EECS 591
Distributed Systems

Manos Kapritsos
Fall 2019

Slides by: Lorenzo Alvisi
PBFT: THE GENERAL IDEA

A

Primary

Replicas

<table>
<thead>
<tr>
<th>A</th>
</tr>
</thead>
</table>
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | ...
Certicates

Protocol steps are justified by certicates
  • Sets (quorums) of signed messages from distinct replicas proving that a property holds

Certicates are of size at least $2f + 1$
  • Any two quorums intersect in at least one correct replica (for safety)
  • There is always a quorum of correct replicas (for liveness)
PBFT: Normal Operation

Three phases:

- **Pre-prepare**: assigns sequence number to request
- **Prepare**: ensures consistent ordering of requests within views
- **Commit**: ensures consistent ordering of requests across views

Each replica maintains the following state:

- Service state
- A message log with all messages sent or received
- An integer representing the replica’s current view
Client issues request

<REQUEST, o, t, c> \(\sigma_c\)

Primary

Replica 1

Replica 2

Replica 3
CLIENT ISSUES REQUEST

<REQUEST, o, t, c> \( \sigma_c \)

- Primary
- Replica 1
- Replica 2
- Replica 3

state machine operation
Client issues request

<REQUEST, o, t, c>_{σ_c}
Client issues request

<REQUEST, o, t, c> \( \sigma_c \)

Primary

Replica 1

Replica 2

Replica 3
Client issues request

<REQUEST, o, t, c> \( \sigma_c \)

Primary

Replica 1

Replica 2

Replica 3

client signature
Primary sends $\langle\langle\text{PRE-PREPARE, } v, n, d>_{\sigma_p}, m\rangle$ to all replicas
Primary sends \( \langle \langle \text{PRE-PREPARE}, v_n d \rangle_{\sigma_p}, m \rangle \) to all replicas.

- Primary
- Replica 1
- Replica 2
- Replica 3

current view
Primary sends \(<\text{PRE-PREPARE, } v, n, d>_{\sigma_p}, m>\) to all replicas.
Primary sends \( \langle\text{PRE-PREPARE, } v, n, d\rangle_{\sigma_p, m} \) to all replicas
Primary sends $\langle\langle \text{PRE-PREPARE}, v, n, d\sigma_p, m \rangle \rangle$ to all replicas

digest of m
**PRE-PREPARE**

Primary sends \( <<\text{PRE-PREPARE}, v, n, d>_{\sigma_p}, m> \) to all replicas

---

Correct backup \( k \) accepts PRE-PREPARE if:

- message is well formed
- \( k \) is in view \( v \)
- \( k \) has not accepted another PRE-PREPARE message for \( v, n \) with a different \( d \)
- \( n \) is between two watermarks \( L \) and \( H \)
  (to prevent sequence number exhaustion)
Primary sends $\langle\langle\text{PRE-PREPARE}, v, n, d\rangle_\sigma, m\rangle$ to all replicas.

Each accepted PRE-PREPARE message is stored in the accepting replica's message log (including the primary's).
ADMINISTRIVIA

Fall study break
- No class on Monday 10/14
- I will still hold OH on 10/15

Problem set #2
- Due Wednesday 10/16, before class

Midterm
- Monday 10/21, 3-5pm, in class

Research part
- Starts on Wednesday 10/23 with Fast Paxos
Replica $k$ sends $<\text{PREPARE, v, n, d, k}>_{\sigma_k}$ to all replicas.

Pre-prepare phase
Replica $k$ sends $\langle \text{PREPARE}, v, n, d, k \rangle_{\sigma_k}$ to all replicas.

Correct backup $k$ accepts PREPARE if:
- message is well formed
- $k$ is in view $v$
- $n$ is between two watermarks $L$ and $H$
Replica **k** sends \(<\text{PREPARE}, v, n, d, k>\sigma_k\>\) to all replicas

- Replicas that send a PREPARE accept the assignment of \(m\) to sequence number \(n\) in view \(v\)
- Each accepted PREPARE message is stored in the accepting replica’s message log
P-Certificates ensure consistent order of requests within views.

A replica produces a P-Certificate($m,v,n$) iff its log holds:
- the request $m$
- A PRE-PREPARE for $m$ in view $v$ with sequence number $n$
- $2f$ PREPARE from distinct backups that match the PRE-PREPARE

A P-Certificate($m,v,n$) means that a quorum agrees to assign $m$ to sequence number $n$ in view $v$
- No two non-faulty replicas with P-Certificate($m,v,n$) and P-Certificate($m',v,n$)
P-Certificates are not enough

- A P-Certificate proves that a quorum of $2f + 1$ replicas has agreed to assign $m$ to sequence number $n$ in view $v$

- Yet that assignment could be modified if a view change happens (the primary changes)
  - The new primary may not be convinced to assign $m$ to $n$ in the new view $v'$
**P-Certificates are not enough**

- Yet that assignment could be modified if a **view change** happens (the primary changes)
  - The new primary may not be convinced to assign \( m \) to \( n \) in the new view \( v' \)
  - \( 2f + 1 \) prepares means at least \( f + 1 \) correct replicas received a pre-prepare for \((m,v,n)\)
After collecting a P-Certificate, replica $k$ sends $\sigma_k^{<\text{COMMIT}, v, n, d, k>}$ to all replicas.
**Commit Certificate**

- C-Certificate: a collection of $2f + 1$ commits

- C-Certificates ensure consistent order of requests across views
  - Cannot miss a P-Certificate during view change

- A replica executes a request when:
  - it gets a C-Certificate for it
  - it has executed all requests with smaller sequence numbers