EECS 591
Distributed Systems

Manos Kapritsos
Winter 2018

Slides by: Lorenzo Alvisi
STATE MACHINE REPLICATION
MODELING FAULTS

- Mean Time To Failure/Mean Time To Recover
  - used mostly for disks
  - of questionable value in expressing reliability

- Threshold: \( f \) out of \( n \)
  - makes condition for correct operation explicit
  - measures fault-tolerance of the architecture, not of individual components

- Enumerate failure scenarios
A HIERARCHY OF FAILURE MODELS

- Fail-stop
- Crash
- Send omission
- Receive omission
- General omission
- Arbitrary (Byzantine) failures

○ = benign failures
A hierarchy of failure models
FAULT TOLERANCE: THE PROBLEM

Clients

Solution: replicate the server

Server
Replication in time

- When a server fails, restart it or replace it
- Failures are detected, not masked
- Lower maintenance, lower availability
- Tolerates only benign failures
Replication in space

- Run multiple copies of a server (replicas)
- Vote on replica output
- Failures are masked
- High availability and can tolerate arbitrary failures
  - but at high cost
THE ENEMY: NON-DETERMINISM

An event is non-deterministic if its output is not uniquely determined by its input.

The problem with non-determinism:

- Replication in time: must reproduce the original outcome of all non-deterministic events.
- Replication in space: each replica must handle non-deterministic events identically.
The solution: state machines

Design the server as a deterministic state machine
The solution: state machines

State machine example: a switch
**State Machine Replication**

**Ingredients:** a server

1. Make server deterministic (state machine)
2. Replicate server
3. Ensure that all replicas go through the same sequence of state transitions
4. Vote on replica outputs
STATE MACHINE REPLICA

Ingredients: a server

1. Make server deterministic (state machine)
2. Replicate server
3. Ensure that all replicas go through the same sequence of state transitions
4. Vote on replica outputs

All state machines receive all commands in the same order
**State Machine Replication**

Ingredients: a server

1. Make server deterministic (state machine)
2. Replicate server
3. Ensure that all replicas go through the same sequence of state transitions
4. Vote on replica outputs
**State Machine Replication**

**Ingredients: a server**

1. Make server deterministic (state machine)
2. Replicate server
3. Ensure that all replicas go through the same sequence of state transitions
4. Vote on replica outputs
1. Evil Lorenzo Speaks French
2. And was born in Corsica
3. Went to Dartmouth instead of Cornell
4. Rides a Ducati instead of a Moto Guzzi
5. Still listens to opera, but doesn't care for Puccini

Evil Lorenzo thinks that 2f+1 is good enough

When in trouble, cheat!

Voter and client share fate!
PREPARING FOR THE “RESEARCH” PART OF THE COURSE

I will finalize the list of papers by tonight

You each pick one to present (email me 4 preferences by Wednesday night)

- If you have a teammate, send me a joined email
- ~30-35 minutes presentation
- Send me the slides two days before your presentation
The research project

Sample topics:

Concrete
- Combining Fast Paxos and Flexible Paxos to reduce latency in a geo-replicated storage system
- Proving the correctness of BitCoin

Motivational
- Why the world needs real-time proofs of distributed systems
- Supporting the equivalent instruction hypothesis
- All the things you can do with Flexible Paxos

Survey
- Applying Byzantine Fault Tolerance to blockchains: theory and practice
Primary-Backup
The model

Failure model: crash

Network model: synchrony

- Reliable, FIFO channels
- All messages are delivered within $\delta$ time

Tolerates $f$ crash failures
THE IDEA

- Clients communicate with a single replica (primary)

- Primary:
  - sequences and processes clients’ requests
  - updates other replicas (backups)

- Backups use timeouts to detect failure of primary

- On primary failure, a backup becomes the new primary
A SIMPLE PRIMARY-BACKUP PROTOCOL

$(f = 1)$

Active replication: sync = client request(s)
Passive replication: sync = state update
A simple primary-backup protocol $(f = 1)$

Active replication: sync = client request(s)
Passive replication: sync = state update
Weakening the model

Failure model: crash

Network model: synchrony
  - Unreliable, FIFO channels
  - Channels may drop messages
  - All messages are delivered within $\delta$ time
    - (looks paradoxical)

Tolerates $f$ crash failures
A slightly different primary-backup protocol $(f = 1)$