Paxos: Fun Facts

- Why is the algorithm called Paxos?
- Leslie Lamport described the algorithm as the solution to a problem of the parliament on a fictitious Greek island called Paxos
- Many readers were so distracted by the description of the activities of the legislators, they did not understand the meaning and purpose of the algorithm. The paper was rejected
- Leslie Lamport refused to rewrite the paper. He later wrote that he *"was quite annoyed at how humorless everyone working in the field seemed to be"*
- After a few years, some people started to understand the importance of the algorithm
- After eight years, Leslie Lamport submitted the paper again, basically unaltered. It got accepted!

Quorum

Paxos used Majority sets: Can this be generalized?

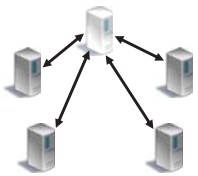
Yes: It`s called Quorum

- In law, a quorum is a the minimum number of members of a deliberative body necessary to conduct the business of the group.
- In our case: substitute "members of a deliberative body" with "any subset of servers of a distributed system"

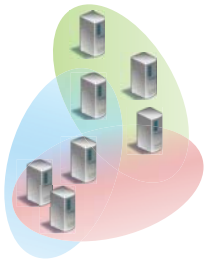
A Quorum does not automatically need to be a majority.

What else can you imagine? What are reasonable objectives?

Quorum: Primary Copy vs. Majority



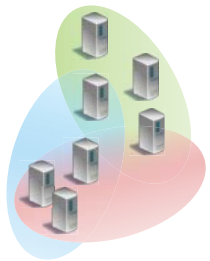
or



?

Quorum: Primary Copy vs. Majority

	Singleton	Majority
How many server failures can be tolerated?	0	$< n/2$
How many servers need to be contacted?	1	$> n/2$
What's the load of the busiest server?	100%	≈ 50%



Definition: Quorum System

Definition

Quorum System

Let $P = \{P_1, \ldots, P_n\}$ be a set of servers.
A quorum system $\mathrm{Q} \subset 2^P$ is a set of subsets of P such that every two subsets intersect. Each $\mathrm{Q} \in \mathrm{Q}$ is called a quorum.

Definition

Minimal Quorum System

A quorum system Q is called minimal if $\forall\, Q, Q' \in \mathrm{Q}: Q \not\subset Q'$

Definition: Load

Definition

Access Strategy

An access strategy W is a random variable on a quorum system Q, i.e. $\sum_{Q \in \mathrm{Q}} P_W(Q) = 1$

Definition

Load

The load induced by access strategy W on a server P_i is:

$$l_W(i) = \sum_{Q \in \mathrm{Q};\, P_i \in Q} P_W(Q)$$

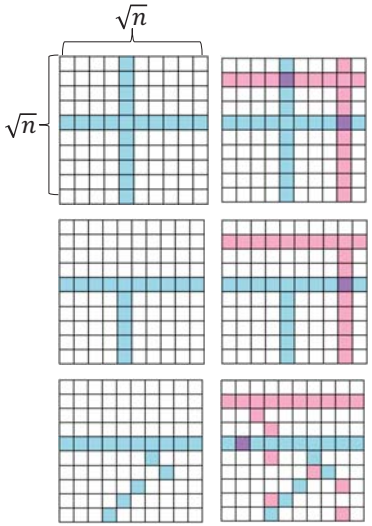
The load induced by W on a quorum system Q is the maximal load induced by W on any server in Q.

$$L_W(\mathrm{Q}) = \max_{P_i \in P} l_W(i)$$

The system load of Q is

$$L(\mathrm{Q}) = \min_W L_W(\mathrm{Q})$$

Quorum: Grid



- Work: $2\sqrt{n} - 1$
- Load: $\frac{2\sqrt{n}-1}{n}$

Definitions: Fault Tolerance

Definition

Resilience

The resilience $R(\mathrm{Q})$ of a quorum system is the largest f such that for all sets $F \subset P$ of cardinality f, there is at least one quorum $Q \in \mathrm{Q}$ which has no common element with F

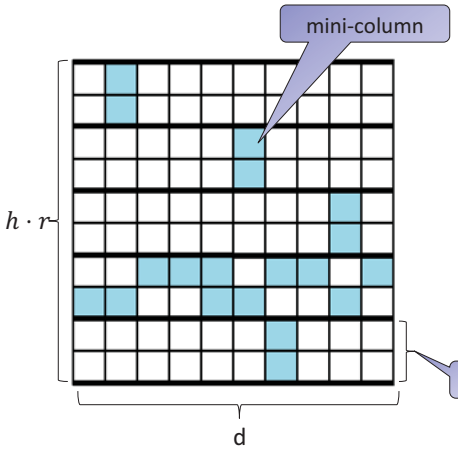
Definition

Failure Probability

The failure probability of a quorum system Q is the probability that at least one server of every quorum fails, i.e.

$$F_p(\mathrm{Q}) = \Pr[\forall Q \in \mathrm{Q}: \exists P_i \in Q \; such \; that \; FAIL(i)]$$

Quorum: B-Grid



- Suppose $n = dhr$ and arrange the elements in a grid with d columns and $h \cdot r$ rows. Call every group of r rows a band and call r elements in a column restricted to a band a mini-column. A quorum consists of one mini-column in every band and one element from each mini-column of one band; thus, every quorum has $d + hr - 1$ elements
- Resilience?

Quorum Systems: Overview

	Singleton	Majority	Grid	B-Grid
Resilience	0	$< n/2$	$\sqrt{n} - 1$	$\theta(\sqrt{n})$
Work	1	$> n/2$	$\theta(\sqrt{n})$	$\theta(\sqrt{n})$
Load	1	$1/2$	$\theta(1/\sqrt{n})$	$\theta(1/\sqrt{n})$
Failure Prob.*	p	$\to 0$	$\to 1$	$\to 0$

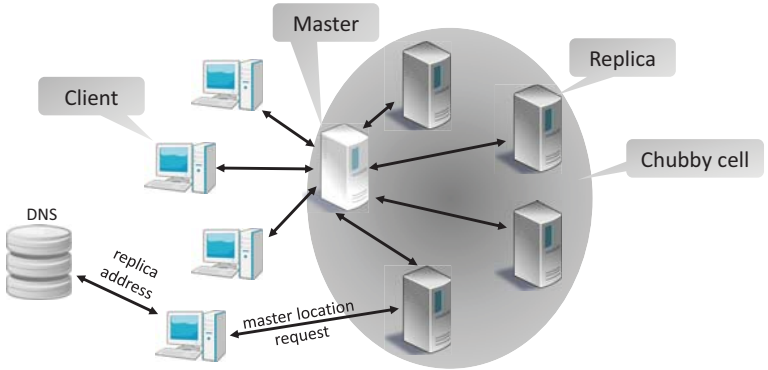
*Assuming p constant but significantly less than ½.

Chubby

- Chubby is a coarse-grained distributed lock service
  - Coarse-grained: Locks are held for hours or even days
- Chubby allows clients to synchronize activities
  - E.g., synchronize access through a leader in a distributed system
  - The leader is elected using Chubby: The node that gets the lock for this service becomes the leader!
- Design goals are high availability and reliability
  - High performance is not a major issue
- Chubby is used in many tools, services etc. at Google
  - Google File System (GFS)
  - BigTable (distributed database)

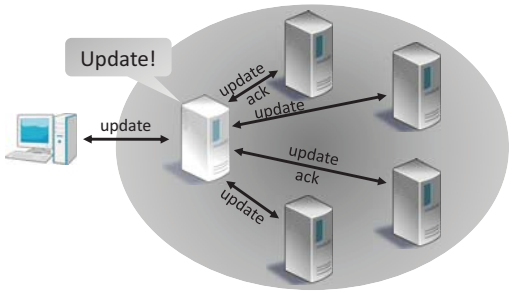
Chubby: System Structure

- A Chubby cell typically consists of 5 servers
  - One server is the master, the others are replicas
  - The clients only communicate with the master
  - Clients find the master by sending master location requests to some replicas listed in the DNS



Chubby: System Structure

- The master handles all read accesses
- The master also handles writes
  - Copies of the updates are sent to the replicas
  - Majority of replicas must acknowledge receipt of update before master writes its own value and updates the official database



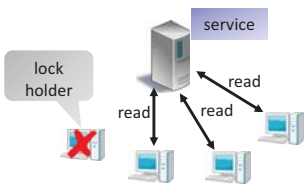
Chubby: Master Election

- The master remains the master for the duration of the master lease
  - Before the lease expires, the master can renew it (and remain the master)
  - It is guaranteed that no new master is elected before the lease expires
  - However, a new master is elected as soon as the lease expires
  - This ensures that the system does not freeze (for a long time) if the master crashed
- How do the servers in the Chubby cell agree on a master?
- They run (a variant of) the Paxos algorithm!

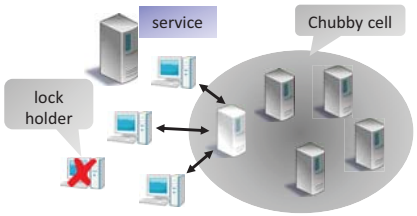
Chubby: Locks

- Locks are advisory (not mandatory)
  - As usual, locks are mutually exclusive
  - However, data can be *read* without the lock!
  - Advisory locks are more efficient than mandatory locks (where any access requires the lock): Most accesses are reads! If a mandatory lock is used and the lock holder crashes, then all reads are stalled until the situation is resolved
  - Write permission to a resource is required to obtain a lock

Advisory:

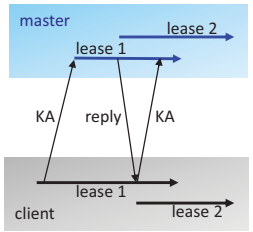


Mandatory:



Chubby: Sessions

- What happens if the lock holder crashes?
- Client initially contacts master to establish a session
  - Session: Relationship between Chubby cell and Chubby client
- Each session has an associated lease
  - The master can extend the lease, but it may not revoke the lease
  - Longer lease times if the load is high
- Periodic KeepAlive (KA) handshake to maintain relationship
  - The master does not respond until the client's previous lease is close to expiring
  - Then it responds with the duration of the new lease
  - The client reacts immediately and issues the next KA
- Ending a session
  - The client terminates the session explicitly
  - or the lease expires

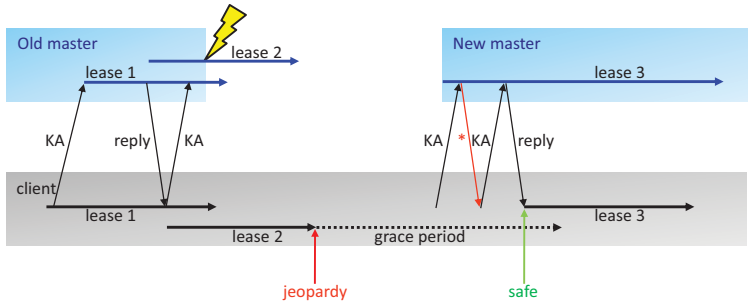


Chubby: Lease Timeout

- The client maintains a local lease timeout (Time when lease expires)
  - The client knows (roughly) when it has to hear from the master again
- If the local lease expires, the session is in jeopardy
- As soon as a session is in jeopardy, the grace period (45s by default) starts
  - If there is a successful KeepAlive exchange before the end of the grace period, the session is saved!
  - Otherwise, the session expired
- This might happen if the master crashed…

Chubby: Master Failure

- The grace period can save sessions



- The client finds the new master using a master location request
- Its first KA to the new master is denied (*) because the new master has a new epoch number (sometimes called view number)
- The next KA succeeds with the new number

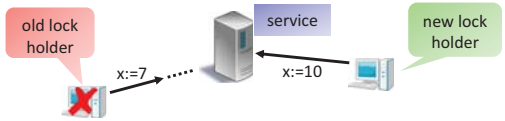
Chubby: Master Failure

- A master failure is detected once the master lease expires
- A new master is elected, which tries to resume exactly where the old master left off
  - Read data that the former master wrote to disk (this data is also replicated)
  - Obtain state from clients
- Actions of the new master
  1. It picks a new epoch number
  - It only replies to master location requests
  2. It rebuilds the data structures of the old master
  - Now it also accepts KeepAlives
  3. It inform all clients about failure → Clients flush cache
  - All operations can proceed

(We omit caching in this lecture!)

Chubby: Locks Reloaded

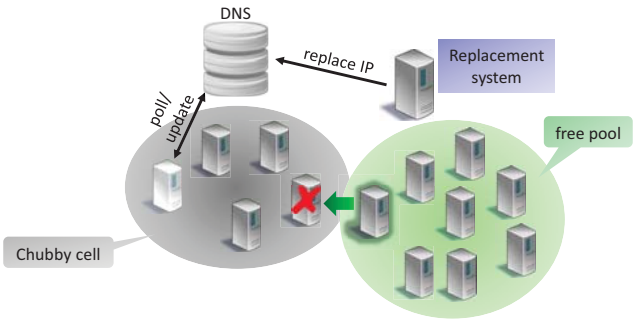
- What if a lock holder crashes and its (write) request is still in transit?
  - This write may undo an operation of the next lock holder!



- Heuristic I: Sequencer
  - Add a sequencer (which describes the state of the lock) to the access requests
  - The sequencer is a bit string that contains the name of lock, the mode (exclusive/shared), and the lock generation number
  - The client passes the sequencer to server. The server is expected to check if the sequencer is still valid and has the appropriate mode
- Heuristic II: Delay access
  - If a lock holder crashed, Chubby blocks the lock for a period called the lock delay

Chubby: Replica Replacement

- What happens when a replica crashes?
  - If it does not recover for a few hours, a replacement system selects a fresh machine from a pool of machines
  - Subsequently, the DNS tables are updated by replacing the IP address of the failed replica with the new one
  - The master polls the DNS periodically and eventually notices the change

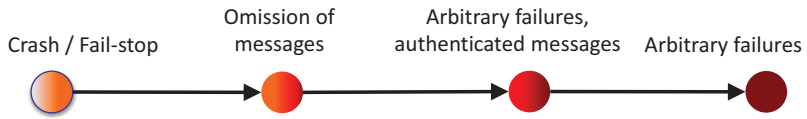


Chubby: Performance

- According to Chubby…
  - Chubby performs quite well
- 90K+ clients can communicate with a single Chubby master (2 CPUs)
- System increases lease times from 12s up to 60s under heavy load
- Clients cache virtually everything
- Only little state has to be stored
  - All data is held in RAM (but also persistently stored on disk)

Practical Byzantine Fault-Tolerance

- So far, we have only looked at systems that deal with simple (crash) failures
- We know that there are other kind of failures:



Practical Byzantine Fault-Tolerance

- Is it reasonable to consider Byzantine behavior in practical systems?
- There are several reasons why clients/servers may behave “arbitrarily”
  - Malfunctioning hardware
  - Buggy software
  - Malicious attacks
- Can we have a practical and efficient system that tolerates Byzantine behavior…?
  - We again need to solve consensus…

PBFT

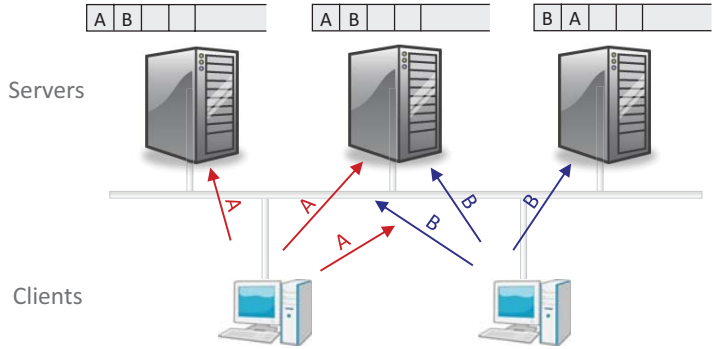
- We are now going to study the Practical Byzantine Fault-Tolerant (PBFT) system
- The system consists of clients that read/write data stored at *n* servers

- Goal
  - The system can be used to implement any deterministic replicated service with a *state* and some *operations*
  - Provide reliability and availability

- Model
  - Communication is asynchronous, but message delays are bounded
  - Messages may be lost, duplicated or may arrive out of order
  - Messages can be authenticated using digital signatures (in order to prevent spoofing, replay, impersonation)
  - At most $f < n/3$ of the servers are Byzantine

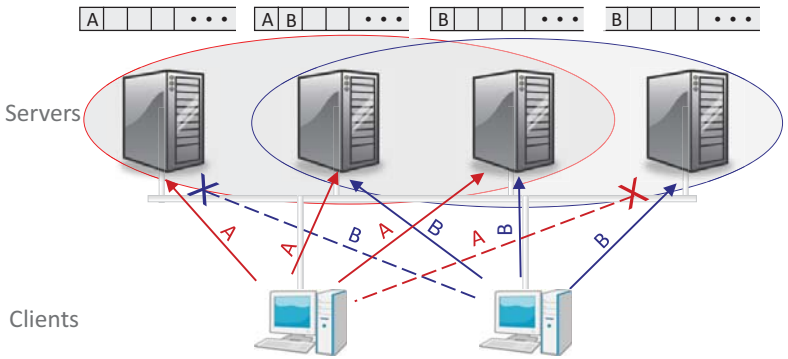
PBFT: Order of Operations

- State replication (repetition): If all servers start in the same state, all operations are deterministic, and all operations are executed in the same order, then all servers remain in the same state!
- Variable message delays may be a problem:



PBFT: Order of Operations

- If messages are lost, some servers may not receive all updates…



PBFT: Basic Idea

- Such problems can be solved by using a coordinator
- One server is the primary
  - The clients send signed commands to the primary
  - The primary assigns sequence numbers to the commands
  - These sequence numbers impose an order on the commands
- The other servers are backups
  - The primary forwards commands to the other servers
  - Information about commands is replicated at a quorum of backups

PBFT is not decentralized like Paxos!

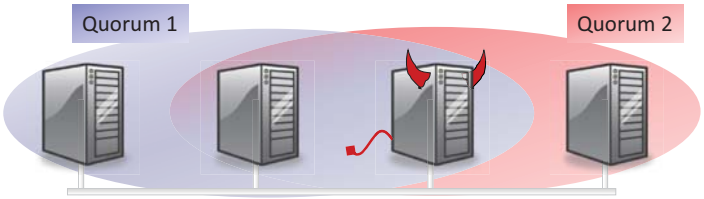
Quorum…?

- Note that we assume in the following that there are *exactly* $n = 3f+1$ servers!

Byzantine Quorums

Now, a quorum is any subset of the servers of size at least $2f+1$

- The intersection between any two quorums contains at least one correct (not Byzantine) server

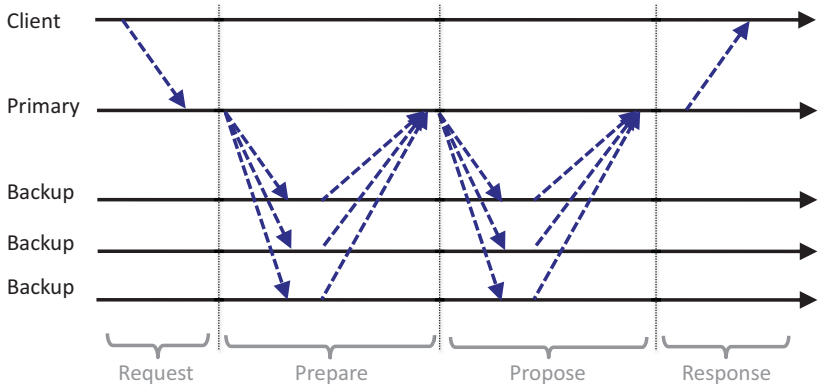


PBFT: Main Algorithm

- PBFT takes 5 rounds of communication
- In the first round, the client sends the command op to the primary
- The following three rounds are
  - Pre-prepare
  - Prepare
  - Propose
- In the fifth round, the client receives replies from the servers
  - If $f+1$ (authenticated) replies are the same, the result is accepted
  - Since there are only f Byzantine servers, at least one correct server supports the result
- The algorithm is somewhat similar to Paxos…

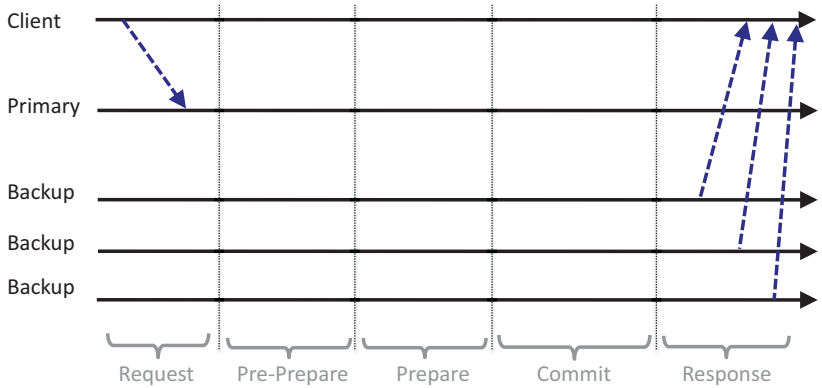
PBFT: Paxos

- In Paxos, there is only a prepare and a propose phase
- The primary is the node issuing the proposal
- In the response phase, the clients learn the final result



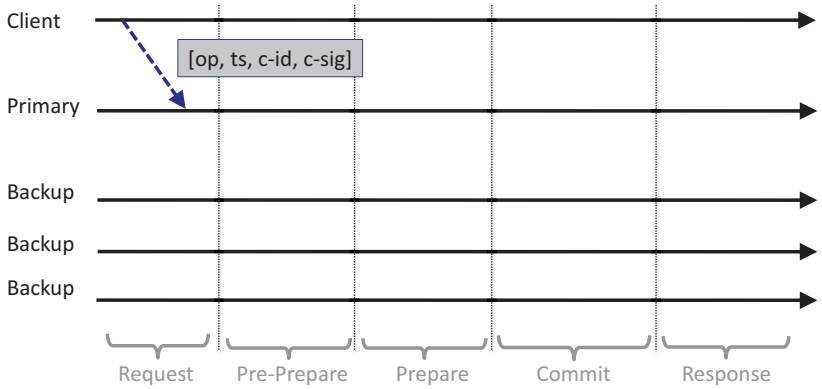
PBFT: Algorithm

- PBFT takes 5 rounds of communication
- The main parts are the three rounds pre-prepare, prepare, and commit



PBFT: Request Phase

- In the first round, the client sends the command op to the primary
- It also sends a timestamp ts, a client identifier c-id and a signature c-sig

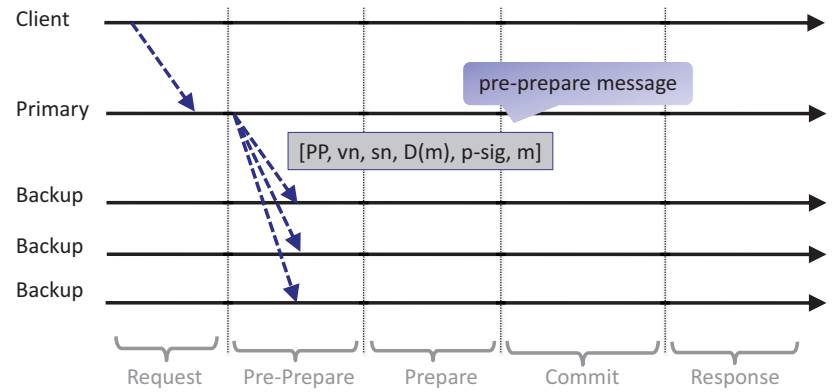


PBFT: Request Phase

- Why adding a timestamp?
  - The timestamp ensures that a command is recorded/executed exactly once
- Why adding a signature?
  - It is not possible for another client (or a Byzantine server) to issue commands that are accepted as commands from client c
  - The system also performs access control: If a client c is allowed to write a variable x but c' is not, c' cannot issue a write command by pretending to be client c!

PBFT: Pre-Prepare Phase

- In the second round, the primary multicasts m = [op, ts, cid, c-sig] to the backups, including the view number vn, the assigned sequence number sn, the message digest D(m) of m, and its own signature p-sig



PBFT: Pre-Prepare Phase

- The sequence numbers are used to order the commands and the signature is used to verify the authenticity as before
- Why adding the message digest of the client's message?
  - The primary signs only [PP, vn, sn, D(m)]. This is more efficient!
- What is a view?
  - A view is a configuration of the system. Here we assume that the system comprises the same set of servers, one of which is the primary
  - I.e., the primary determines the view: Two views are different if a different server is the primary
  - A view number identifies a view
  - The primary in view vn is the server whose identifier is vn mod *n*
  - Ideally, all servers are (always) in the same view
  - A view change occurs if a different primary is elected

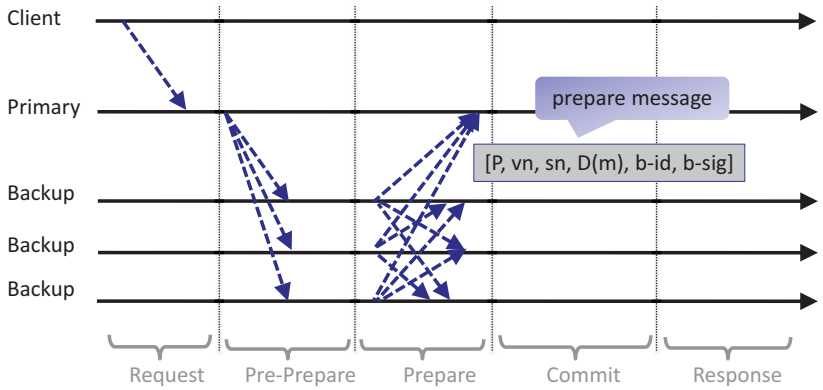
More on view changes later…

PBFT: Pre-Prepare Phase

- A backup accepts a pre-prepare message if
  - the signatures are correct
  - D(m) is the digest of m = [op, ts, cid, c-sig]
  - it is in view vn
  - It has not accepted a pre-prepare message for view number vn and sequence number sn containing a different digest
  - the sequence number is between a low water mark h and a high water mark H
  - The last condition prevents a faulty primary from exhausting the space of sequence numbers
- Each accepted pre-prepare message is stored in the local log

PBFT: Prepare Phase

- If a backup b accepts the pre-prepare message, it enters the prepare phase and multicasts [P, vn ,sn, D(m), b-id, b-sig] to all other replicas and stores this prepare message in its log

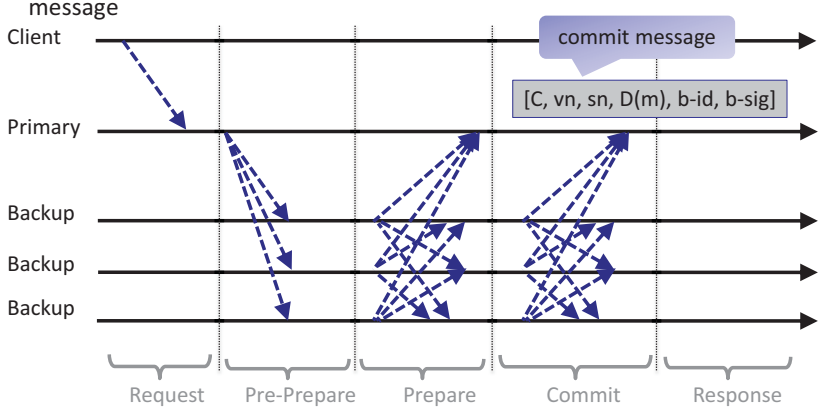


PBFT: Prepare Phase

- A replica (including the primary) accepts a prepare message if
  - the signatures are correct
  - it is in view vn
  - the sequence number is between a low water mark h and a high water mark H
- Each accepted prepare message is also stored in the local log

PBFT: Commit Phase

- If a backup b has message m, an accepted pre-prepare message, and *2f* accepted prepare messages from different replicas in its log, it multicasts [C, vn, sn, D(m), b-id, b-sig] to all other replicas and stores this commit message

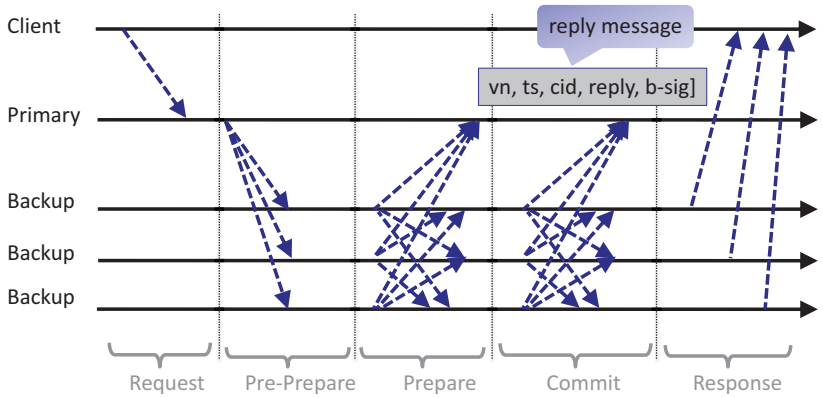


PBFT: Commit Phase

- A replica (including the primary) accepts a commit message if
  - the signatures are correct
  - it is in view vn
  - the sequence number is between a low water mark h and a high water mark H
- Each accepted commit message is also stored in the local log

PBFT: Response Phase

- If a backup b has accepted *2f+1* commit messages, it performs op ("commits") and sends a reply to the client



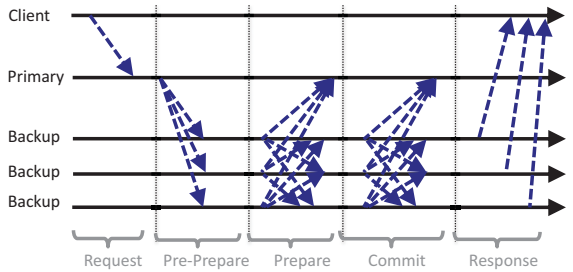
PBFT: Garbage Collection

- The servers store all messages in their log
- In order to discard messages in the log, the servers create checkpoints (snapshots of the state) every once in a while
- A checkpoint contains the 2*f*+1 signed commit messages for the committed commands in the log
- The checkpoint is multicast to all other servers
- If a server receives 2*f*+1 matching checkpoint messages, the checkpoint becomes stable and any command that preceded the commands in the checkpoint are discarded
- Note that the checkpoints are also used to set the low water mark h
  - to the sequence number of the last stable checkpoint

  and the high water mark H
  - to a "sufficiently large" value

PBFT: Correct Primary

- If the primary is correct, the algorithm works
  - All 2*f*+1 correct nodes receive pre-prepare messages and send prepare messages
  - All 2*f*+1 correct nodes receive 2*f*+1 prepare messages and send commit messages
  - All 2*f*+1 correct nodes receive 2*f*+1 commit messages, commit, and send a reply to the client
  - The client accepts the result



PBFT: No Replies

- What happens if the client does not receive replies?
  - Because the command message has been lost
  - Because the primary is Byzantine and did not forward it
- After a time-out, the client multicasts the command to all servers
  - A server that has already committed the result sends it again
  - A server that is still processing it ignores it
  - A server that has not received the pre-prepare message forwards the command to the primary
  - If the server does not receive the pre-prepare message in return after a certain time, it concludes that the primary is faulty/Byzantine

This is how a failure of the primary is detected!

PBFT: View Change

- If a server suspects that the primary is faulty
  - it stops accepting messages except checkpoint, view change and new view messages
  - it sends a view change message containing the identifier i = vn+1 mod n of the next primary and also a certificate for each command for which it accepted 2f+1 prepare messages
  - A certificate simply contains the 2*f*+1 accepted signatures

The next primary!

- When server i receives 2*f* view change messages from other servers, it broadcasts a new view message containing the signed view change
- The servers verify the signature and accept the view change!
- The new primary issues pre-prepare messages with the new view number for all commands with a correct certificate

PBFT: Ordered Commands

- Commands are totally ordered using the view numbers and the sequence numbers
- We must ensure that a certain (vn,sn) pair is always associated with a unique command m!
- If a correct server committed [m, vn, sn], then no other correct server can commit [m', vn, sn] for any m≠ m' s.t. D(m) ≠ D(m')
  - If a correct server committed, it accepted a set of 2*f*+1 authenticated commit messages
  - The intersection between two such sets contains at least *f*+1 authenticated commit messages
  - There is at least one correct server in the intersection
  - A correct server does not issue (pre-)prepare messages with the same vn and sn for different m!

PBFT: Correctness

Theorem

If a client accepts a result, no correct server commits a different result

Proof:

- A client only accepts a result if it receives *f*+1 authenticated messages with the same result
- At least one correct server must have committed this result
- As we argued on the previous slide, no other correct server can commit a different result

PBFT: Liveness

Theorem

PBFT terminates eventually

Proof:

- The primary is correct
  - As we argued before, the algorithm terminates after 5 rounds if no messages are lost
  - Message loss is handled by retransmitting after certain time-outs
  - Assuming that messages arrive eventually, the algorithm also terminates eventually

PBFT: Liveness

Theorem

PBFT terminates eventually

Proof continued:

- The primary is Byzantine
  - If the client does not accept an answer in a certain period of time, it sends its command to all servers
  - In this case, the system behaves as if the primary is correct and the algorithm terminates eventually!
- Thus, the Byzantine primary cannot delay the command indefinitely. As we saw before, if the algorithm terminates, the result is correct!
  - i.e., at least one correct server committed this result

PBFT: Evaluation

- The Andrew benchmark emulates a software development workload
- It has 5 phases:

1. Create subdirectories recursively
2. Copy a source tree
3. Examine the status of all the files in the tree without examining the data
4. Examine every byte in all the files
5. Compile and link the files

- It is used to compare 3 systems
  - BFS (PBFT) and 4 replicas and BFS-nr (PBFT without replication)
  - BFS (PBFT) and NFS-std (network file system)
- Measured normal-case behavior (i.e. no view changes) in an isolated network

PBFT: Evaluation

- Most operations in NFS V2 are not read-only (r/o)
  - E.g., *read* and *lookup* modify the time-last-accessed attribute
- A second version of PBFT has been tested in which lookups are read-only
- Normal (strict) PBFT is only 26% slower than PBFT without replication
  → Replication does not cost too much!
- Normal (strict) PBFT is only 3% slower than NFS-std, and PBFT with read-only lookups is even 2% faster!

phase	BFS		BFS-nr
	strict	r/o lookup	
1	0.55 (57%)	0.47 (34%)	0.35
2	9.24 (82%)	7.91 (56%)	5.08
3	7.24 (18%)	6.45 (6%)	6.11
4	8.77 (18%)	7.87 (6%)	7.41
5	38.68 (20%)	38.38 (19%)	32.12
total	64.48 (26%)	61.07 (20%)	51.07

Times are in seconds

phase	BFS		NFS-std
	strict	r/o lookup	
1	0.55 (-69%)	0.47 (-73%)	1.75
2	9.24 (-2%)	7.91 (-16%)	9.46
3	7.24 (35%)	6.45 (20%)	5.36
4	8.77 (32%)	7.87 (19%)	6.60
5	38.68 (-2%)	38.38 (-2%)	39.35
total	64.48 (3%)	61.07 (-2%)	62.52

PBFT: Discussion

- PBFT guarantees that the commands are totally ordered
- If a client accepts a result, it knows that at least one correct server supports this result

- Disadvantages:
- Commit not at all correct servers
  - It is possible that only one correct server commits the command
  - We know that f other correct servers have sent commit, but they may only receive $f+1$ commits and therefore do not commit themselves…
- Byzantine primary can slow down the system
  - Ignore the initial command
  - Send pre-prepare always after the other servers forwarded the command
  - No correct server will force a view change!

Beating the Lower Bounds…

- We know several crucial impossibility results and lower bounds
  - No deterministic algorithm can achieve consensus in asynchronous systems even if only one node may crash
  - Any deterministic algorithm for synchronous systems that tolerates f crash failures takes at least $f+1$ rounds

- Yet we have just seen a deterministic algorithm/system that
  - achieves consensus in asynchronous systems and that tolerates $f < n/3$ Byzantine failures
  - The algorithm only takes five rounds…?

- So, why does the algorithm work…?

Beating the Lower Bounds…

- So, why does the algorithm work…?

- It is not really an asynchronous system
  - There are bounds on the message delays
  - This is almost a synchronous system…
- We used authenticated messages
  - It can be verified if a server really sent a certain message
- The algorithm takes more than 5 rounds in the worst case
  - It takes more than f rounds!

Messages do not just "arrive eventually"

Why?

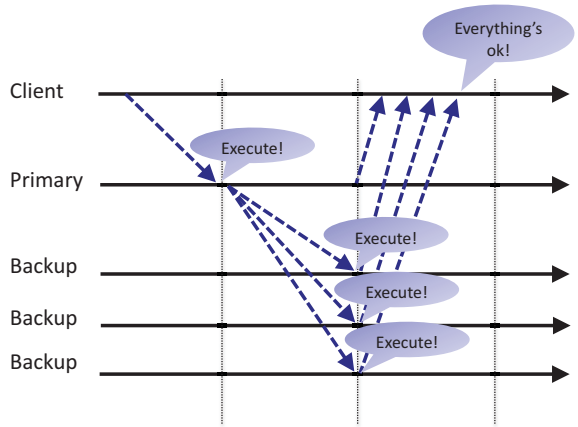
Zyzzyva

- Zyzzyva is another BFT protocol
- Idea
  - The protocol should be very efficient if there are no failures
  - The clients speculatively execute the command without going through an agreement protocol!

- Problem
  - States of correct servers may diverge
  - Clients may receive diverging/conflicting responses
- Solution
  - Clients detect inconsistencies in the replies and help the correct servers to converge to a single total ordering of requests

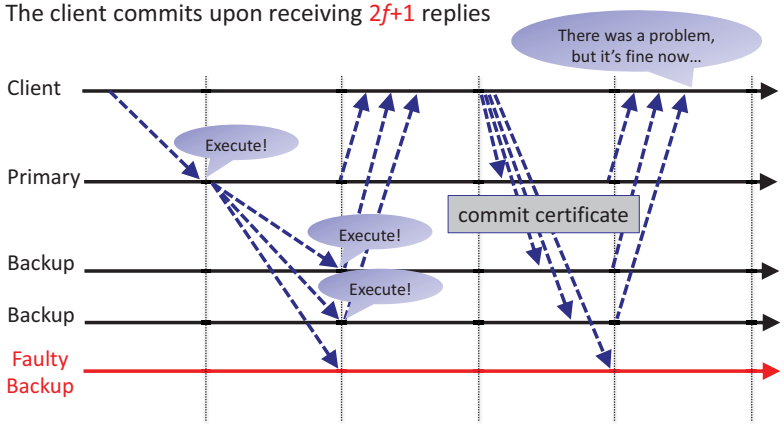
Zyzzyva

- Normal operation: Speculative execution!
- Case 1: All $3f+1$ report the same result



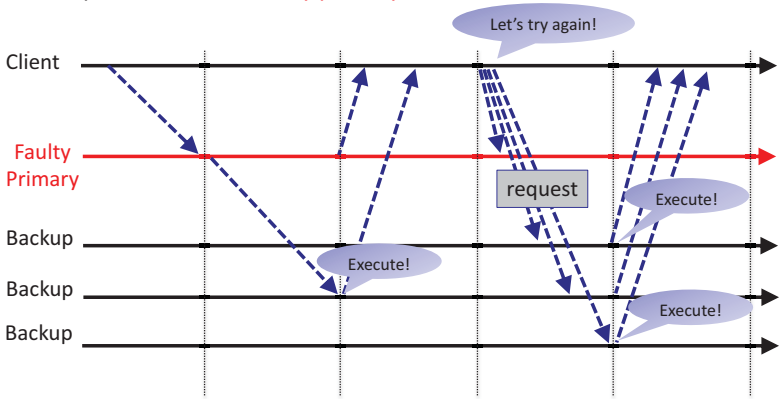
Zyzzyva

- Case 2: Between $2f+1$ and $3f$ results are the same
- The client broadcasts a commit certificate containing the $2f+1$ results
- The client commits upon receiving $2f+1$ replies



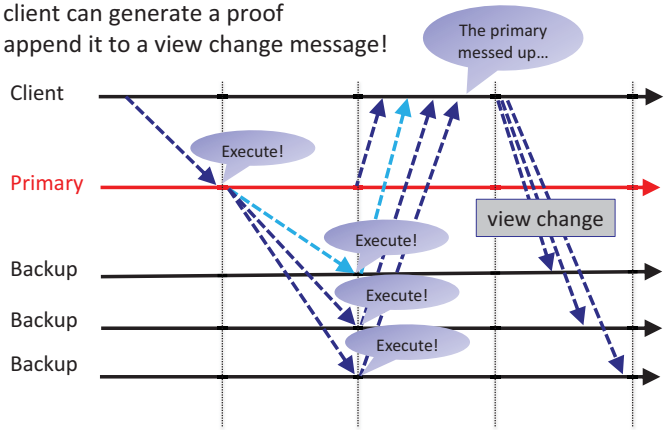
Zyzzyva

- Case 3: Less than $2f+1$ replies are the same
- The client broadcasts its request to all servers
- This step circumvents a faulty primary



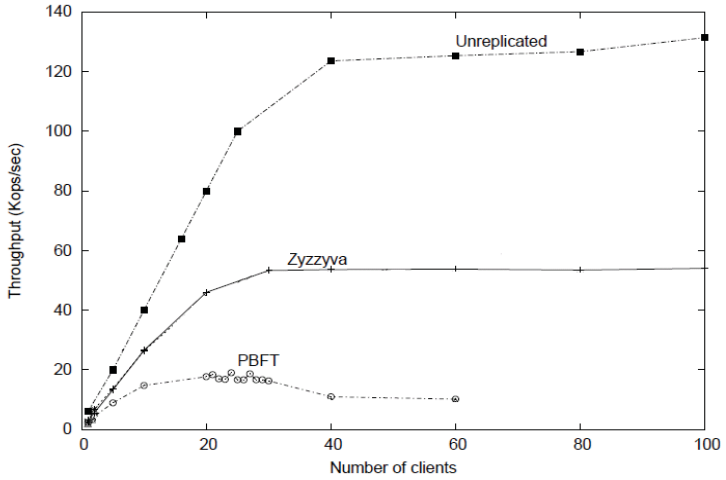
Zyzzyva

- Case 4: The client receives results that indicate an inconsistent ordering by the primary
- The client can generate a proof and append it to a view change message!



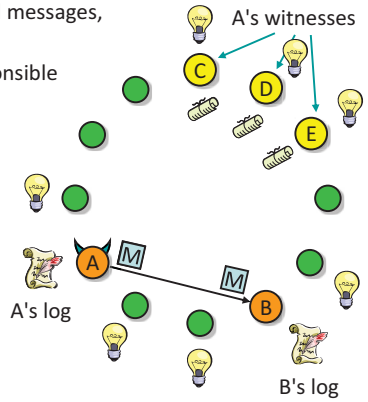
Zyzzyva: Evaluation

- Zyzzyva outperforms PBFT because it normally takes only 3 rounds!



More BFT Systems in a Nutshell: PeerReview

- The goal of PeerReview is to provide accountability for distributed systems
  - All nodes store I/O events, including all messages, in a local log
  - Selected nodes ("witnesses") are responsible for auditing the log
  - If the witnesses detect misbehavior, they generate evidence and make the evidence available
  - Other nodes check the evidence and report the fault
- What if a node tries to manipulate its log entries?
  - Log entries form a hash chain creating secure histories



More BFT Systems in a Nutshell: PeerReview

- PeerReview has to solve the same problems…
  - Byzantine nodes must not be able to convince correct nodes that another correct node is faulty
  - The witness sets must always contain at least one correct node
- PeerReview provides the following guarantees:

1. Faults will be detected
   - If a node commits a fault and it has a correct witness, then the witness obtains a proof of misbehavior or a challenge that the faulty node cannot answer
2. Correct nodes cannot be accused
   - If a node is correct, then there cannot be a correct proof of misbehavior and it can answer any challenge

More BFT Systems in a Nutshell: FARSITE

- "**F**ederated, **A**vailable, and **R**eliable **S**torage for an **I**ncompletely **T**rusted **E**nvironment"
- Distributed file system without servers
- Clients contribute part of their hard disk to FARSITE

- Resistant against attacks: It tolerates $f < n/3$ Byzantine clients
- Files
  - f+1 replicas per file to tolerate f failures
  - Encrypted by the user
- Meta-data/Directories
  - 3f+1 replicas store meta-data of the files
  - File content hash in meta-data allows verification
  - How is consistency established? FARSITE uses PBFT!

More efficient than replicating the files!