q$ is possible.
- (b) $(1, P) ? (0, Q)$: $P \parallel Q$ or $Q \rightarrow P$ are possible.
- (c) $(0, P) ? (1, Q)$: $Q \parallel P$ or $P \rightarrow Q$ are possible.

Note: The Lamport timestamps are not enough to determine causal order relationships between independent processes, unless we have the context of the execution trace. In the processes were ordered

Answer 3.2: Given the following Vector clocks, determine the causal order relationships:

- (a) $[0, 0] ? [0, 0]$: $[0, 0] \parallel [0, 0]$
- (b) $[1, 0] ? [0, 0]$: $[1, 0] \not\rightarrow [0, 0]$
- (c) $[0, 1] ? [1, 0]$: $[0, 1] <> [1, 0]$ or $[0, 1] \parallel [1, 0]$

1.4 Implementing RPCs with Go

Before we can make RPC calls in Go, we need to understand Go's typing system.

1.4.1 Typing in Go

Definition 4.1: Typing in Go

Go is a statically typed language, meaning every variable has a fixed type at compile time. The language provides several built-in types, including: **Basic Types:** `int`, `float64`, `string`, `bool`, and **Composite Types:** `array`, `slice`, `map`, `struct`, `interface`.

A **type** in Go can also be user-defined using the `type` keyword. For example, we can define a custom structure:

```
type Person struct {
    Name string
    Age  int
}
```

A `struct` groups related fields together, allowing us to represent objects with multiple attributes. We can create and use instances of this struct:

```
p := Person{Name: "Alice", Age: 25}
fmt.Println(p.Name) // Outputs: Alice
```

Definition 4.2: Typing Functions in Go

Functions are explicitly typed, meaning parameters and return values must be declared:

```
// Takes two ints and returns an int
func Add(x int, y int) int {
    return x + y
}

// We can shorthand parameters of the same type:
func Divide(x, y int) (int, error) { // Returns an int and an error
    if y == 0 {
        return 0, fmt.Errorf("division by zero")
    }
    return x / y, nil
}
```

Here, `fmt.Errorf()` generates an error message.

Definition 4.3: Methods in Go: Pointers vs. Values

In Go, methods can be defined for both pointer receivers (`*T`) and value receivers (`T`). A method with a pointer receiver can modify the original struct, while a method with a value receiver works on a copy. Consider a struct representing a rectangle:

```
package main; import "fmt"

type rect struct {
    width, height int
}

// Pointer receiver: Can modify the original struct
func (r *rect) area() int {
    return r.width * r.height
}

// Value receiver: Works on a copy of rect
func (r rect) perim() int {
    return 2*r.width + 2*r.height
}

func (r *rect) setWidth(w int) {
    r.width = w
}

func (r rect) setHeight(h int) {
    r.height = h
}

func main() {

    r := rect{width: 10, height: 5}
    // Converts r.area() to (&r).area() automatically
    fmt.Println("area: ", r.area()) // 50
    fmt.Println("perim:", r.perim()) // 30

    rp := &r
    // Converts rp.perim() to (*rp).perim() automatically
    fmt.Println("perim:", rp.perim()) // 30
    fmt.Println("area: ", rp.area()) // 50

    r.setWidth(20)
    r.setHeight(10)
    fmt.Println("width:", r.width) // 20
    fmt.Println("height:", r.height) // 5
}
```

1.4.2 Go's RPC Package

In Go, we can use the `net/rpc` package to implement server-client RPCs.

Definition 4.4: Setting Up an RPC Server and Client in Go

Setting Up an RPC Server: To create an RPC server in Go, we need to,

- Define a **service type** (a struct) that will handle remote calls.
- Implement **methods** that satisfy the RPC requirements.
- Register the service using `rpc.Register()`.
- Set up a network listener using `net.Listen()`.
- Handle requests using `rpc.HandleHTTP()` and `http.Serve()`.

RPC Method Requirements: Methods must satisfy these constraints or be ignored:

- The method itself and its **type** must be exported.
- The method must return a single value of type `error`, and have exactly **two arguments**, both of which must be **exported or built-in types**.
- The method's **second argument must be a pointer**.

E.g. Signature,

```
func (t *T) MethodName(argType T1, replyType *T2) error
```

Setting Up an RPC Client: To communicate with the RPC server, a client must,

- Establish a connection using `rpc.DialHTTP()`.
- Call remote methods synchronously with `client.Call()`.
- Call remote methods asynchronously with `client.Go()`.

Example 4.1: Simple Local Arithmetic RPC (Part 1)

Let's create a simple Arithmetic RPC. First let's setup our file structure:

```
arith/
  └── server/
    └── server.go
  └── client.go
```

Example 4.2: Simple Local Arithmetic RPC (Part 2)

Now to setup the server:

```
package main; import "fmt"; import "log"; import "net"; import "net/rpc"

type ArithService int // Define the service type

type Args struct { A, B int } // Define the arguments struct

type Quotient struct { Quo, Rem int } // Reply struct for Dividing

// Both Multiply and Divide methods satisfy the RPC requirements
func (t *ArithService) Multiply(args *Args, reply *int) error {
    *reply = args.A * args.B
    return nil
}

func (t *ArithService) Divide(args *Args, quo *Quotient) error {
    if args.B == 0 { return fmt.Errorf("divide by zero") }
    quo.Quo = args.A / args.B
    quo.Rem = args.A % args.B
    return nil
}

func main() {
    arith := new(ArithService)
    err := rpc.Register(arith)
    if err != nil { log.Fatal("Register Error:", err) }

    listener, err := net.Listen("tcp", ":8080")
    if err != nil { log.Fatal("RPC Start Error:", err) }

    defer listener.Close()
    log.Println("Arithmetic RPC server is running on port 8080")

    for {
        // Server waits until a connection is made
        conn, err := listener.Accept()
        if err != nil {
            log.Println("Connection Error:", err)
            continue
        }
        // Serves the connection on a new goroutine
        go rpc.ServeConn(conn)
    }
}
```

Example 4.3: Simple Local Arithmetic RPC (Part 3)

After the following client code, run the server first, then client on a separate terminals:

```
package main; import "fmt"; import "log"; import "net/rpc"

// Define the same structures as the server
type Args struct {
    A, B int
}

type Quotient struct {
    Quo, Rem int
}

func main() {
    // Connect to the RPC server
    client, err := rpc.Dial("tcp", "localhost:8080")
    if err != nil {
        log.Fatal("Error dialing:", err)
    }
    defer client.Close()

    // Perform a multiplication RPC call
    args := &Args{7, 8}
    var reply int
    err = client.Call("ArithService.Multiply", args, &reply)
    if err != nil {
        log.Fatal("Multiply error:", err)
    }

    // Perform a division RPC call
    var quotient Quotient
    err = client.Call("ArithService.Divide", args, &quotient)
    if err != nil {
        log.Fatal("Divide error:", err)
    }

    // Output: 7 * 8 = 56
    fmt.Printf("ArithService: %d * %d = %d\n", args.A, args.B, reply)

    // Output: 7 / 8 = 0, remainder = 7
    fmt.Printf("ArithService: %d / %d = %d, remainder = %d\n", args.A,
               args.B, quotient.Quo, quotient.Rem)
}
```

2.1 Saving System State: Snapshots

This section discusses saving state. This is useful for fault-tolerance and system migrations.

Definition 1.1: Snapshot

A **snapshot** is a consistent global state of a distributed system at a specific point in time.

Definition 1.2: Consistent vs. Inconsistent Snapshots

To evaluate a snapshot's consistency, we compare events in the system **pre-snapshot** (events before the snapshot) and **post-snapshot** (events after the snapshot). The snapshot itself is instantaneous, like a photograph. Given an event ordering r :

- **Consistent Snapshots:** Respect causal dependencies. Let there be events e_1 and e_2 ; If $e_1 \rightarrow_r e_2$, then e_1 must be included in the snapshot if e_2 is present.
- **Inconsistent Snapshots:** Violate causal dependencies. If e_2 is included without the causally preceding event e_1 , then the snapshot is inconsistent.

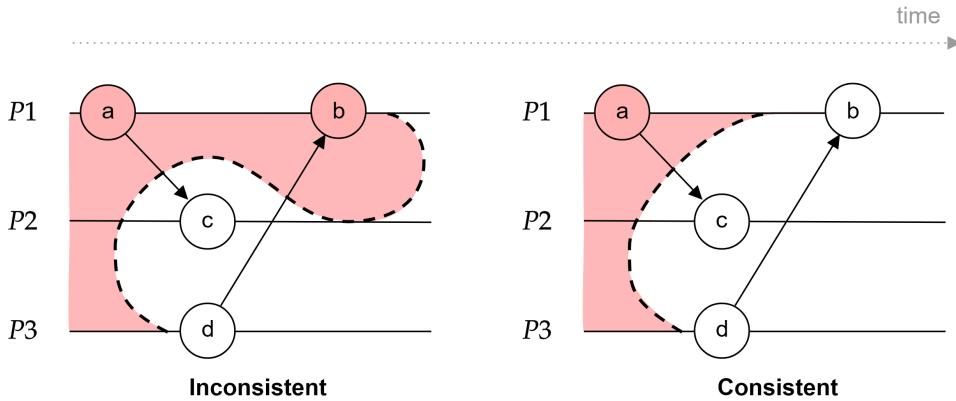


Figure 2.1: Inconsistent vs. Consistent Snapshots (pre-snapshot highlighted in red)

Here, in the inconsistent snapshot, b is included without d , its causally preceding event. In the consistent snapshot, only a is included. In this snapshot c and d could be added without violating causality.

There are many snapshot protocols, but we will focus on the **Chandy-Lamport Algorithm**.

Definition 1.3: Chandy-Lamport Algorithm

The **Chandy-Lamport Algorithm** is a snapshot algorithm that is used in distributed systems for recording a consistent global state of an asynchronous system. In this protocol:

- The snapshot procedure does not disrupt other processes.
- Each process records its local state.
- Any process can initiate the snapshot.

This model requires:

- **No Failures:** No failures during the snapshot.
- **First in, First out Channels (FIFO):** no lost or duplicated messages.
- **Strongly Connected Network:** All processes can reach every other process.
- **Single Initiator:** Only one process can initiate the snapshot.

The initiator then:

- Sends a **Marker** message to all outgoing channels.
- Records local and incoming channel data.

Recipients of the marker message:

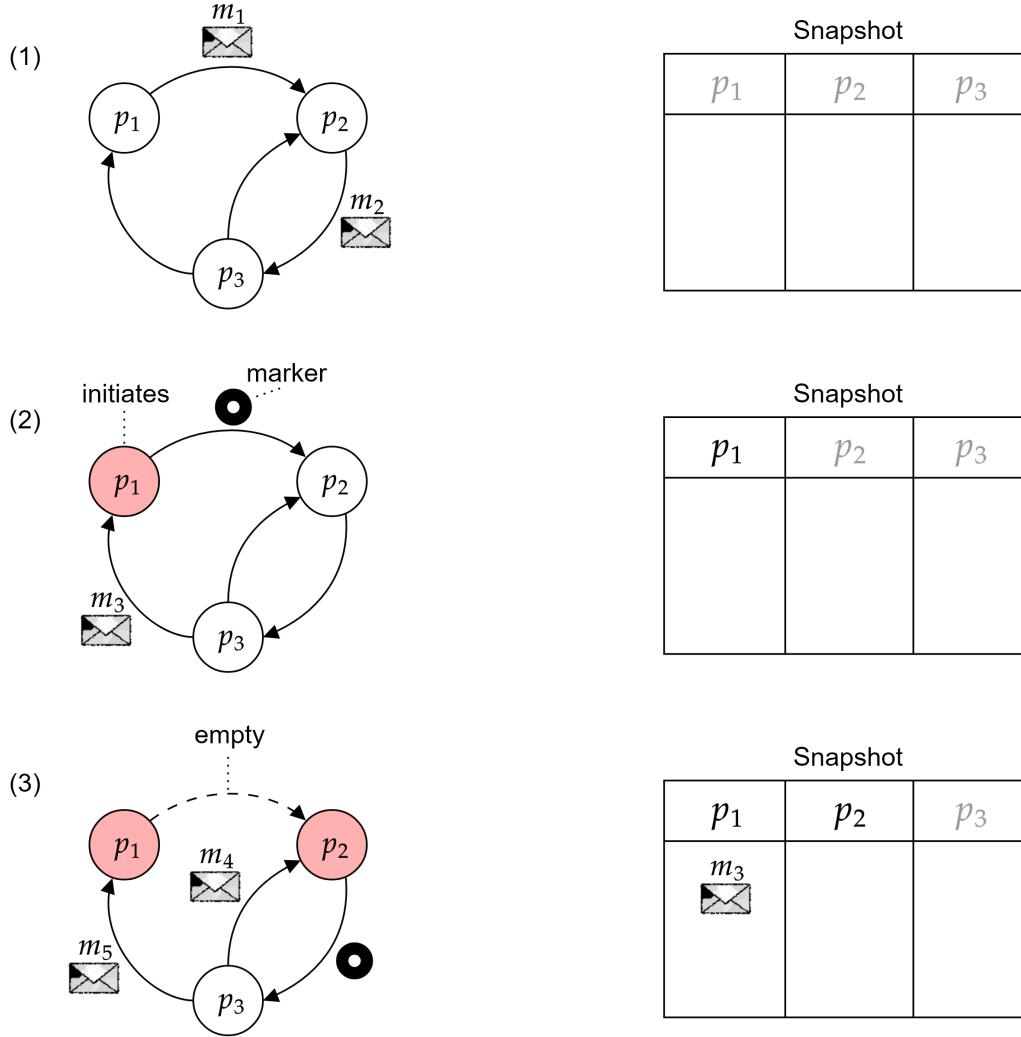
- Designates the channel which the marker arrived as **empty**.
- Records their local state and all other incoming channels except the empty one.
- Sends the marker to all outgoing channels.

Completion: When all processes have received and sent a marker, the snapshot is complete. This means every processes incoming channel is empty, hence the conclusion of the snapshot.

Important Notes:

- Subsequent markers received are ignored, **only the first** marker received is acted upon.
- Without the FIFO property, restarting the system may not be possible, as it relies on the order of messages to reproduce state.

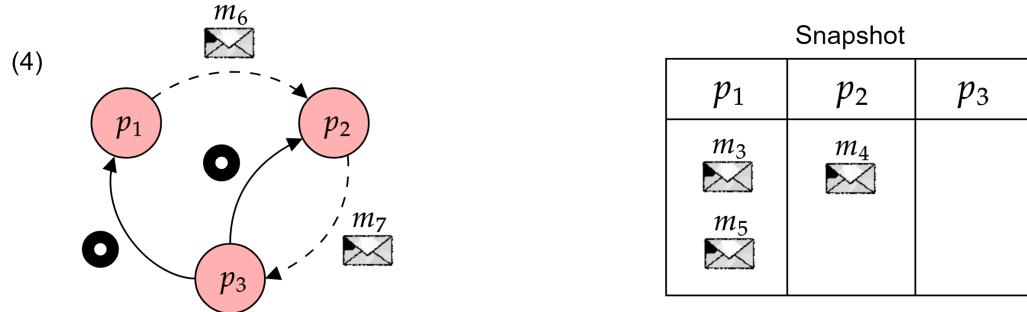
Observe the following illustration of the Chandy-Lamport Algorithm given processes p_1, p_2, p_3 :



Examining these first three steps: (1) The system before the snapshot. (2) p_1 initiates the snapshot sending a marker to p_2 and recording its local and incoming channel's state. (3) p_2 receives the marker, designating the incoming channel from p_1 as empty, sends the marker on outgoing channels, and begins recording. We also see, p_1 has recorded an incoming message m_3 from (2).

We continue on the next page.

We continue the snapshot process with the following steps:



(4) p_3 received the marker from p_2 , and designates that incoming channel as empty. It begins recording state, sending out the marker to p_1 and p_2 . Take notice that messages m_4 and m_5 have been recorded from (3). (5) All channels are now empty, concluding the snapshot. Take note that m_6 and m_7 were not recorded as they were sent on empty channels.

Theorem 1.1: Replay of a Snapshot

When replaying a snapshot captured using the Chandy-Lamport protocol, messages in transit (i.e., sent by a task whose state was recorded, but not yet received by others) **must also be replayed**.

In particular, if task A sends a message to task B , and only A is included in the snapshot, then the message is considered in transit and must be delivered to B upon replay.

Consider the following illustrations and think about how many messages may be replayed:

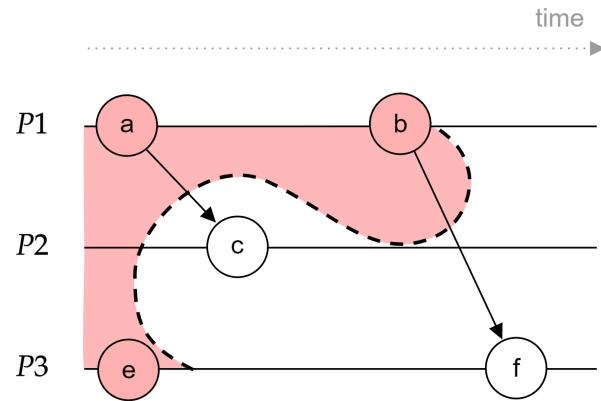


Figure 2.2: tasks *a*, *b* and *e* are recorded in the snapshot. Both *a* and *b* send messages from which aren't recorded in the snapshot. After reloading the snapshot, *a* and *b*'s messages **must be replayed**.

2.2 Replication: Synchronizing State

This section discusses replicating state across distributed systems.

Definition 2.1: Replication

Replication is the process of maintaining multiple copies of the same data on different nodes (machines). This is done for fault-tolerance, load balancing, and data locality.

Problem Space:

Consider we are running a money transfer service, from which Alice and Bob interact with:

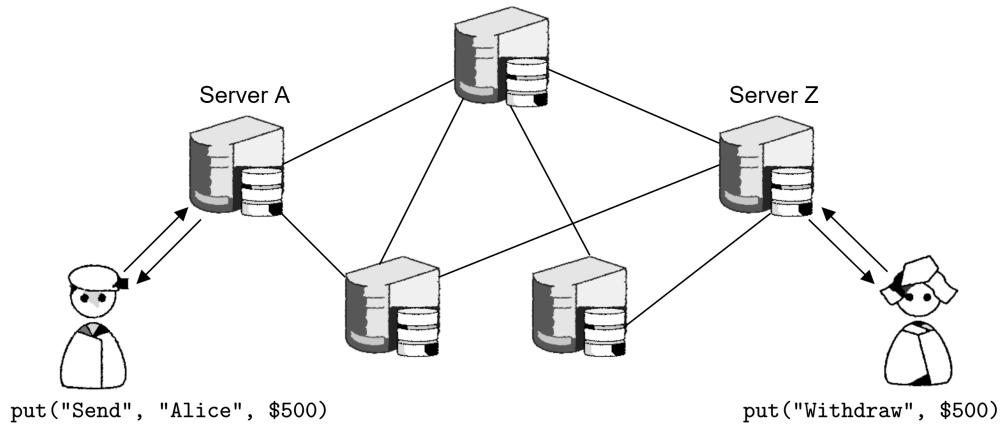


Figure 2.3: Bob sending money to Alice, while she withdraws money on two different servers.

In this scenario Many problems can arise: What if,

- The respective servers crash while or after Alice or Bob make requests?
- Alice and Bob share the same account, how do we ensure consistency?

In all, how do we ensure the propagation and synchronization of state across multiple servers? We consider two models:

Definition 2.2: Active vs. Passive Replication

Active Replication: Client sends requests to all servers at the same time and waits for acknowledgments. This method must ensure that all requests are processed in the same order (expensive).

Passive Replication: Client sends requests to a primary server, which then forwards the request to backup servers.

We'll move forward with **Passive Replication** as the preferred choice in this section. Though it is less expensive than active replication, it still has its challenges:

- **Consistency:** How do we ensure all backups are consistent?
- **Failure Handling:** What if the primary server fails?
- **Performance:** How do we ensure that the system is performant as we scale backups?

We consider two methods of replication:

Definition 2.3: State vs. Request Replication

State Replication: Forward the entire state to backups. This results in large message sizes, but is relatively simple depending on the system.

Request Replication: Forward only requests to backups. This results in smaller message sizes, but adds complexity when requests are not deterministic (e.g., random number generation).

Still again, we run into the following problem:

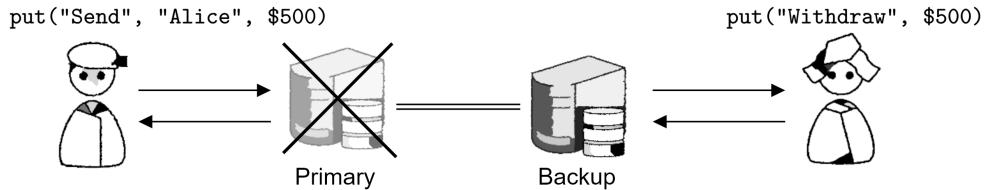


Figure 2.4: Bob's primary server failing, as Alice accesses the backup server.

How do we ensure both parties receive consistent feedback, even when the primary server fails?

Definition 2.4: Commit Point

The point at which the client is committed to a transaction, goes as follows:

1. The client sends a request to the primary server and waits for an acknowledgment.
2. The primary server forwards the request to the backup servers.
3. The backup servers process the request and send an acknowledgment to the primary server.
4. The primary server sends an acknowledgment to the client.

Step 4 is considered the **commit point**.

The following Figure illustrates the commit point in action:

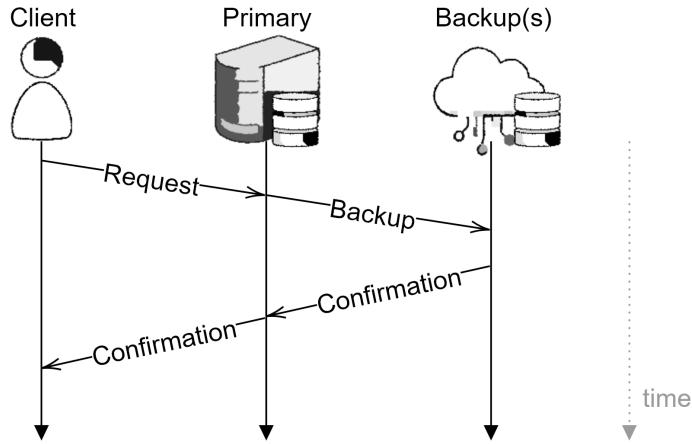


Figure 2.5: Client's requests propagating through the primary and backup sever.

We now can confirm that the client's request has been processed by the primary and backup servers, though this creates a lot of overhead as we scale the number of backups.

Now that some level of consistency is ensured, we want to orchestrate designating a new primary server when the current primary server fails:

Definition 2.5: Arbitration Server and Primary Election via Test-And-Set

In a distributed system with a primary-backup model, maintaining a **single active primary** is crucial. When a primary server fails, a **backup must be promoted**, but only if the failure is confirmed. An **Arbitration Server**, also called **Configuration Service Provider (CFG)** ensures a controlled **failover** (switching to a backup) process using **test-and-set**. It goes as follows:

- If the **primary fails** (or becomes unreachable), the CFG.
- The backup executes a `test_and_set(f, 1)` operation on a shared variable f .
 - If $f = 0$, the backup sets $f = 1$ and becomes the new **primary**.
 - If $f = 1$, another server has already taken over, preventing duplicate primaries.
- Clients redirect their requests to the new primary.

To demonstrate test-and-set in action, consider the following illustration:

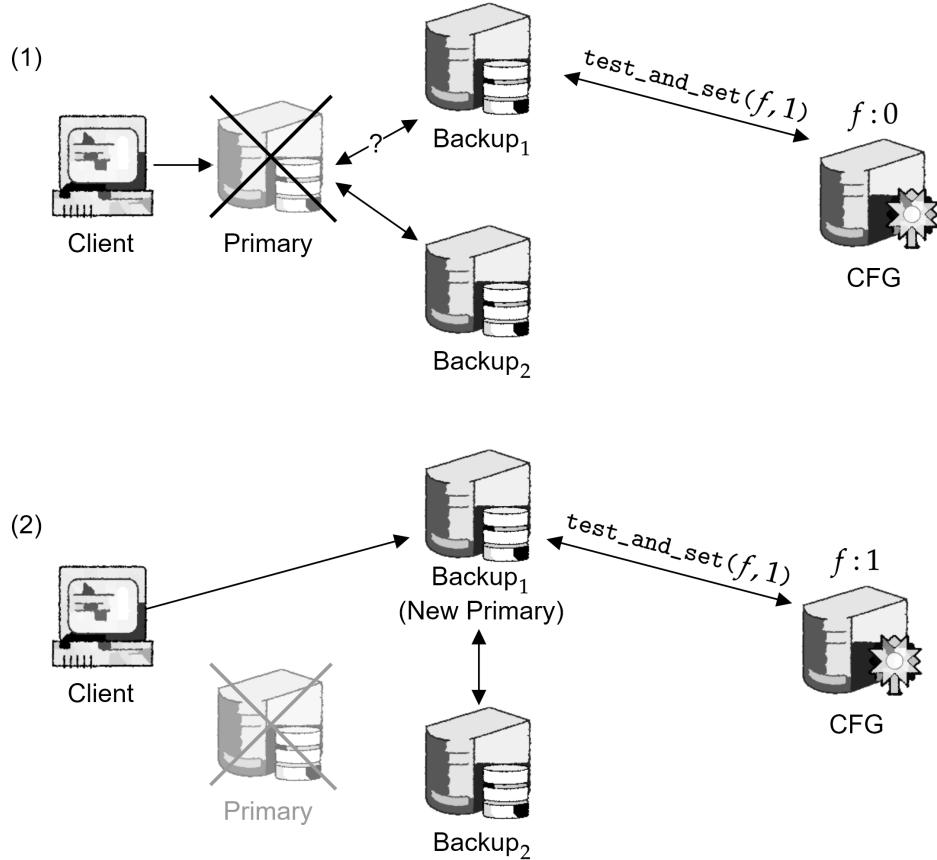


Figure 2.6: Backup server 1, is unable to connect to the primary, sends promotion request to CFG. All servers are connected to the CFG with access to the shared variable f . While the primary server is reachable, $f = 0$. When the primary fails, the backup server executes `test_and_set(f, 1)`.

A mutex is crucial to avoid data-races with multiple servers attempting promotion:

```

1  var flag int = 0; var mu sync.Mutex;
2  func (s *Server) TestAndSet(f *int, reply *int) error {
3      mu.Lock()
4      *reply = 0
5      if flag != *f {
6          flag = *f
7          *reply = 1
8      }
9      lock.Unlock()
10     return nil
11 }
```

Now what if the primary server partially processed a request before failing? Consider the following:

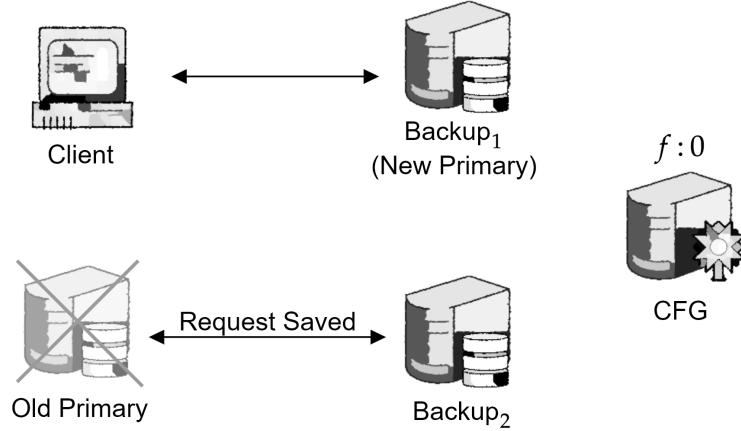


Figure 2.7: Primary server failing after partially processing a request.

To combat this issue, we discuss the **Chain Replication** method:

Definition 2.6: Chain Replication

The **Chain Replication** architecture servers are organized in a linear sequence, forming a chain: s_1, s_2, \dots, s_n , where s_1 is the **head** and s_n is the **tail**:

- **Write Operations:** Clients send write requests to the head (s_1), which is forwarded to s_2 , and so on, until the tail (s_n).
- **Read Operations:** Clients directly read from the tail (s_n), ensuring fully processed states; Specifically as s_i states may not be fully propagated through the chain.

Failover Handling:

- If server s_i fails, s_{i-1} forwards to the successor s_{i+1} instead.
- If the head (s_1) fails, the CFG promotes s_2 to be the new head.
- If the tail (s_n) fails, the CFG designates s_{n-1} as the new tail.

Tip: OSDI 2004 featured Robbert van Renesse and Fred B. Schneider from Cornell University, presenting Chain Replication. Van Renesse, originally from the Netherlands, and Schneider, from the U.S., have significantly influenced distributed systems research.

Consider the comparison between **Primary-Backup** vs. **Chain Replication**:

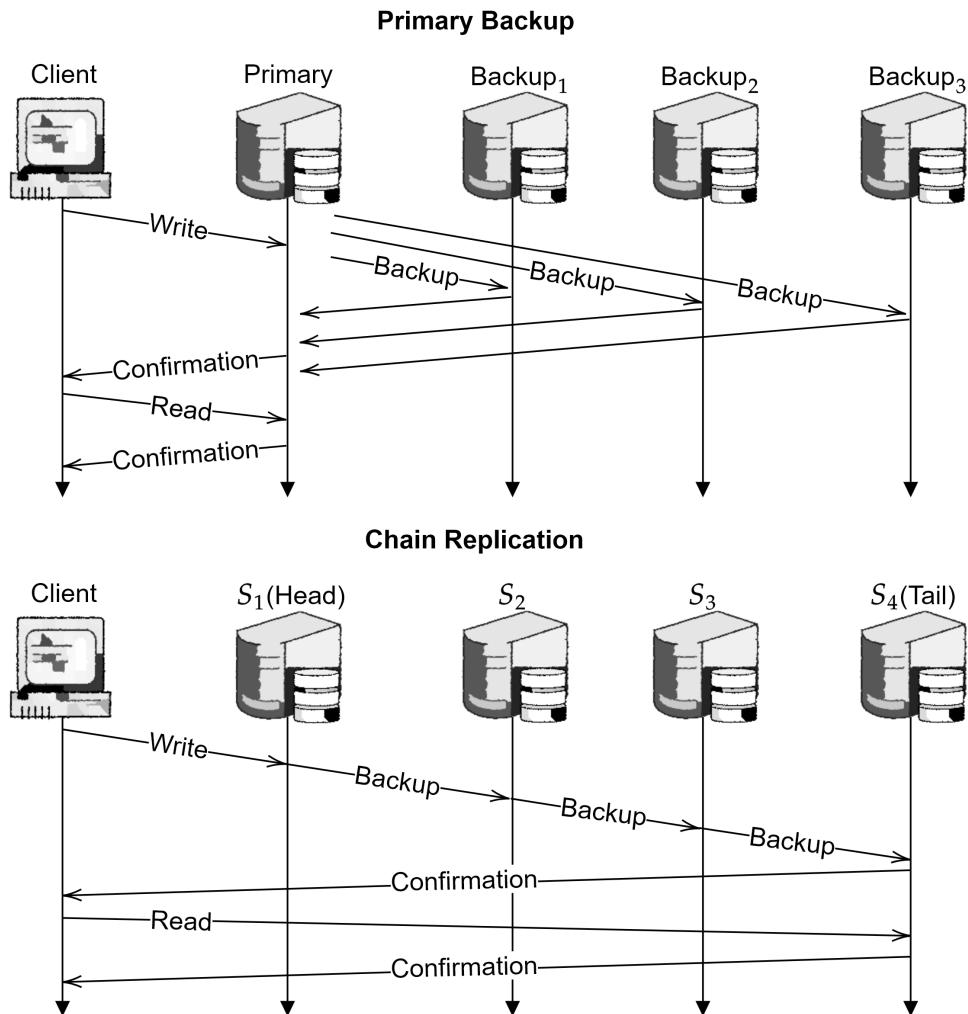


Figure 2.8: Primary-Backup vs. Chain Replication

Theorem 2.1: Primary-Backup vs. Chain Replication

In terms of **Latency**, **Primary-Backup** is faster as the request can theoretically be processed in parallel. However, **Chain Replication** is more **Fault-Tolerant**. If Reads/Writes are 50/50, then Chain Replication has **faster throughput** as it can split up reads (to the tail) and writes (to the head).

The Big Problem with Chain Replication:

There are some major assumptions that the Chain Replication Protocol makes:

- **Fail-Stop Servers:** Failures are complete and detectable (does not address slow servers).
- **Centralized Configuration Service:** Assumed never to fail.
- **Failure Detection:** Failures are easy to identify; **However**, distinguishing temporary vs. permanent failures is challenging.
 - Quick failure detection prevents stalls.
 - Over-eager detection may cause unnecessary data copying.

We will address these in the next section, discussing the **Raft Consensus Algorithm**.

2.3 Raft Protocol: Consensus Replication

As we've discussed so far, replication consistency is crucial and a difficult task. The next strategy, **Raft**, deals with such problem. There are other solutions, however, Raft is considered safe and easier to implement correctly than other solutions:

Definition 3.1: State Machines & Consensus Algorithms

A **State Machine** processes sequence of inputs from a **log** and saves them in state. **Replicated state machines** are implemented via **replicated logs** across multiple servers utilizing a **Consensus Model**, which facilitate data agreement between nodes.

Hence, commands need be **deterministic**, and the consensus algorithm **must be** flexible enough to handle and replicate application specific data.

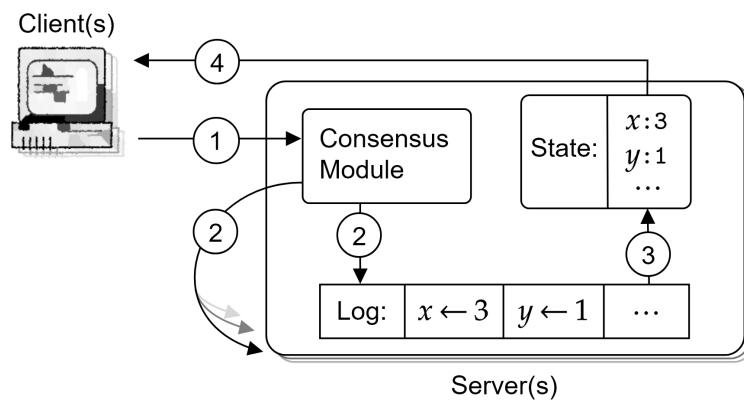


Figure 2.9: High-level framework for replicated state machines.

In the above the **Consensus Module** communicates with other servers to serve consistent logs.

2.3.1 Procedure Outline: Heartbeats, Elections, & Log Replication

Now to define the parts which make up a consensus algorithm:

Definition 3.2: Consensus Algorithm Components

A **Consensus Algorithm** involves the following components:

- **Safety:** Always returns correct results in spite of, network delays, partitions, duplications, and reorderings.
- **Liveness:** A **Cluster** (group of servers) must tolerate a subset of server failures (e.g., A cluster of 5 servers can tolerate 2 failures). Offline servers may later recover and rejoin the cluster.
- **Time Agnostic:** The algorithm must not rely on synchronized clocks.
- **Majority Rule:** A majority of servers must agree on a value before it is committed. Minority slow servers must not block the system.

At a high level, The Raft Algorithm:

Definition 3.3: Raft Abstract

The Raft Algorithm involves three main components:

- **Leader Election:** A leader ℓ is elected to manage the replication process of backups β .
- **Heartbeats:** Where ℓ and β exchange consistent pulses of data to ensure liveness.
- **Assurance:** Commit points (2.4) are established between the client, ℓ , and β .

In Raft, servers are given roles to manage the replication process.

Definition 3.4: Raft Server States

A Raft server can be in one of the following states:

- **Follower:** A server that listens to the leader.
- **Candidate:** A server that is running for leader.
- **Leader:** A server that is managing the replication process.

Followers are passive and simply listen to the leader.

Definition 3.5: Raft Heartbeats

Raft servers exchange **heartbeats** to ensure liveness. Each server on boot is randomly assigned a timeout value. After the initial timeout, a server sends a signal akin to the out-and-in beats of a heart. **Out-beats** are origin signals, and **in-beats** are return signals. If a server β receives an out-beat from another server ℓ before its in-beat, the β timer resets and follows ℓ 's beat. The leader of the beat's timer is not used.

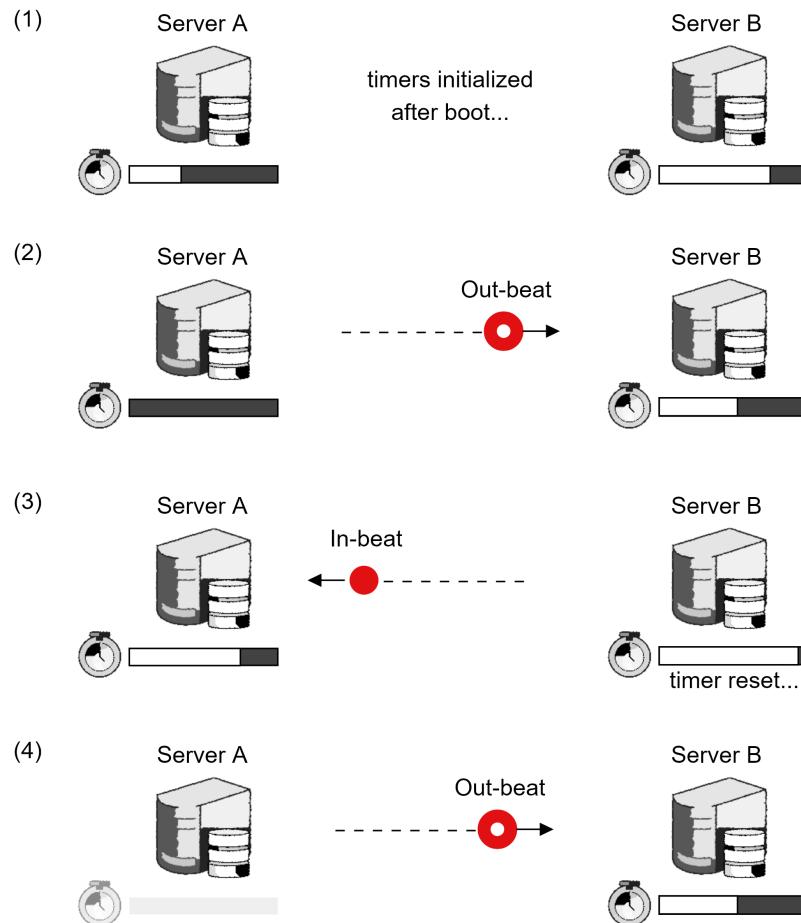


Figure 2.10: Server A and B exchanging heartbeats after system initialization (1). Server A 's timer runs out first, sending the first signal to B (2). B has received A 's out-beat, and now follows them. B resets their timer, and returns the signal to A (3). A 's timer isn't effect, nor is it used whilst a leader of the beat (4).

Definition 3.6: Raft Election Terms

Let us denote an arbitrary server as β , then β is either of class γ (**Follower**), ς (**Candidate**), or ℓ (**Leader**). I.e., we have types $\{\beta : \gamma \mid \varsigma \mid \ell\}$. The Raft Election Process involves the following:

1. **Timeout** ($\gamma \rightarrow \varsigma$): First, all β initialize as γ with a random timer (e.g., 150-300ms). Once the timer runs out, γ transitions to ς .
2. **Candidate Election** ($\varsigma \rightarrow \ell$): ς votes for itself and sends a **RequestVote RPC** to all β , informing them of the new term and approval request. All β vote once per term for the RPC they received first. If ς receives the majority vote, it transitions to ℓ .
3. **Split Vote**: If all top ς tie in votes, all ς timeout, waiting for the next term. At this point, if a γ 's timer runs out before the ς turn-around, then $\gamma \rightarrow \varsigma$ starting a new term election.
4. **Behind Leaders & Candidates** ($\ell \rightarrow \gamma, \varsigma \rightarrow \gamma$): If a ℓ or ς ever receives a heartbeat from a higher term β , it reverts to γ . Additionally, lower term signals are **ignored**.
5. **Leader Timeout** ($\gamma \rightarrow \varsigma$): If γ does not receive a heartbeat from the ℓ (server failure) whilst their timer runs out, they transition to ς and start a new election.

Let's observe a cluster of three servers:

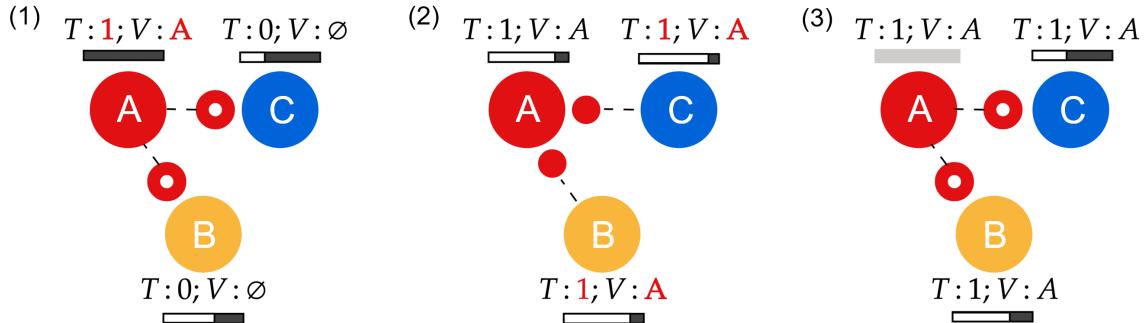


Figure 2.11: First Election after boot where A , B , and C nodes in a cluster of three have been initialized. (1) A times out first, becoming a Candidate and sending a RequestVote RPC to B and C . (2) B and C vote for A . (3) A receives the majority vote and becomes the Leader. Now A 's timer is no longer used.

Observe a cluster of four who is dealing with a split vote:

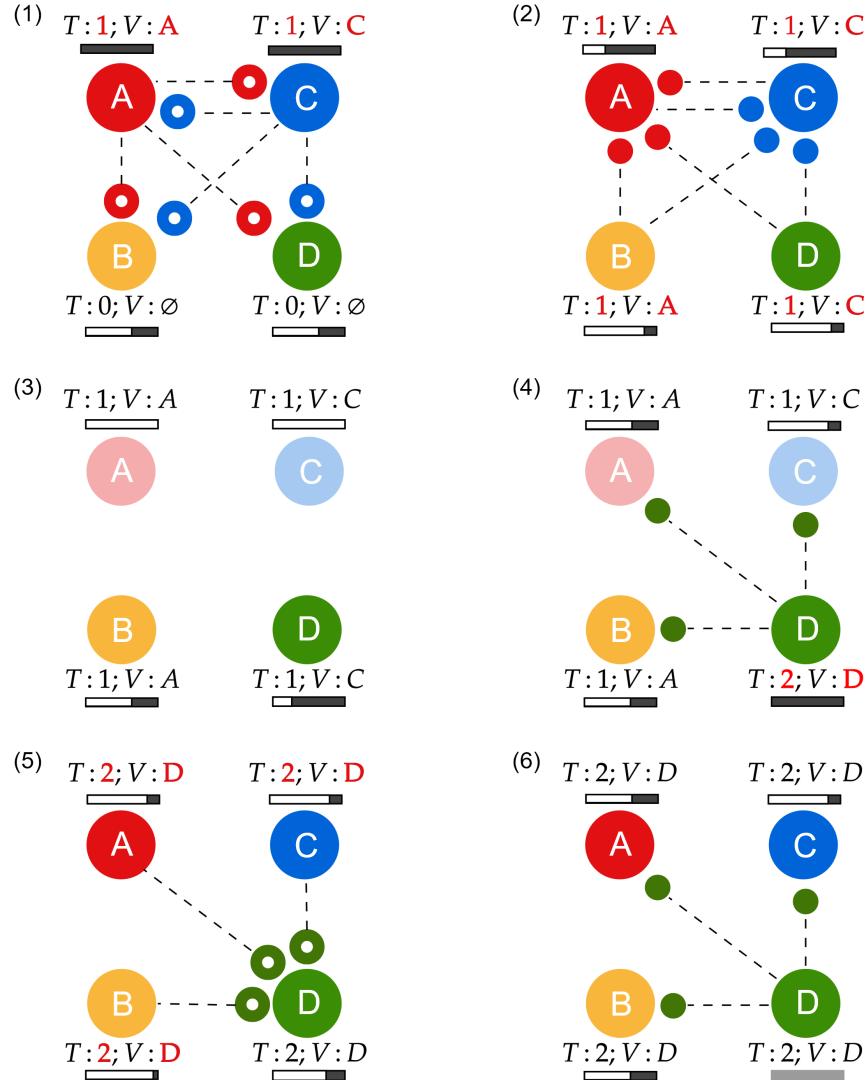


Figure 2.12: (1) A and C 's timers run out first, sending out a RequestVote RPC to all other servers. (2) B votes for A , and D votes for C . (3) A and C timeout as per split vote. (4) D times out starting a new term, casting their Request RPC. (5) All servers respond to the higher order term. (6) D becomes the Leader (Their time no longer used).

Definition 3.7: Log Replication

Given $\{ \beta \text{ (Server)} : \gamma \text{ (Follower)} | \varsigma \text{ (Candidate)} | \ell \text{ (Leader)} \}$, the Raft Log Replication Process involves the following:

- **Leader Log Replication:** Once a ℓ is elected, it services command requests from the client. The ℓ first appends the command to its **indexed log**, then sends an **AppendEntries RPC** in parallel to all β . Once a majority of β have replicated the log, the command is **committed**, and thus, safe to apply to their state machine.
- **Order of Execution:** The ℓ maintains a `nextIndex[]`, indicating the next expected log entry for all β . The ℓ sends missing log entries to lagging β ; otherwise, it sends empty **AppendEntries RPCs** as heartbeats.
- **Leader Redirection:** All β that receive client requests, redirect the client to the ℓ .

Theorem 3.1: Log Matching Property

If two entries in different logs have the same index and term, then:

- They store the same command.
- All proceeding entries are the same.

Definition 3.8: Log Correction

Given $\{ \beta \text{ (Server)} : \gamma \text{ (Follower)} | \varsigma \text{ (Candidate)} | \ell \text{ (Leader)} \}$,

If ℓ fails, the new ℓ may be missing log entries. Let i denote the last index of ℓ 's log entry, and j the last index of other β 's logs. If after an **AppendEntries RPC**, β 's $j \neq i$, then ℓ decrements β 's **nextIndex** to $(j - 1)$ **per call**, until $j = i$. Once majority coherence is met, the log is committed.

This holds as we assume the preceding Theorem (3.1) is true. This is to say, the leader has a—**Append Only Property**—that it does not modify its own entries, but **only** appends new ones.

Note: A slight optimization for log correction, is for γ to tell ℓ the conflicting term and its first index. With that, ℓ can skip the log entries of such term. Though this isn't necessary as in real world applications, these conflicts are infrequent.

2.3.2 Safety: Restricting Leader Election

Since previously discussed, a new leader may be missing log entries, from which they tell other servers to discard mismatches. This can become problematic when the previous leader and other servers have committed several entries.

To fix this, we introduce constraints on leader election:

Definition 3.9: Leader Completeness

All proceeding leaders must have all committed entries of the previous leader.

Definition 3.10: Leader Election Restriction

Given $\{ \beta \text{ (Server)} : \gamma \text{ (Follower)} | \varsigma \text{ (Candidate)} | \ell \text{ (Leader)} \}$,

New ℓ must demonstrate **Leader Completeness**. To ensure this, β upon receiving a **RequestVote RPC** from ς checks:

- If ς 's last log is from a lower term, the vote is **rejected**.
- If ς 's last log is from the same term, but shorter, the vote is **rejected**.
- If ς 's last log is from a higher term, then the vote is **accepted**.
- If ς 's log is **at least** as up-to-date as the voter, the vote is **accepted**.

Theorem 3.2: Present Term Commitment

Leaders cannot assume previous term entries are committed, as they may not have been fully replicated. Hence they only commit entries from their term, and by the **Log Matching Property (3.1)**, previous entries are also committed.

Theorem 3.3: Ensuring a Leader Through Timing

To ensure that there is always a leader in the face of probabilistic timing, these magnitude requirements should be satisfied:

$$\textit{broadcastTime} \ll \textit{electionTimeout} \ll \textit{MTBF}$$

$\textit{broadcastTime}$ measures the average time of heartbeat reciprocation (e.g., .5ms-20ms). Then $\textit{electionTimeout}$, the randomly assigned timeout times for each server (e.g, 10ms-500ms), and \textit{MTBF} , the average time between failures of a server (e.g., several months or more).

2.3.3 AppendEntries, State, and RequestVote RPC Schema

The next two pages are a summary of the Raft RPCs **AppendEntries**, **RequestVote**, and **state** for implementation:

AppendEntries RPC	
Invoked by leader to replicate log entries; also used as heartbeat.	
Arguments	Description
term	Leader's term.
leaderId	Allows follower to redirect clients.
prevLogIndex	Index of log entry immediately preceding new ones.
prevLogTerm	Term of prevLogIndex , entry.
entries[]	Log entries to store (empty for heartbeat; may send more than one for efficiency).
leaderCommit	Leader's commitIndex .
Results	
term	Current term, for leader to update itself.
success	true if follower contained entry matching prevLogIndex and prevLogTerm .
Receiver Implementation	
	<ol style="list-style-type: none"> 1. Reply false if term < currentTerm. 2. Reply false if log doesn't contain an entry at prevLogIndex whose term matches prevLogTerm. 3. If an existing entry conflicts with a new one (same index but different terms), delete the existing entry and all that follow it. 4. Append any new entries not already in the log. 5. If leaderCommit > commitIndex, set commitIndex = min(leaderCommit, index of last new entry).

STATE	
Persistent state on all servers (stored on stable storage before responding to RPCs)	
currentTerm	Latest term server has seen (initialized to 0 on first boot, increases monotonically).
votedFor	Candidate ID that received vote in current term (or null if none).
log[]	Log entries; each entry contains a command for the state machine and the term when it was received by the leader (first index is 1).
Volatile state on all servers	
commitIndex	Index of the highest log entry known to be committed (initialized to 0, increases monotonically).
lastApplied	Index of highest log entry applied to the state machine (initialized to 0, increases monotonically).
Volatile state on leaders (Reinitialized after election)	
nextIndex[]	For each server, index of the next log entry to send to that server (initialized to leader last log index + 1).
matchIndex[]	For each server, index of highest log entry known to be replicated on server (initialized to 0, increases monotonically).
RequestVote RPC	
Invoked by candidates to gather votes.	
term	Candidate's term.
candidateId	Candidate requesting vote.
lastLogIndex	Index of candidate's last log entry.
lastLogTerm	Term of candidate's last log entry.
Results	
term	Current term, for candidate to update itself.
voteGranted	true means candidate received vote.
Receiver Implementation	
<ol style="list-style-type: none"> 1. Reply false if term < currentTerm. 2. If votedFor is null or candidateId, and candidate's log is at least as up-to-date as receiver's log, grant vote. 	

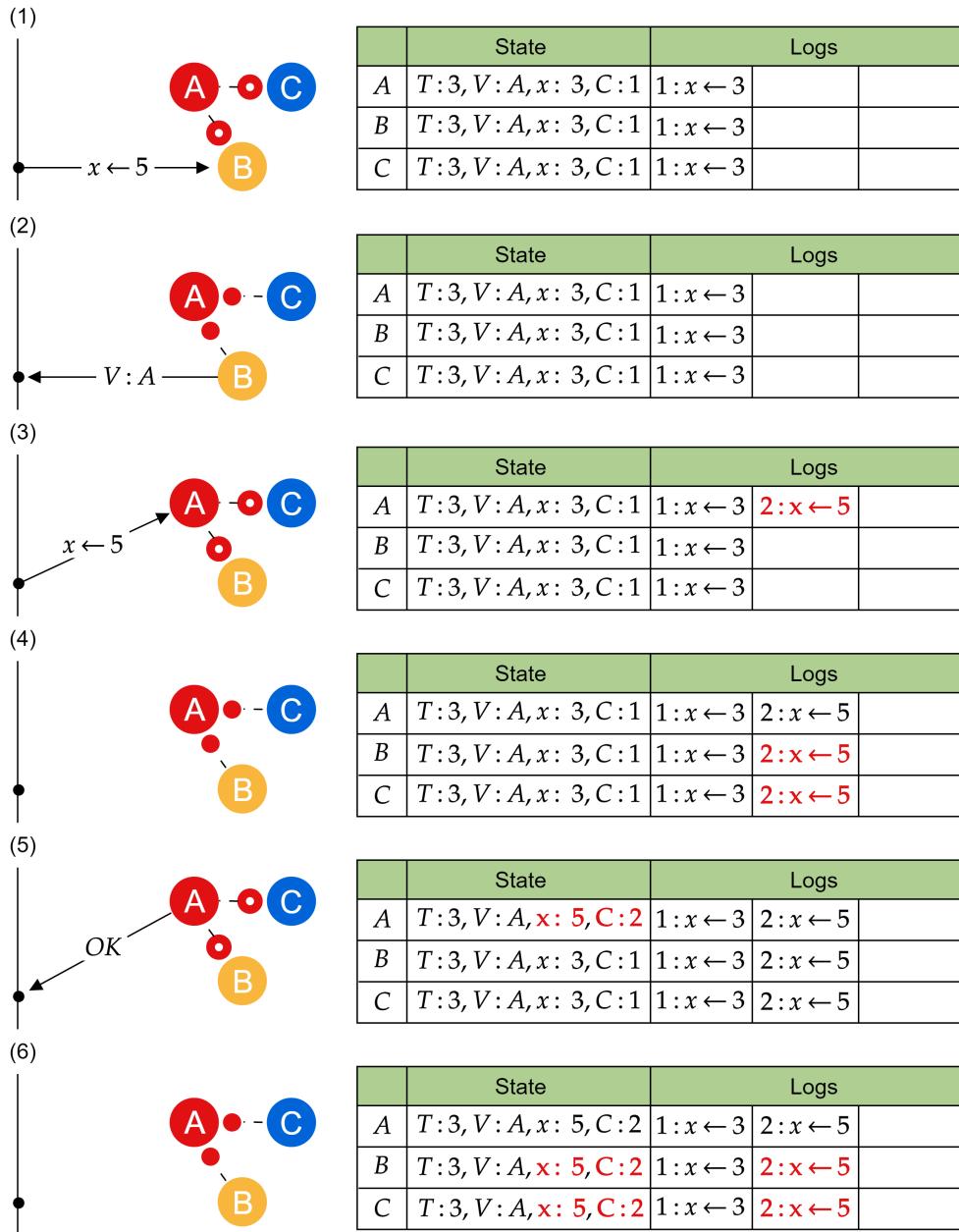


Figure 2.13: Simplified Raft server log replication where T is the current term, V is the voted for candidate, and C is the commit index. (1-3) The client's request is redirected from B to A . Then A propagates the command $x \leftarrow 5$ as log index 2. (4) both B and C replicate the log and send back a heartbeat. (5) A commits log 2, applies it to state, and sends an acknowledgment back to the client. (6) B and C update the newly committed index and apply it to state.

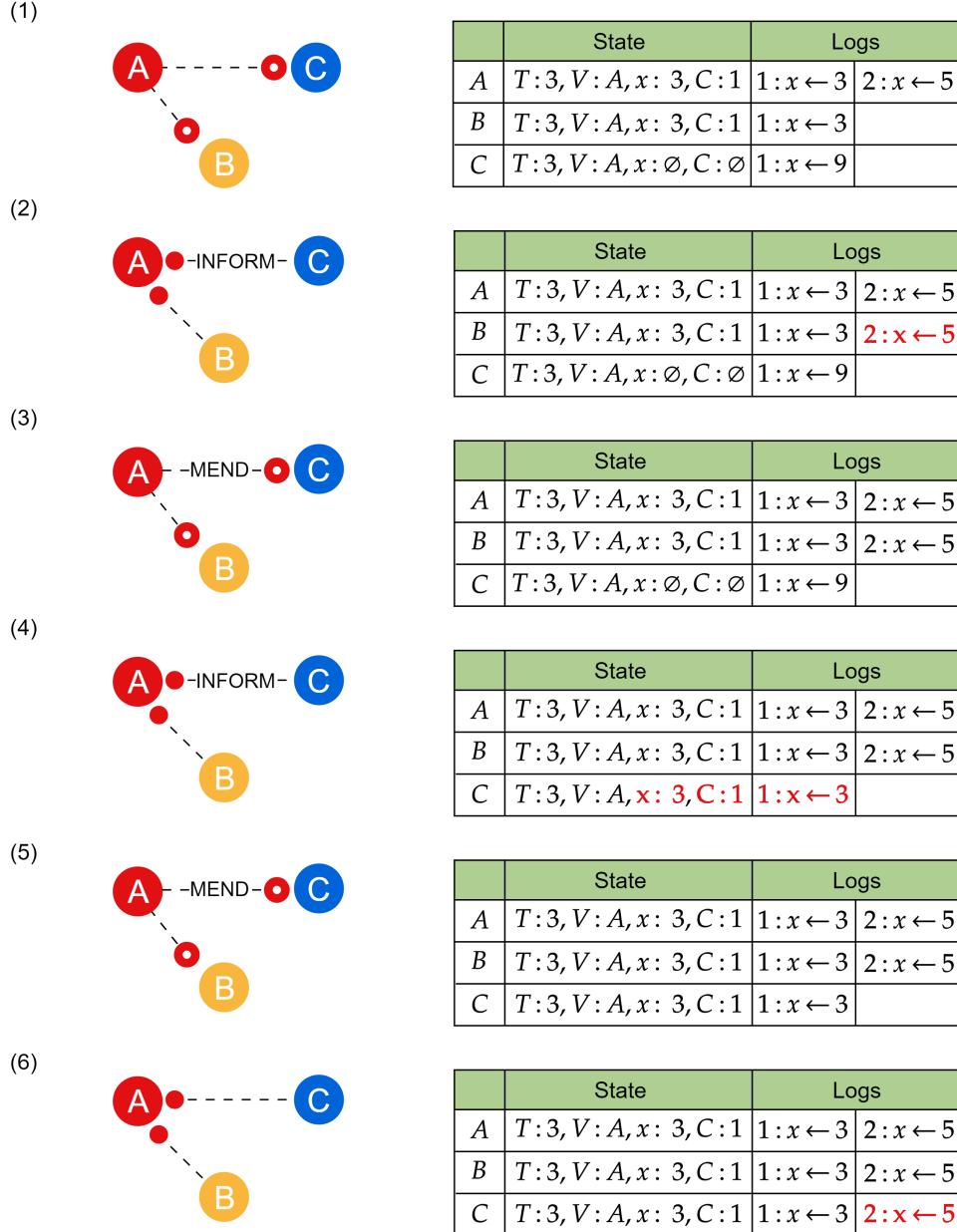


Figure 2.14: (1) Server C has an incorrect log, whilst server A sends out the command $x \leftarrow 5$. (2) Server C receives the command, but rejects it and informs A where they differ. (3) Server A re-sends the log from which they differ. (4) Server C corrects such entry, and informs A again. (5) Server A retries once more. (6) Server C receives the log and replicates it, and is thus up-to-date.

2.3.4 Cluster Reconfiguration (Adding, Removing, and Replacing Servers)

Cluster reconfiguration involves adding, removing, and replacing servers, or adjusting routines:

Definition 3.11: Stop and Restart & Pause and Resume

- **Stop and Restart:** Immediately halt operations, save state, reconfigure, and restart.
- **Pause and Resume:** Pause executing incoming requests and instead save them in a buffer, finish current processes, reconfigure, restart and resume.

Raft takes a two phase approach to reconfigure the cluster:

Definition 3.12: Raft Joint Consensus Reconfiguration (Part 1)

Joint consensus allows for a cluster to service requests while reconfiguring. It does this via a transition state, which is a combination of the old and new configurations (Joint consensus).

Reconfiguration: Once a leader receives a configurations request to swap from C_{old} to C_{new} , it appends $C_{old,new}$ to its logs, proceeding as follows:

- **Phase 1:**

1. Propagate $C_{old,new}$ to all servers.
2. Servers receiving new configurations (e.g., $C_{old,new}$), acknowledge it (**regardless** if it's committed or not).
3. A leader crashing leads to reelections under C_{old} or $C_{old,new}$.
4. Once $C_{old,new}$ is committed, the leader appends C_{new} to its log and replicates it.

- **Phase 2:**

1. C_{new} is committed and live, any server not under C_{new} can be shut down.
2. The leader who committed C_{new} becomes a follower to account for **Deletion**.

Deletion: To remove servers, they are excluded in the C_{new} configuration; **With the only exception** that leaders who do not exist in C_{new} may continue to manage the cluster until C_{new} is committed.

Addition: New Servers enter with **empty logs** and are added in the following manner:

1. The leader upon receiving the C_{new} request is notified of new server existence.
2. New servers **cannot vote**, but only receive logs (as to avoid splitting the cluster).
3. Once new servers are synchronized in logs, $C_{old,new}$ may be committed.
4. $C_{old,new}$ clusters require the majority vote from **both old and new servers**.

Though we must address the situation where servers outside of C_{new} try to cast votes:

Definition 3.13: Raft Joint Consensus Reconfiguration (Part 2)

Since Raft processes requests asynchronously, during the transition to the new configuration from $C_{old,new}$ to C_{new} , the following could happen:

- A server not yet notified of the new configuration may not know it has been removed from it. Hence, it may not receive heartbeats and attempt to start a new election.
- A server with some other configuration comes online and begins to send out RPCs to the cluster.

In any such case of unwanted votes, Raft enacts the following protocol to ensure safety:

- **Leader Liveliness:** If servers within C_{new} believe they have a leader, they ignore votes (regardless of higher order terms).
- **Majority Rule:** Even if an election occurs, C_{new} members will vote amongst themselves, for which servers outside of C_{new} will never win.

Ideally, we would like to immediately terminate servers that are removed, but this may not always be possible or practical.

2.3.5 Log Compaction & Snapshotting

In finite systems, logs cannot expand without bound. So as to clean log entries without losing state consensus, snapshotting is employed.

Definition 3.14: Raft - Log Compaction & Snapshots

Raft compacts logs via state snapshots (1.1). Servers independently take snapshots, and are not initiated by the leader. Snapshots employ the following protocol:

1. Trigger snapshots when logs reach a fixed size in **bytes** or time (e.g., see below tip).
2. The snapshot replaces **committed** log entries and retains entries yet to be committed.
3. The snapshot includes last committed log index, term of last index, and machine state (e.g., key-value pairs).

Tip: Simple Snapshot Initiate Protocol - One possible solution suggested by the Raft paper is to initiate the snapshot when the log reaches a certain size in bytes. Ideally a size significantly larger than the size of the expected snapshot. Since snapshots have a **fixed cost** in setup, it reduces the overall cost overtime if snapshots occur at fewer intervals.

Proof 3.1: Raft - Log Compaction & Snapshots

Leader initiated snapshots may overcomplicate its RPC and or make costly use of bandwidth. Since the leader has already vetted committed log entries, it is safe to assume that the resulting independent snapshots consisting of committed logs—are also safe. ■

Though it's not ideal to send an snapshots over the network, in some cases we might have to.

Definition 3.15: Raft Resynchronization (InstallSnapshot RPC)

There are two cases where a snapshot may be sent:

1. **Missing Entry:** If the leader has already compacted a log entry it needs to send to a server, it sends the snapshot instead.
2. **Server Recovery:** A server coming online may be severely behind, and thus, simply sending AppendEntries RPCs may take too long.

In such cases the leader will send a **InstallSnapshot RPC**. Such RPC includes the last log index. The receiver discards all its logs before the last log index, and retains the rest.

Copying the entire dataset is costly, as to avoid this, Raft recommends a lazy snapshotting approach:

Definition 3.16: Copy-on-Write (CoW)

Servers mark existing data as shared between a snapshot and live state. Only when a write (modification) touches shared data, does the server copy to preserve the snapshot.

Definition 3.17: Raft - Handling Stale Logs in Client Interactions

Raft ensures clients receive up-to-date logs from leaders via the following methods:

- **Command IDs:** Each command is assigned unique serial numbers. Each server maintains the last processed command ID and associated response for each client.
- **Syncing Leader Logs:** To avoid stale leader logs, two precautionary are taken:
 1. At the beginning of each term, the leader replicates a dummy *no-op* command log and waits for it to be committed.
 2. The leader sends responses only after exchanging heartbeats with the majority.

2.3.6 InstallSnapshot RPC Schema

Below is a summary of the Raft RPC **InstallSnapshot** for implementation:

InstallSnapshot RPC	
Invoked by leader to send chunks of a snapshot to a follower. Leaders always send chunks in order.	
Arguments	Description
term	Leader's term.
leaderId	Allows follower to redirect clients.
lastIncludedIndex	The snapshot replaces all entries up through and including this index.
lastIncludedTerm	Term of lastIncludedIndex .
offset	Byte offset where chunk is positioned in the snapshot file.
data[]	Raw bytes of the snapshot chunk, starting at offset .
done	True if this is the last chunk.
Response to Sender	
term	Current term, for leader to update itself.
Reciever Implementation	
<ol style="list-style-type: none"> 1. Reply immediately if term < currentTerm. 2. Create new snapshot file if first chunk (offset = 0). 3. Write data into snapshot file at given offset. 4. Reply and wait for more data chunks if done is false. 5. Save snapshot file, discard any existing or partial snapshot with a smaller index. 6. If existing log entry has the same index and term as snapshot's last included entry, retain log entries following it and reply. 7. Discard the entire log. 8. Reset state machine using snapshot contents (and load snapshot's cluster configuration). 	

Consider the below figure which illustrates a snapshot taken from a server A :

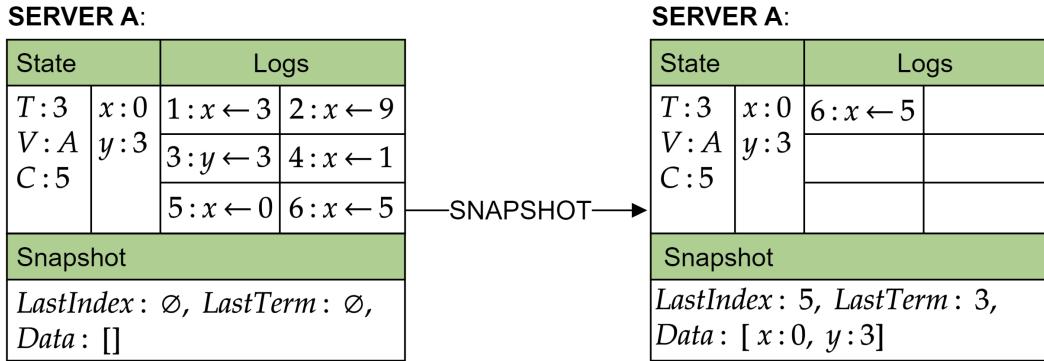


Figure 2.15: Server A begins with the initial state where, T is the current term, V is the voted for candidate, and C is the commit index. There is a separate partition for the snapshot where, $LastIndex$ is the last committed log entry's index, and $LastTerm$, the term from which $LastIndex$ resided. The arrow points towards after the snapshot, from which we see 6, the remaining log entry that wasn't committed.

2.3.7 Raft Algorithm Paper

Below is the original source with additional proofs and explanations [2]:

- <https://www.usenix.org/system/files/conference/atc14/atc14-paper-ongaro.pdf>
- **Extended Version:** <https://raft.github.io/raft.pdf>

And these two websites which offer interactive visualizations of the Raft Algorithm:

- <http://thesecretlivesofdata.com/raft/>
- <https://raft.github.io/>

2.4 Failure Models

2.4.1 Defining Failures

This section discusses, what to do or how to classify when failures occur.

Definition 4.1: Failure Model

A **Failure Model** is a set of assumptions about the types of failures that can occur in a distributed system. Such failures may be **Correlated** or **Independent**. Processes that do not fail are considered **Correct**.

In particular, we are concerned with the following types of failures:

- Crash and/or omission of responses, and how we might recover from them.
- Deviation protocol, arbitrarily or maliciously.

Definition 4.2: Crash-Stop Failure

A **Crash-Stop Failure** occurs when a process halts and does not resume (e.g., power failure, software crash). The failure is not explicitly detectable by other processes. **Permanent hardware failures** typically fall under category of crash-stop failures.

Definition 4.3: Fail-Stop Failure

A **Fail-Stop Failure** is a detectable crash-stop failure (e.g., timeouts, heartbeats).

Definition 4.4: Omission Failure

An **Omission Failure** can be categorized into two types:

- **Send Omission:** The process fails to send messages according to the protocol.
- **Receive Omission:** The process fails to receive messages that were sent by other processes.

In particular, The process itself may still be operational but unable to correctly communicate (e.g., network disruptions, software errors, or buffer overflows).

Definition 4.5: Crash-Recovery Failure

A **Crash-Recovery Failure** occurs when a process halts due to a crash but retains the ability to recover and resume execution.

- **Crash Phase:** The process halts in some way (e.g., stops sending or receiving messages).
- **Recovery Phase:** The processes may recover to the last correct state via snapshot (2.1). Certain types of memory may persist through the crash:
 - **Volatile memory** is lost during a crash (e.g., mid-execution variables).
 - **Stable storage** is retained through a crash (e.g., disk storage, assuming no disk failure).

However, if the processes crashes indefinitely, it is considered a **Crash-Stop** failure (4.2).

Definition 4.6: Byzantine Failure

A **Byzantine Failure** occurs when a process exhibits arbitrary or malicious behavior, leading to unpredictable system behavior.

- **Arbitrary Behavior:** The process deviates from the expected protocol, such as:
 - Sending corrupted or inconsistent messages to different nodes.
 - Updating its state in an unintended or unpredictable manner.
- **Malicious Behavior:** The process actively attempts to disrupt the system, such as:
 - Exploiting protocol vulnerabilities to manipulate outcomes (e.g., double-spending in a blockchain).

In short, these failures occur due to **bugs** (unintentional) or **attacks** (intentional).

Tip: The term **Byzantine** in computer science comes from the *Byzantine Generals Problem*, introduced by Leslie Lamport in 1982. It describes a scenario where generals must coordinate an attack but cannot trust all messengers—some may be traitors sending conflicting information.

The name *Byzantine* is inspired by the Byzantine Empire, which was historically known for its complex and often deceptive political intrigues. While the term is widely used in distributed systems, some argue it unfairly portrays Byzantine history.

In computing, a **Byzantine failure** refers to a system component acting unpredictably, whether due to bugs, faults, or malicious intent, making consensus difficult.

2.4.2 Failures Model Hierarchy

Here we compare failure models as extensions of each other, though we will omit **fail-stop** as it is more of a detection mechanism. To quickly recap:

- **Crash-Stop:** Process halts and cannot resume (undetectable).
- **Omission:** Process fails to properly communicate.
- **Crash-Recovery:** Process halts but can recover and resume.
- **Byzantine:** Process exhibits arbitrary or malicious behavior.

Theorem 4.1: Failure Model Hierarchy

The failure models can be arranged in a hierarchy, where each model is an extension of the previous one. The hierarchy is as follows:

$$\text{Crash-Stop} \subset \text{Omission} \subset \text{Crash-Recovery} \subset \text{Byzantine}$$

In particular,

- **Crash-Stop \subset Omission:** A crash-stop is a full crash rather than a partial communication failure.
- **Omission \subset Crash-Recovery:** During recovering the last correct state, volatile memory lost may exhibit omission-like behavior. Meaning some messages are lost due to “amnesia.”
- **Crash-Recovery \subset Byzantine:** as a process may recover and exhibit arbitrary or malicious behavior. Moreover, a Byzantine failure may be **any type of failure**.

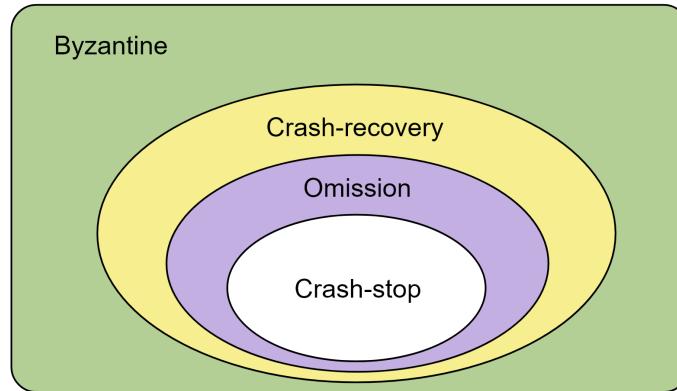


Figure 2.16: Failure Hierarchy depicting: $\text{Crash-stop} \subset \text{Omission} \subset \text{Crash-Recovery} \subset \text{Byzantine}$

2.5 Consistency Models

2.5.1 Introduction

Moving on from consensus models, which handle agreement on data values, we now shift our focus to the ordering of operations and their validity:

Definition 5.1: Consistency Model

Within a distributed system, a consistency model acts as a contract between systems on valid operation ordering (e.g., reads or writes) on shared data within the network.

These models are critical for ensuring that a system behaves as expected when replicating data across multiple nodes. [3]

Definition 5.2: Global Total Order

A Global Total Order is a sequence of operations that is agreed upon by all nodes in a distributed system. This means that such sequence given some client inputs, could reproduce the same state on a single machine.

Now we begin to list some common consistency models:

Definition 5.3: Strong Consistency

The client observes that all nodes *appear* to agree on the order of execution. This means all node reads of shared data are identical (Global Total Order of operations).

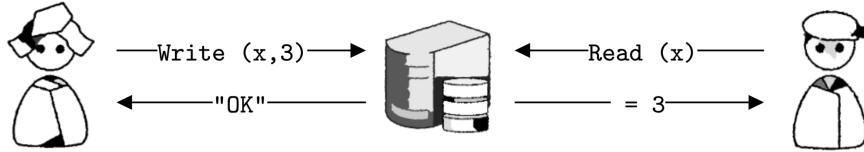
Definition 5.4: Weak Consistency

The client **temporarily** observes that nodes disagree on shared data values at some point in time (no Global Total Order).

Theorem 5.1: Strong Consistency vs Weak Consistency

Even though strong consistency is desirable, it can be costly on the network and difficult to implement. Weak consistency is easier to implement and may provide better performance, but at the cost of data integrity and difficulty in debugging interactions.

Consider this example interaction, which interaction has the weakest consistency model?



Find The Weakest Consistency

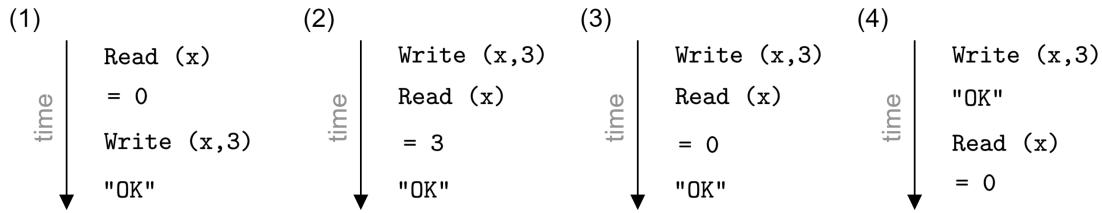


Figure 2.17: (1) Strong Consistency, as the read x could be 0, and the write responds with “OK”. (2) This is okay, as our write of 3 is read. The “OK” does come a little late, but its order isn’t bizarre. (3) Still okay, as the “OK” comes after the read of 0, meaning the write could not have happened yet. (4) The weakest, as our read completely disregards our acknowledged write of 3.

2.5.2 Strong Consistency Models: Linearizability & Sequential Consistency

We discuss two strong consistency models, Linearizability and Sequential Consistency:

Definition 5.5: Linearizability

Replicas produce a Global Total Order, which preserves real-time ordering of events. Moreover, every read must return the value of the **most completed** write. In particular:

- If operation A completes before B , then $A \rightarrow B$ in real-time.
- If tasks A and B overlap, there is no real-time ordering.

Theorem 5.2: Raft and Linearizability

Raft’s leader-commit design provides Linearizability, **except** in specific scenarios such as:

- A committed log entry is lost within the majority of the cluster.
- The newly elected leader somehow bypasses the Leader Completeness property.

In all other cases, Raft is considered to have a strong consistency model.

Consider the following examples and determine whether it is linearizable:

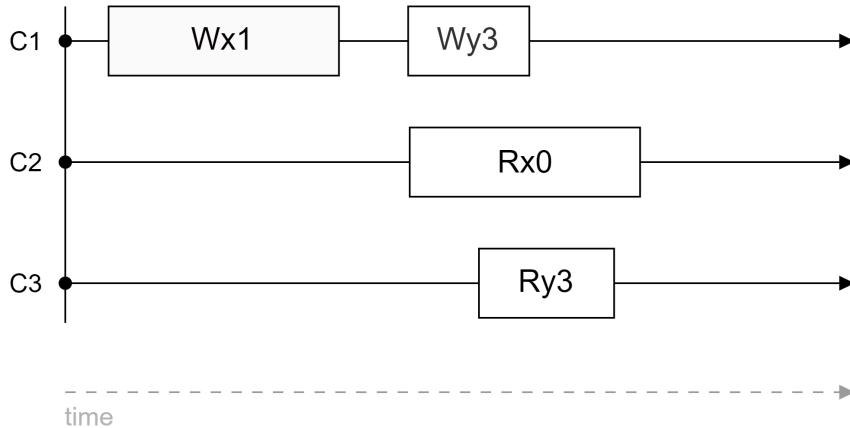


Figure 2.18: A distributed system with client C1, C2, and C3 interactions. Where $Wx1$ reads, “Write 1 to x” and $Rx0$ reads, “0 read from x.” This system is not linearizable because the read $Rx0$ should have returned 1, in respect to real-time.

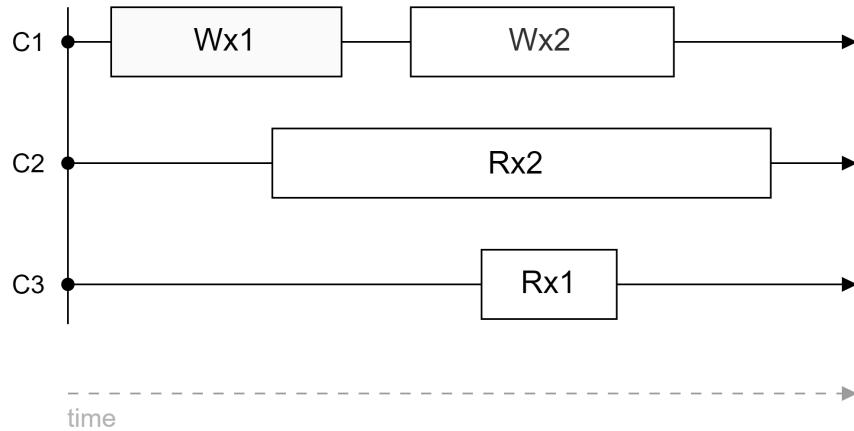


Figure 2.19: A distributed system with client C1, C2, and C3 interactions. Where $Wx1$ reads, “Write 1 to x” and $Rx1$ reads, “1 read from x.” This system is linearizable. This is because each box represents a period of time when the action could take place. Therefore, $Wx1$ and $Rx1$ can happen, while still having enough time for $Wx2$ and $Rx2$ to occur.

The [next page](#) includes an elaboration of the above.

Here we elaborate on Figure (2.19):

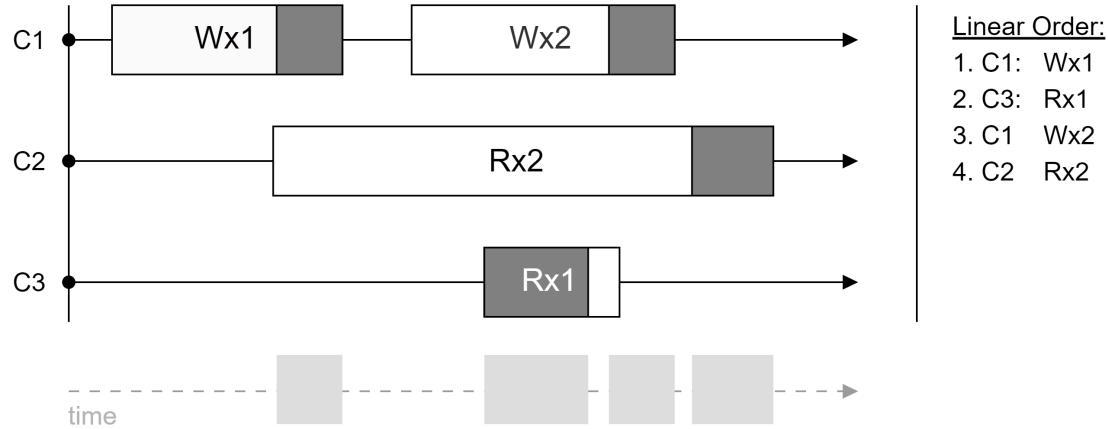


Figure 2.20: The gray boxes represent a mutex on the shared data, x . Here we see the order of operations, $Wx1$, $Rx1$, $Wx2$, and $Rx2$, from which **no overlaps occur** and logically make sense.

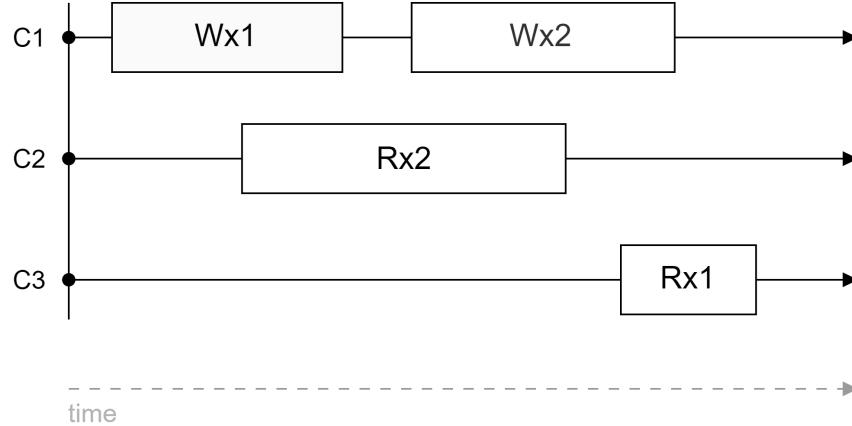


Figure 2.21: A distributed system with client C1, C2, and C3 interactions. Where $Wx1$ reads, “Write 1 to x ” and $Rx1$ reads, “1 read from x .” This system is not linearizable. The $Rx1$ is not possible as it’s too far disjointed from $Wx1$ after being overwritten by $Wx2$.

We've talked about another replication method before that is also linearizable:

Theorem 5.3: Chain Replication and Linearizability

Chain Replication is linearizable. The duration of a write operation extends until the tail, from which an acknowledgment is sent. Reads also adhere to reading the last completed write, and is thus—a strong consistency model.

The next model is Sequential Consistency:

Definition 5.6: Sequential Consistency

Weaker than Linearizability. All replicas execute all operations in **some** Global Total Order. Each client **observes the same order** once such order is agreed upon. Therefore a single machine may replicate the system if given such order. In particular:

- If a system process issues A before B , then $A \rightarrow B$ in the Global Total Order.
- If A and B happen on different processes, there is no real-time ordering (concurrent).

Consider the following examples and determine whether it is sequentially consistent:

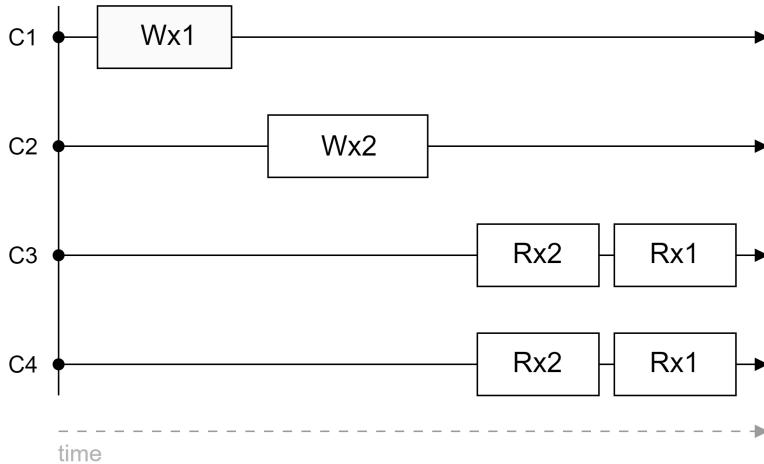


Figure 2.22: A distributed system with client C1, C2, C3, C4 interactions. Where $Wx1$ reads, “Write 1 to x” and $Rx0$ reads, “0 read from x.” This figure may or may not be sequentially consistent.

Next page includes an elaboration of the above.

We elaborate on the previous example:

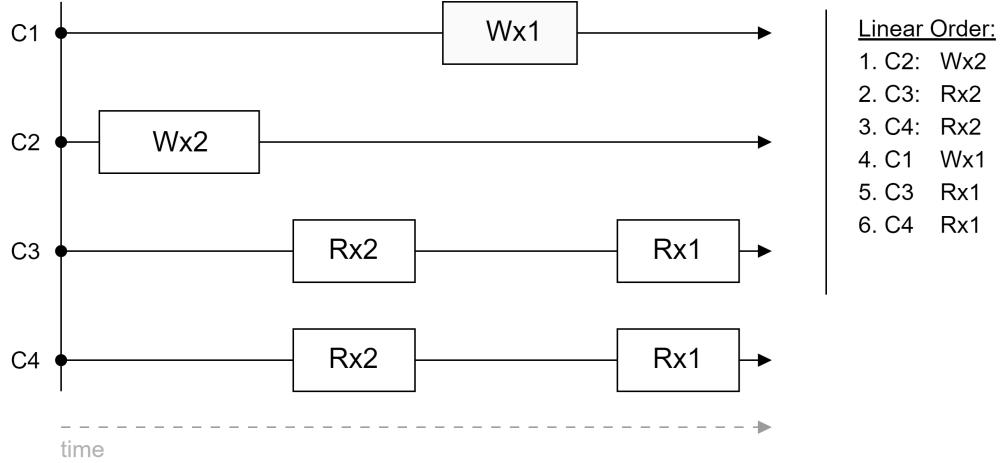


Figure 2.23: A distributed system with client C1, C2, C3, C4 interactions. Where $Wx1$ reads, “Write 1 to x” and $Rx1$ reads, “1 read from x.”. This figure is sequentially consistent as there exists some Global Total Order that can be agreed upon. Here listed, $Wx2$, $Rx2$, $Rx2$, $Wx1$, $Rx1$, $Rx1$.

And lastly, consider the following example:

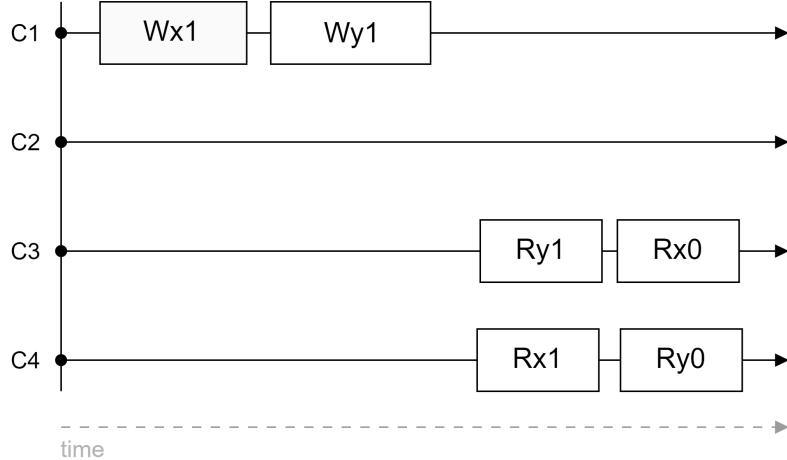


Figure 2.24: A distributed system with client C1, C2, C3, C4 interactions. Where $Wx1$ reads, “Write 1 to x” and $Rx1$ reads, “1 read from x.”. This figure is not sequentially consistent as there is no Global Total Order that can be agreed upon. Here $Rx0$ cannot happen after $Ry1$ because of the relation $Wx1 \rightarrow Wy1$.

Consider the following theorem:

Theorem 5.4: Linearizability vs Sequential Consistency

If a system is linearizable, it is also sequentially consistent. As, adhering to real-time order naturally satisfies sequential consistency. **However**, the reverse is not true. In particular:

- **Linearizability:** Relies on real-time.
- **Sequential Consistency:** Relies on program order.

2.5.3 Handling Shared Data via Mutex: Release & Lazy-release Consistency

Consider the following **Weak** methods of handling shared data:

Definition 5.7: Release Consistency vs. Lazy-release Consistency

These methods require explicit use of locks to propagate updates:

Release Consistency: Push updates to all nodes after releasing the lock.

Lazy-release Consistency: Push updates to all nodes after the lock is acquired. *Though this may provide less stress to the system, if the client does not lock data, it may become stale (weak consistency). Alternatively, if the client locks all data (strong consistency) though it may be costly. These both are weak consistency models.*

Determine whether the following system is release or lazy-release consistent:

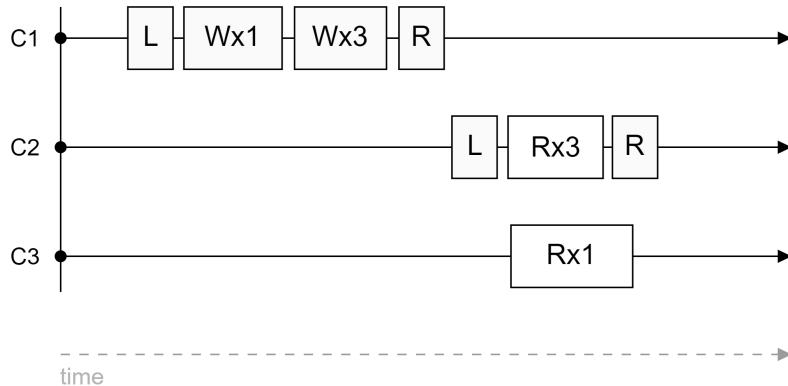


Figure 2.25: A distributed system with client C1, C2, and C3 interactions. Where $Wx1$ reads, “Write 1 to x” and $Rx1$ reads, “1 read from x.” With L (lock) and R (Release). This is Lazy-release consistent, as C3’s read of x wasn’t updated to 3 in absence of a lock.

2.5.4 Weak Consistency Models: Causal & Eventual Consistency

We now discuss two weak consistency models, Causal Consistency and Eventual Consistency:

Definition 5.8: Causal Consistency

Weaker form of Sequential Consistency. That the Global Total Order of operations adhere logically to their causal dependencies. In particular:

- If A causes B , then $A \rightarrow B$ should be in the Global Total Order.

For example, only one of these sentences makes causal sense:

(1)

1. Alice: Lunch?
2. Bob: Yes.
3. Carl: No.

(2)

1. Bob: Yes.
2. Alice: Lunch?
3. Carl: No.

(3)

1. Alice: Lunch?
2. Carl: No.
3. Bob: Yes.

Here, statements (1) and (3) are causally consistent, as the responses are in order of the question. Take that same intuition for this next example and determine whether the below system is causally consistent:

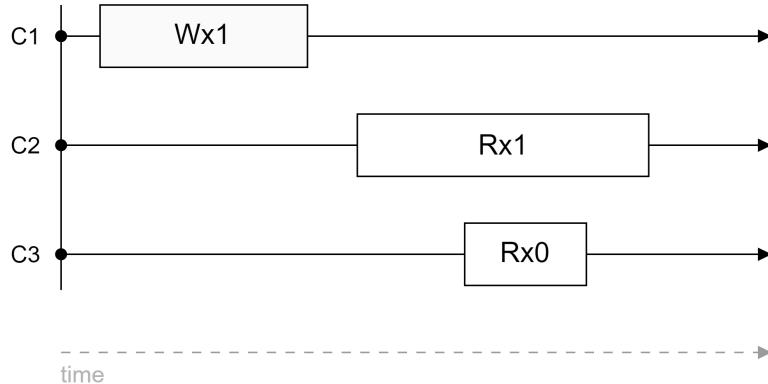


Figure 2.26: A distributed system with client C1, C2, and C3 interactions. Where $Wx1$ reads, “Write 1 to x ” and $Rx1$ reads, “1 read from x .” This system is causally consistent as the Global Total Order of operations adhere to their causal dependencies. Here, $Wx1 \rightarrow Rx1$, where $Rx0$ must come before $Wx1$.

Consider the following example and determine whether it is causally consistent:

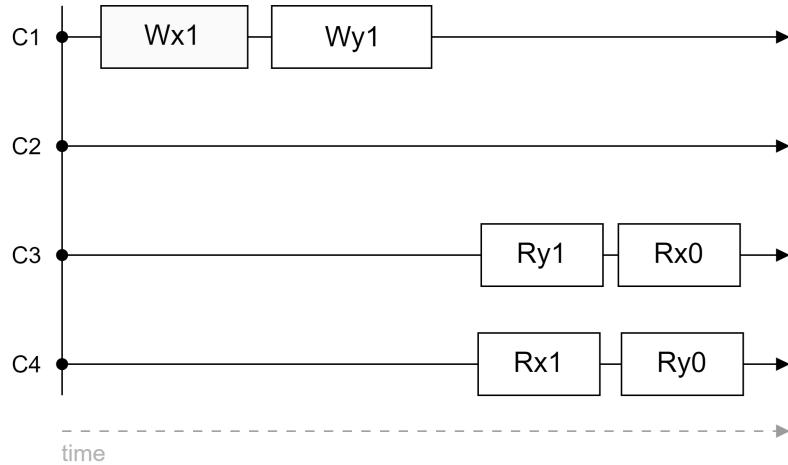


Figure 2.27: This System is causally consistent.

Definition 5.9: Eventual Consistency

Weaker than Causal Consistency. Given there are no new writes, replicas will **eventually** agree on the same value after some period of time.

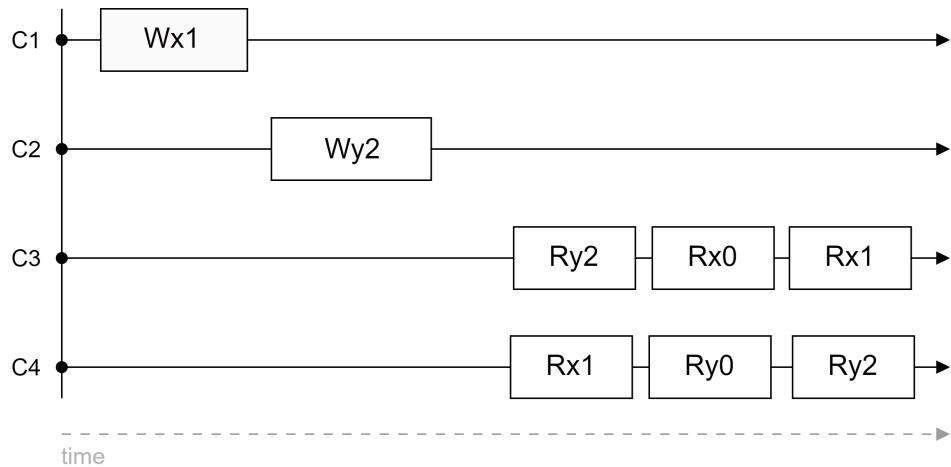


Figure 2.28: This system is eventually consistent as it eventually agrees on *x* and *y*.

Causal Consistency implies two properties:

Theorem 5.5: Causal Consistency → FIFO & RYW

Causal Consistency implies FIFO (First-In-First-Out) and RYW (Read-Your-Writes):

- **FIFO:** All writes are read in the order they were issued.
- **RYW:** If the client writes to x , then their next read of x must return the value they just previously wrote.

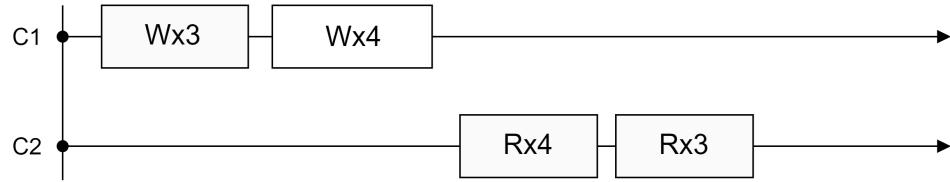


Figure 2.29: This system is not FIFO nor Causally Consistent. Writes **must be read in monotonic order**. Hence, Rx3 must come before Rx4.



Figure 2.30: This system is not RYW nor Causally Consistent. As the next read should be Rx4 alone. The order, Rx0 → Wx4 → Rx4, satisfies RYW and is Causally Consistent.

Bibliography

- [1] Ioannis Liagouris. Cs351: Distributed systems. Lecture notes, Boston University, Spring Semester, 2025. Boston University, CS Department.
- [2] Diego Ongaro and John Ousterhout. In search of an understandable consensus algorithm. In *2014 USENIX Annual Technical Conference (USENIX ATC 14)*, pages 305–319, Philadelphia, PA, June 2014. USENIX Association.
- [3] ScyllaDB. Consistency models definition, 2025. Accessed: 2025-03-25.