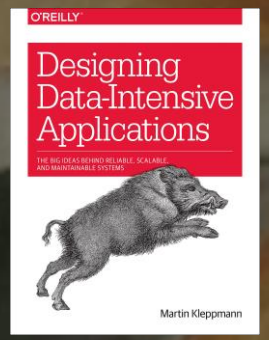
MSFT Sys Meetup



http://

Freely accessible resources

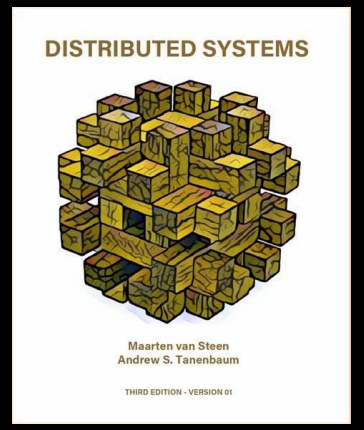
[Code]

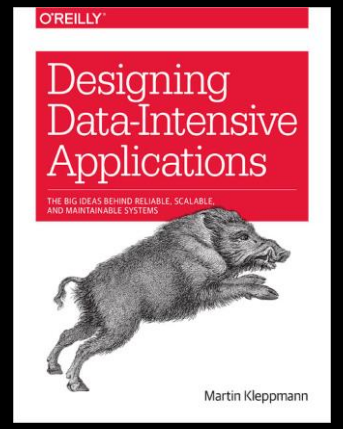
[Zoom]

[Course]

[DDIA (O'Reilly)]

[Distributed System 3rd edition]





Calendar: https://docs.google.com/spreadsheets/d/1RsbGpq1cwNSmYn5hcmT8Hv5O4qssl2HXsTcG82RHVQk/edit?usp=sharing

(Internal) Teams: g078pwd

(Public) Discord

(Public) WeChat: add mossaka or Lin1991Wen

Notion: https://www.notion.so/invite/cd6df70a94e7f67f6d21f4c509783d3c9cfd0e69

YouTube: https://www.youtube.com/playlist?list=PL1voNxn5MODMJxAZVvgFHZ0jZ-fuSut68

Company Privacy

Microsoft

Topic Covered

- Fault Tolerance
- Replication
- **Questions in Replication**
- **VMware Fault-Tolerance**

Fault Tolerance

- Fault tolerance is the property that enables a system to continue operating properly in the event of the failure of some of its components.
- Why fault tolerance:
  - High-Availability
- How to achieve Fault Tolerance
  - Replication
  - Redundancy
    - Load Balancing (Server)
    - RAID (Storage)

Replication

- Failures that replication can handle:
  - Fail stop failures of a single replica
    - Power off / Disk out of space and stops
- Failures that replication CANNOT handle:
  - Hardware defects
  - Software bugs or Human configuration errors
- Replication Approaches:
  - State transfer
    - Memory & CPU & I/O devices
  - Replicated state machine
    - Operations (Deterministic & Non-deterministic)

	State Transfer	Replicated sate machine
Bandwidth	High	Low
Client Operation	Only primary	Both primary and backup
Implementation	Simpler	Complicated

Questions in Replication

- What state to replicate?
- Does primary have to wait for backup?
- When to cut over to backup?
- How to bring a replacement backup up to speed?

VMware Fault-Tolerance: Basic Design

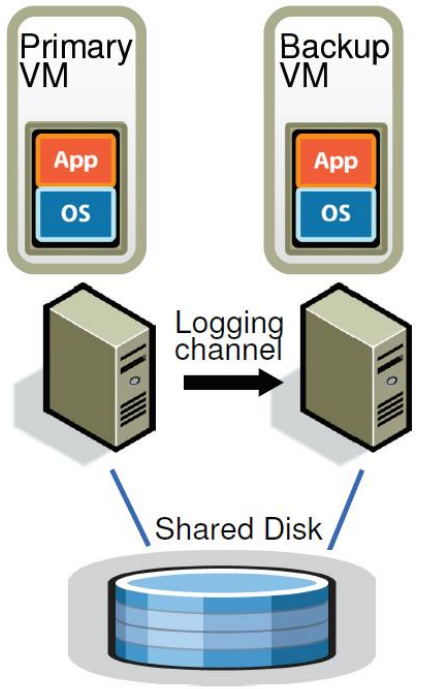


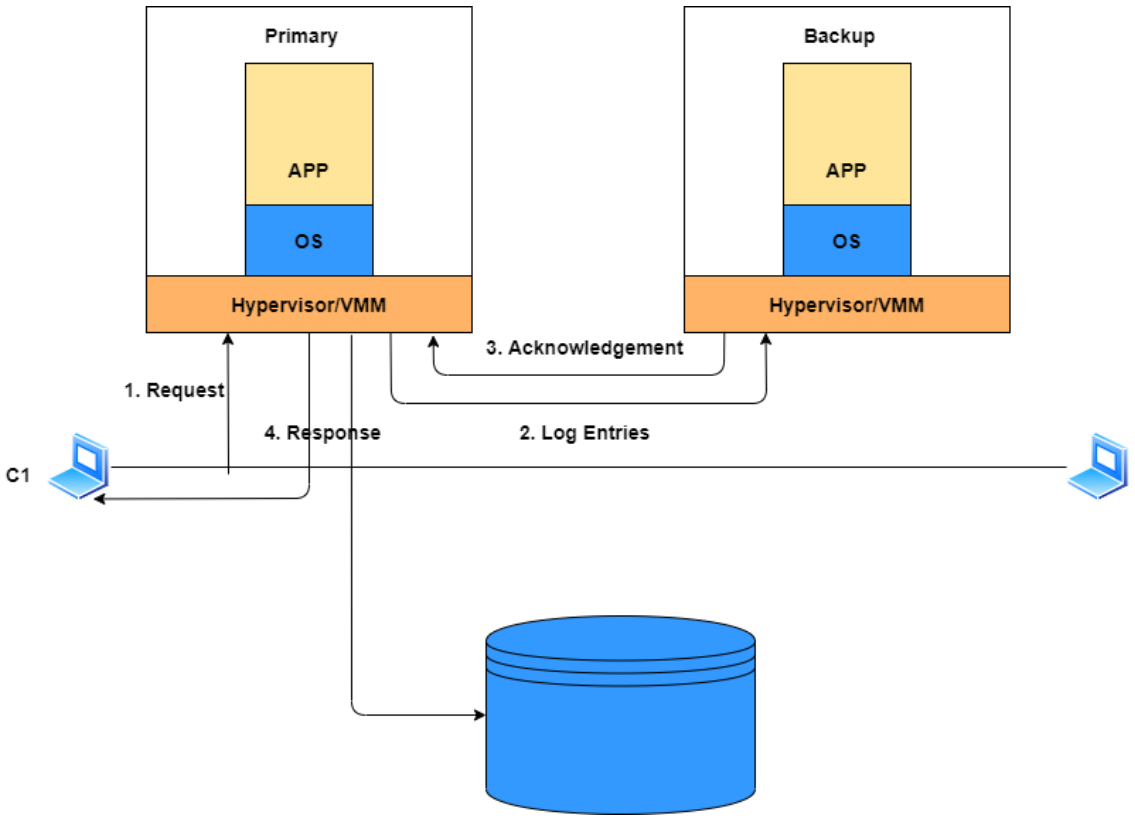
Figure 1: Basic FT Configuration.

- Based on Deterministic-Replay
- Only supports Uni-Processor VMs
- Only the primary replica advertises its presence on the network.
- Only the primary replica produces actual outputs that are returned to clients.

VMware Fault-Tolerance: FT Protocol

- Output Requirement:
  - If the backup VM ever takes over after a failure of the primary, the backup VM will continue executing in a way that is entirely consistent with all outputs that the primary VM has sent to the external world.
- Output Rule:
  - The primary VM may not send an output to the external world, until the backup VM has received and acknowledged the log entry associated with the operation producing the output.

VMware Fault-Tolerance: Output Rule



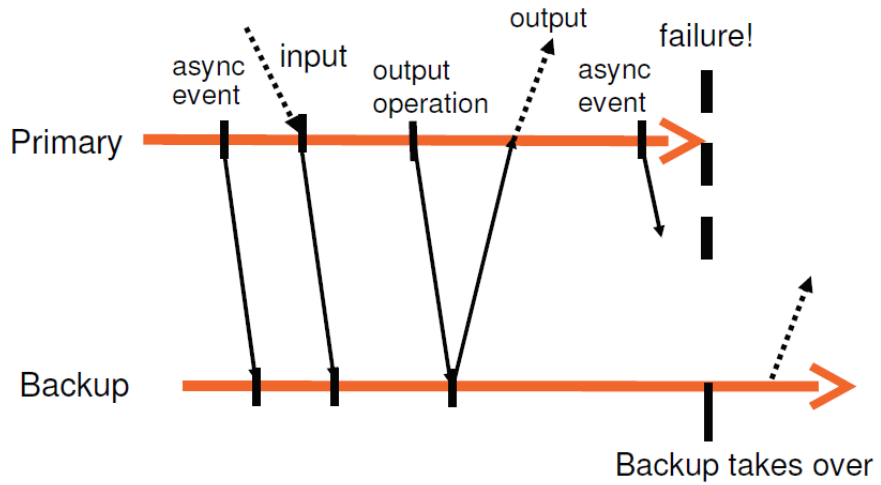
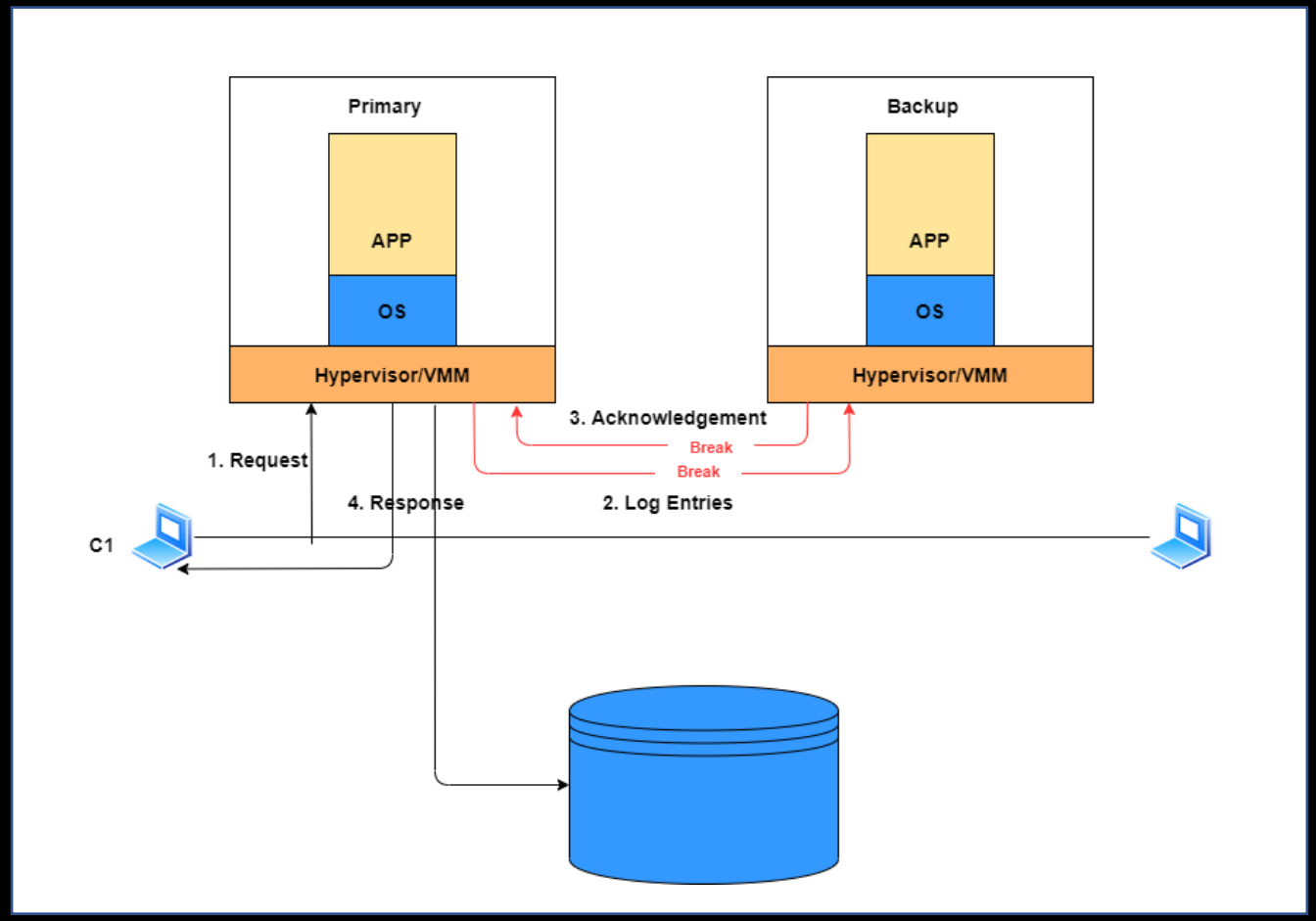


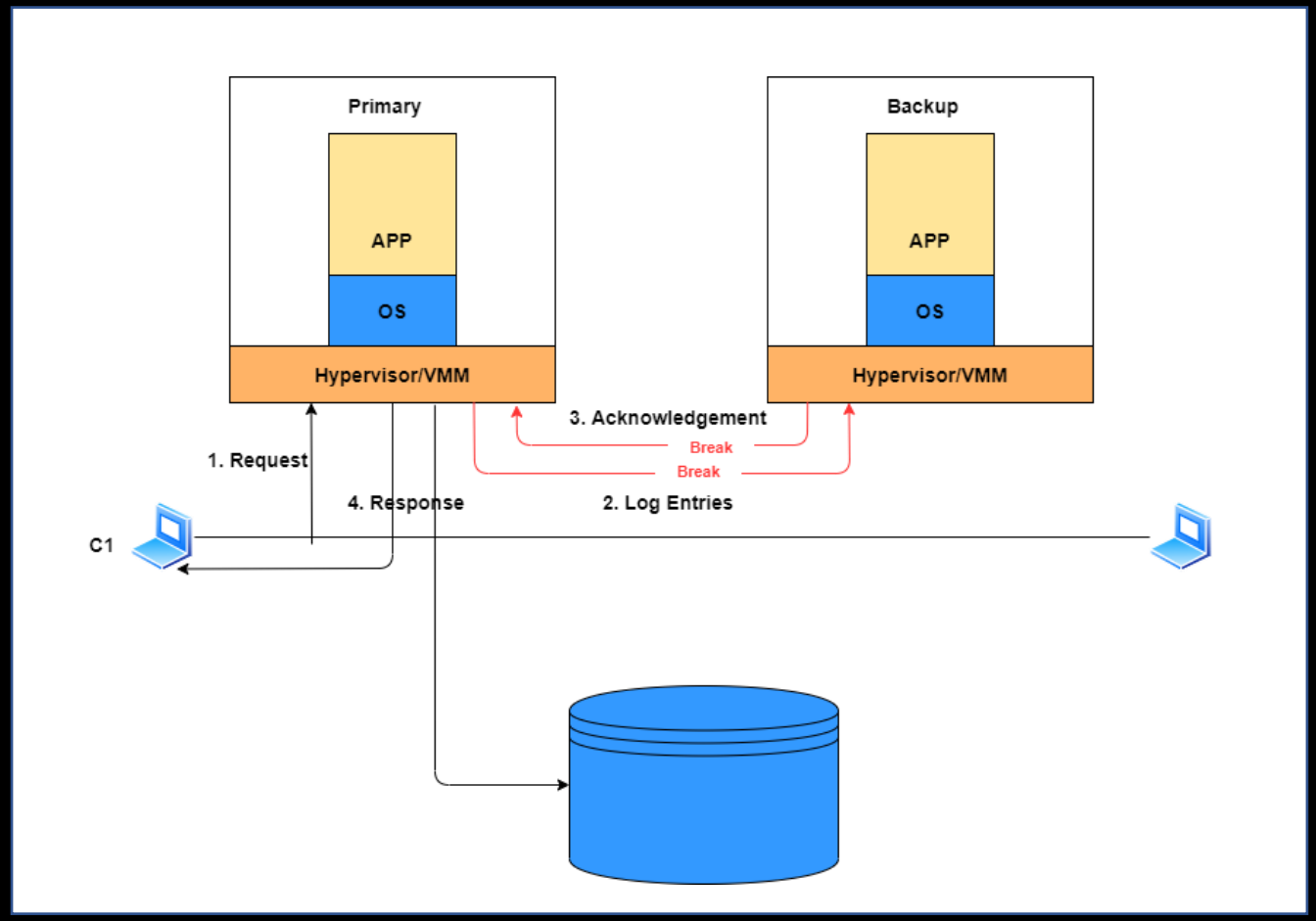
Figure 2: FT Protocol.

VMware Fault-Tolerance: Cut-Over



- Failure detection
  - UDP heartbeating between servers that are running fault-tolerant VMs.
  - Regular timer interrupts: the logging traffic should be regular and never stop for a functioning guest OS.

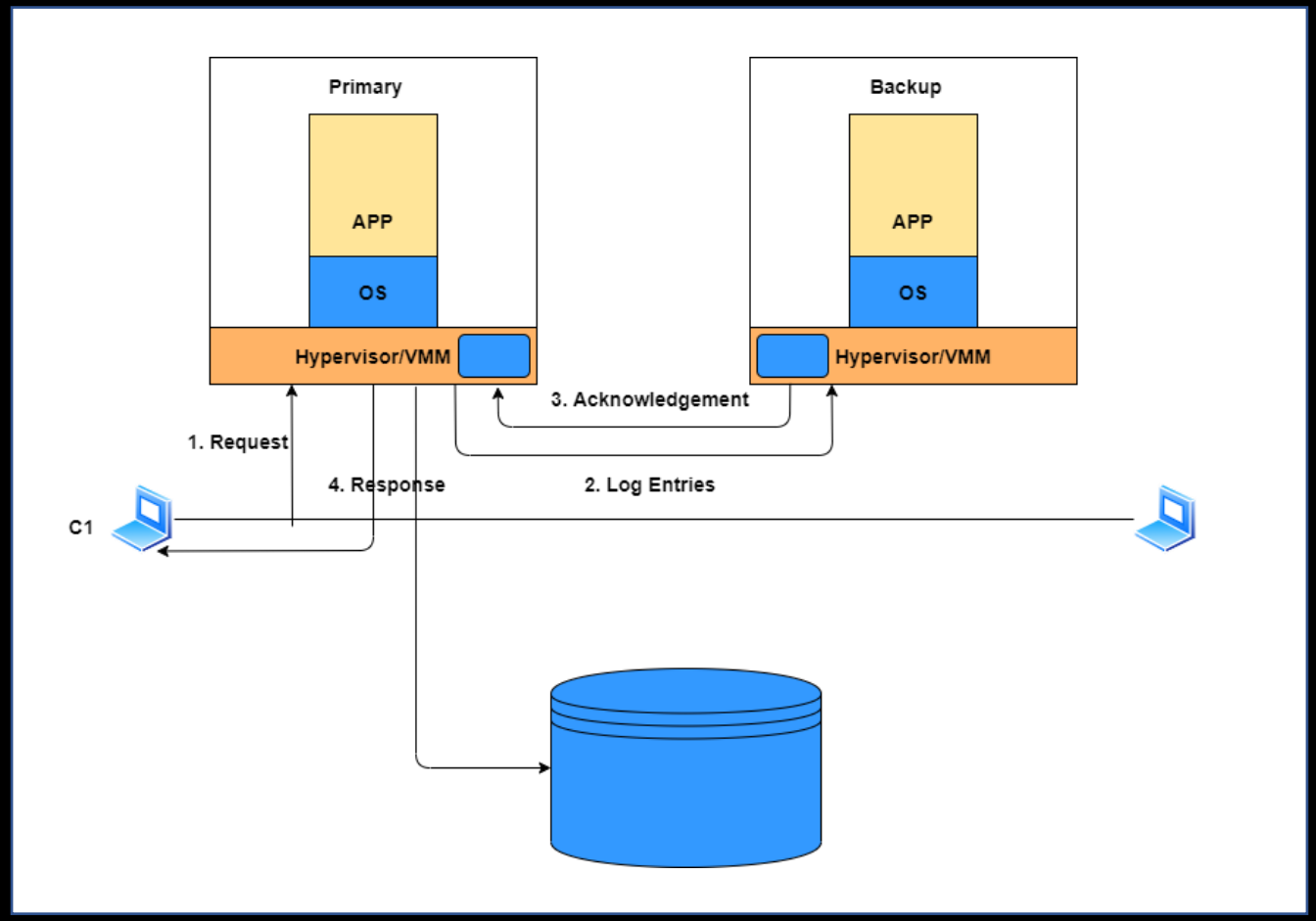
VMware Fault-Tolerance: Split-Brain



- Test-and-Set Service
  - If the primary OR backup thinks the other server is dead, and thus that it should take over by itself, it first sends a test-and-set operation to the disk server.

```
test-and-set() {
  acquire_lock()
  if flag == true:
    release_lock()
    return false
  else:
    flag = true
    release_lock()
    return true
```

VMware Fault-Tolerance: Speed up Backup



- Primary Log Buffer
  - The contents of the primary's log buffer are flushed out to the logging channel as soon as possible.
  - If the primary VM encounters a full log buffer when it needs to write a log entry, it must stop execution until log entries can be flushed out.

VMware Fault-Tolerance: Disk IO & Network IO

- Forcing potential racing disk operations to execute sequentially in the same way on the primary and backup.
- Page protection
  - Interrupt the VM until disk operation completes.
- Bounce buffer
  - Read into bunce buffer
  - Copy from bunce buffer to memory

- Disable the asynchronous network optimization
- Reducing delay for transmitted packets
  - Sending and receiving log entries and acknowledgements can all be done without any thread context switch.
- “Bounce Buffer”
  - NIC will copy packets to memory (DMA)
  - Optimization: Copy to private memory of hypervisor and interrupt VM (49:00 – 52:00 in Lec4)

Questions?

Freely accessible resources

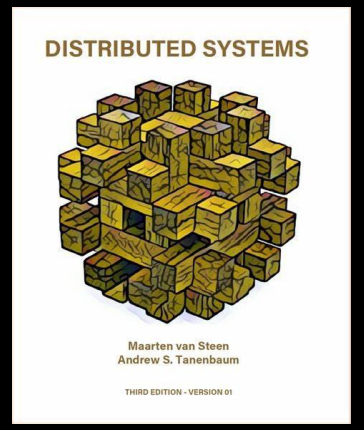
[Code]

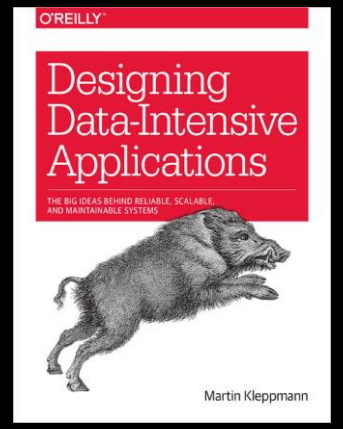
[Zoom]

[Course]

[DDIA (O'Reilly)]

[Distributed System 3rd edition]





Calendar: https://docs.google.com/spreadsheets/d/1RsbGpq1cwNSmYn5hcmT8Hv5O4qssl2HXsTcG82RHVQk/edit?usp=sharing

(Internal) Teams: g078pwd

(Public) Discord

(Public) WeChat: add mossaka or Lin1991Wen

Notion: https://www.notion.so/invite/cd6df70a94e7f67f6d21f4c509783d3c9cfd0e69

YouTube: https://www.youtube.com/playlist?list=PL1voNxn5MODMJxAZVvgFHZ0jZ-fuSut68