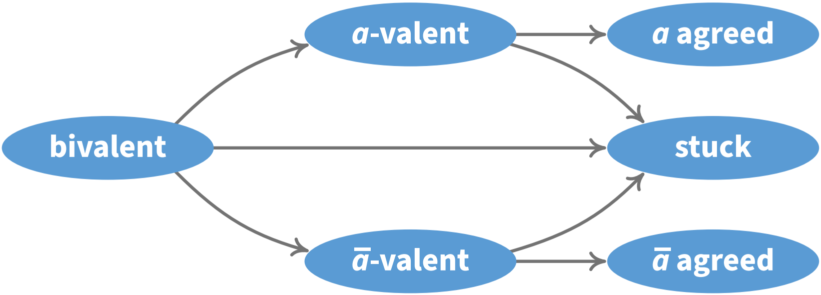
What voting gives us



- **You might get system-wide agreement or you might get stuck**
- **Can’t vote directly on consensus question (i.e., log entry)**
- **What can we vote on without jeopardizing liveness?**
  1. Statements that never get stuck (irrefutable), and
  2. Statements whose hold on consensus question can be broken if stuck (neutralizable)

Paxos [Lamport]

- **A *ballot* is a pair $\langle n, x \rangle$**
  - n – a counter to ensure arbitrarily many ballots exist
  - x – a candidate output value for the consensus protocol
- **Conceptually vote to *commit* and *abort* ballots**
  - If a quorum votes to commit $\langle n, x \rangle$ for any n, it is safe to output x
- **Invariant: all committed and stuck ballots must have same x**
- **To preserve: can't vote to commit a ballot before *preparing* it**
  - Prepare $\langle n, x \rangle$ by aborting all $\langle n', x' \rangle$ with $n' \leq n$ and $x' \neq x$.
  - PREPARED message votes to abort all lower ballots not containing x (or all lower ballots period if previous is NULL)
- **If ballot $\langle n, x \rangle$ stuck, neutralize by restarting with $\langle n+1, x \rangle$**
  - Can prepare $\langle n+1, x \rangle$ even if $\langle n, x \rangle$ is stuck

Paxos example

candidate values

counter	***a***	***b***	***c***	***d***	***e***	***f***	***g***	***h***
1	?	?	?	?	?	?	?	?
2	?	?	?	?	?	?	?	?
3	?	?	?	?	?	?	?	?
4	?	?	?	?	?	?	?	?

0. **Initially, all ballots are bivalent**
1. **Agree that $\langle 1, g \rangle$ is prepared and vote to commit it**
2. **Lose vote on $\langle 1, g \rangle$; agree $\langle 2, f \rangle$ prepared and vote to commit it**
3. **$\langle 2, f \rangle$ is stuck, so agree $\langle 3, f \rangle$ prepared and vote to commit it**
4. **See T votes to commit $\langle 3, f \rangle$ (commit-valent) and externalize f**
    - At this point nobody cares about $\langle 2, f \rangle$—neutralized
5. **Node failure makes $\langle 3, f \rangle$ stuck, prepare and commit $\langle 4, f \rangle$**

Paxos example

candidate values

counter	***a***	***b***	***c***	***d***	***e***	***f***	***g***	***h***
1	✗	✗	✗	✗	✗	✗	?	✗
2	?	?	?	?	?	?	?	?
3	?	?	?	?	?	?	?	?
4	?	?	?	?	?	?	?	?

0. Initially, all ballots are bivalent

1. Agree that $\langle 1, g\rangle$ is prepared and vote to commit it

2. Lose vote on $\langle 1, g\rangle$; agree $\langle 2, f\rangle$ prepared and vote to commit it

3. $\langle 2, f\rangle$ is stuck, so agree $\langle 3, f\rangle$ prepared and vote to commit it

4. See T votes to commit $\langle 3, f\rangle$ (commit-valent) and externalize f

- At this point nobody cares about $\langle 2, f\rangle$—neutralized

5. Node failure makes $\langle 3, f\rangle$ stuck, prepare and commit $\langle 4, f\rangle$

Paxos example

candidate values

counter	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*
1	✗	✗	✗	✗	✗	✗	✗	✗
2	✗	✗	✗	✗	✗	?	✗	✗
3	?	?	?	?	?	?	?	?
4	?	?	?	?	?	?	?	?

0. **Initially, all ballots are bivalent**
1. **Agree that $\langle 1, g \rangle$ is prepared and vote to commit it**
2. **Lose vote on $\langle 1, g \rangle$; agree $\langle 2, f \rangle$ prepared and vote to commit it**
3. **$\langle 2, f \rangle$ is stuck, so agree $\langle 3, f \rangle$ prepared and vote to commit it**
4. **See T votes to commit $\langle 3, f \rangle$ (commit-valent) and externalize f**
    - At this point nobody cares about $\langle 2, f \rangle$—neutralized
5. **Node failure makes $\langle 3, f \rangle$ stuck, prepare and commit $\langle 4, f \rangle$**

Paxos example

candidate values

counter	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*
1	✗	✗	✗	✗	✗	✗	✗	✗
2	✗	✗	✗	✗	✗	?	✗	✗
3	✗	✗	✗	✗	✗	?	✗	✗
4	?	?	?	?	?	?	?	?

0. **Initially, all ballots are bivalent**
1. **Agree that $\langle 1, g\rangle$ is prepared and vote to commit it**
2. **Lose vote on $\langle 1, g\rangle$; agree $\langle 2, f\rangle$ prepared and vote to commit it**
3. **$\langle 2, f\rangle$ is stuck, so agree $\langle 3, f\rangle$ prepared and vote to commit it**
4. **See T votes to commit $\langle 3, f\rangle$ (commit-valent) and externalize f**
   - At this point nobody cares about $\langle 2, f\rangle$—neutralized
5. **Node failure makes $\langle 3, f\rangle$ stuck, prepare and commit $\langle 4, f\rangle$**

Paxos example

candidate values

counter	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*
1	✗	✗	✗	✗	✗	✗	✗	✗
2	✗	✗	✗	✗	✗	?	✗	✗
3	✗	✗	✗	✗	✗	✓	✗	✗
4	?	?	?	?	?	?	?	?

0. **Initially, all ballots are bivalent**
1. **Agree that $\langle 1, g\rangle$ is prepared and vote to commit it**
2. **Lose vote on $\langle 1, g\rangle$; agree $\langle 2, f\rangle$ prepared and vote to commit it**
3. **$\langle 2, f\rangle$ is stuck, so agree $\langle 3, f\rangle$ prepared and vote to commit it**
4. **See T votes to commit $\langle 3, f\rangle$ (commit-valent) and externalize f**
    - At this point nobody cares about $\langle 2, f\rangle$—neutralized
5. **Node failure makes $\langle 3, f\rangle$ stuck, prepare and commit $\langle 4, f\rangle$**

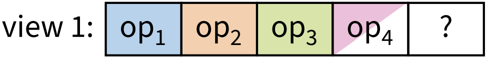
Paxos example

candidate values

counter	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*
1	✗	✗	✗	✗	✗	✗	✗	✗
2	✗	✗	✗	✗	✗	?	✗	✗
3	✗	✗	✗	✗	✗	✓	✗	✗
4	✗	✗	✗	✗	✗	✓	✗	✗

0. **Initially, all ballots are bivalent**
1. **Agree that $\langle 1, g \rangle$ is prepared and vote to commit it**
2. **Lose vote on $\langle 1, g \rangle$; agree $\langle 2, f \rangle$ prepared and vote to commit it**
3. **$\langle 2, f \rangle$ is stuck, so agree $\langle 3, f \rangle$ prepared and vote to commit it**
4. **See T votes to commit $\langle 3, f \rangle$ (commit-valent) and externalize f**
    - At this point nobody cares about $\langle 2, f \rangle$—neutralized
5. **Node failure makes $\langle 3, f \rangle$ stuck, prepare and commit $\langle 4, f \rangle$**

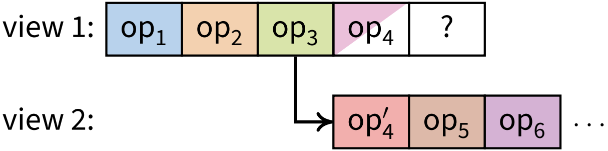
Viewstamped replication [Oki]



view 1:	op$_1$	op$_2$	op$_3$	op$_4$	?

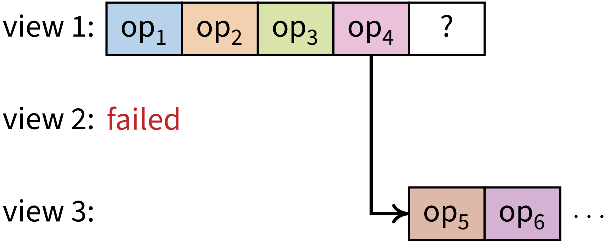
- **Instead of voting on op$_1, \ldots$ directly, vote on $\langle$view 1, op$_1\rangle, \ldots$**
  - Each $\langle$view, op$\rangle$ selected by a single *leader* for view, so irrefutable
  - E.g., chose leader by round-robin using *view#* mod *N*
  - Really, a vote is a promise to include $\langle$view 1, op$_1\rangle$ in future views
- **What if votes on op$_4$ and op$_5$ are stuck (e.g., leader fails)?**
  - Neutralize by agreeing view 1 had only 3 meaningful operations
  - Vote to form view 2 that immediately follows $\langle$view 1, op$_3\rangle$
- **Failed to form view 2 (e.g., a node wants $\langle$view 1, op$_4\rangle$)?**
  - Just go on to form view 3 after $\langle$view 1, op$_4\rangle$

Viewstamped replication [Oki]



- **Instead of voting on $\mathbf{op}_1, \ldots$ directly, vote on $\langle$view 1, $\mathbf{op}_1\rangle, \ldots$**
  - Each $\langle$view, op$\rangle$ selected by a single *leader* for view, so irrefutable
  - E.g., chose leader by round-robin using *view#* $\bmod N$
  - Really, a vote is a promise to include $\langle$view 1, $\mathrm{op}_1\rangle$ in future views
- **What if votes on $\mathbf{op}_4$ and $\mathbf{op}_5$ are stuck (e.g., leader fails)?**
  - Neutralize by agreeing view 1 had only 3 meaningful operations
  - Vote to form view 2 that immediately follows $\langle$view 1, $\mathrm{op}_3\rangle$
- **Failed to form view 2 (e.g., a node wants $\langle$view 1, $\mathbf{op}_4\rangle$)?**
  - Just go on to form view 3 after $\langle$view 1, $\mathrm{op}_4\rangle$

Viewstamped replication [Oki]



- **Instead of voting on $\mathbf{op}_1, \ldots$ directly, vote on $\langle \textbf{view 1}, \mathbf{op}_1 \rangle, \ldots$**
  - Each $\langle$view, op$\rangle$ selected by a single *leader* for view, so irrefutable
  - E.g., chose leader by round-robin using *view#* mod *N*
  - Really, a vote is a promise to include $\langle$view 1, $\text{op}_1\rangle$ in future views
- **What if votes on $\mathbf{op}_4$ and $\mathbf{op}_5$ are stuck (e.g., leader fails)?**
  - Neutralize by agreeing view 1 had only 3 meaningful operations
  - Vote to form view 2 that immediately follows $\langle$view 1, $\text{op}_3\rangle$
- **Failed to form view 2 (e.g., a node wants $\langle \textbf{view 1}, \mathbf{op}_4 \rangle$)?**
  - Just go on to form view 3 after $\langle$view 1, $\text{op}_4\rangle$