ATTRIBUTION

• These slides are released under an Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) Creative Commons license

• These slides incorporate material from:
  • Tanenbaum and Van Steen, Dist. Systems: Principles and Paradigms
  • Kyle Jamieson, Princeton University (also under a CC BY-NC-SA 3.0 Creative Commons license)
ANNOUNCEMENTS
RECALL OUR 2PC COMMIT PROBLEM

1. C → TC: “go!”

2. TC → A, B: “prepare!”

3. A, B → P: “yes” or “no”

4. TC → A, B: “commit!” or “abort!”
RECALL OUR 2PC COMMIT PROBLEM

- Who acts as TC?
- Which server(s) own the account of A? B?
- Who takes over if TC fails? What about if A or B fail?
DOING FAILOVER "CORRECTLY" ISN’T EASY

Which node takes over as backup?
DOING FAILOVER "CORRECTLY" ISN’T EASY

Okay, so specify some ordering
(manually, using some identifier)
DOING FAILOVER “CORRECTLY” ISN’T EASY

But who determines if 1 failed?
DOING FAILOVER "CORRECTLY" ISN’T EASY

Transaction Coordinator TC

Easy, right? Just ping and timeout!
Doing failover "correctly" isn’t easy.

Is the server or the network actually dead/slow?
WHAT CAN GO WRONG?

Two nodes think they are TC:
“Split brain” scenario
WHAT CAN GO WRONG?

Transaction Coordinator TC

Two nodes think they are TC: “Split brain” scenario
WHAT CAN GO WRONG?

Safety invariant:
Only 1 node is TC at any single time

Another problem:
A and B need to know (and agree upon) who the TC is...
CONSENSUS (NOUN):

1. A general agreement about something
2. An idea or opinion that is shared by all the people in a group
CONSENSUS

- Given a set of processors, each with an initial value:
  - **Termination**: All non-faulty processes eventually decide on a value
  - **Agreement**: All processes that decide do so on the same value
  - **Validity**: The value that has been decided must have proposed by some process
Group of servers attempting:

• Make sure all servers in group receive the same updates in the same order as each other

• Maintain own lists (views) on who is a current member of the group, and update lists when somebody leaves/fails

• Elect a leader in group, and inform everybody

• Ensure mutually exclusive (one process at a time only) access to a critical resource like a file
STEP ONE: DEFINE YOUR SYSTEM MODEL

- **Network model:**
  - Synchronous (time-bounded delay) or asynchronous (arbitrary delay)
  - Reliable or unreliable communication
  - Unicast or multicast communication

- **Node failures:**
  - Fail-stop (correct/dead) or Byzantine (arbitrary)
STEP ONE: DEFINE YOUR SYSTEM MODEL

• Network model:
  • Synchronous (time-bounded delay) or asynchronous (arbitrary delay)
  • Reliable or unreliable communication
  • Unicast or multicast communication

• Node failures:
  • Fail-stop (correct/dead) or Byzantine (arbitrary)
IMPORTANCE OF THIS STUFF

• **Werner Vogels, Amazon CTO**

• Job openings in my group

• What kind of things am I looking for in you?

• “You know your distributed systems theory: You know about logical time, snapshots, stability, message ordering, but also acid and multi-level transactions. You have heard about the FLP impossibility argument. You know why failure detectors can solve it (but you do not have to remember which one diamond-w was). You have at least once tried to understand Paxos by reading the original paper.”
PAXOS

- **Safety**
  - Only a single value is chosen
  - Only a proposed value can be chosen
  - Only chosen values are learned by processes

- **Liveness**
  - Some proposed value eventually chosen if fewer than half of processes fail
  - If value is chosen, a process eventually learns it
Leader election to decide transaction coordinator
USING PAXOS IN SYSTEM

New leader election protocol

Still have split-brain scenario!
• Tells mythical story of Greek island of Paxos with “legislators” and “current law” passed through parliamentary voting protocol


• Lamport won the Turing Award in 2013
As Paxos prospered, legislators became very busy.

Parliament could no longer handle all details of government, so a bureaucracy was established.

Instead of passing a decree to declare whether each lot of cheese was fit for sale, Parliament passed a decree appointing a cheese inspector to make those decisions.

Cheese inspector $\approx$ leader using quorum-based voting protocol
Parliament passed a decree making Δικστρα the first cheese inspector. After some months, merchants complained that Δικστρα was too strict and was rejecting perfectly good cheese.

Parliament then replaced him by passing the decree 1375: Γωυδα is the new cheese inspector.

But Δικστρα did not pay close attention to what Parliament did, so he did not learn of this decree right away.

There was a period of confusion in the cheese market when both Δικστρα and Γωυδα were inspecting cheese and making conflicting decisions.

Split-brain!
To prevent such confusion, the Paxons had to guarantee that a position could be held by at most one bureaucrat at any time.

To do this, a president included as part of each decree the time and date when it was proposed.

A decree making Δίκστρα the cheese inspector might read

2716: 8:30 15 Jan 72 – Δίκστρα is cheese inspector for 3 months.

Leader gets a lease!
A bureaucrat needed to tell time to determine if he currently held a post. Mechanical clocks were unknown on Paxos, but Paxons could tell time accurately to within 15 minutes by the position of the sun or the stars.

If 

If Δικστρα’s term began at 8:30, he would not start inspecting cheese until his celestial observations indicated that it was 8:45.

Handle clock skew:
Lease doesn’t end until expiry + max skew
New leader election protocol

Solution
If L isn’t part of majority electing $L_{new}$
$L_{new}$ waits until L’s lease expires before accepting new ops
UC San Diego