Paxos made Live
How Google employs paxos to build a replicated log

Qing Zheng
15799 – Adv Topics in DB Systems
Distributed Consensus
Distributed Consensus

Lock Service

- Exclusive Access
- Synchronization
- ...

Distributed Consensus

**Lock Service**
- Exclusive Access
- Synchronization
- ...

**Name Service**
- Primary Copy
- Partition Table
- Leader / Master
- Membership
- Global Metadata
- ...

Chubby

• Help clients …
  – synchronize activities
  – agree on basic information about their environment
What should Chubby Offer?
What should Chubby Offer?

• Agreement
What should Chubby Offer?

• Agreement
• High Throughput
What should Chubby Offer?

- Agreement
- High Throughput
What should Chubby Offer?

• Agreement
• High Throughput
• Massive Storage
What should Chubby Offer?

- Agreement
- High Throughput
- Massive Storage
What should Chubby Offer?

- Agreement
- High Throughput
- Massive Storage
- Availability
What should Chubby Offer?

• Agreement
• High Throughput
• Massive Storage
• Availability
• Reliable and Fault Tolerant
Why use Paxos?
Why use Paxos?

• Safety
  – bad things never happen
**Why use Paxos?**

- **Safety**
  - bad things never happen

- **Liveness**
  - good things eventually happen
    - as long as only 1 proposer exists eventually
Why use Paxos?

- **Safety**
  - bad things never happen

- **Liveness**
  - good things eventually happen
    - as long as only 1 proposer exists eventually

- **Fault-Tolerant**
  - won’t block
    - as long as a majority of nodes are still live
Why use Paxos?

• Safety
  – bad things never happen

• Liveness
  – good things eventually happen
  • as long as only 1 proposer exists eventually

• Fault-Tolerant
  – won’t block
  • as long as a majority of nodes are still live

No other choices...
Chubby Overview

Chubby

DB

Log

Log Data

local file system

chubby interface (RPC)

insert, delete, lookup, cas

snapshot
Chubby Overview

Chubby

DB

Log

Log Data

Multi-Paxos

local file system

snapshot

chubby interface (RPC)
Chubby Overview

Client

Chubby
Chubby
Chubby
Master
Chubby
Chubby
Chubby

DB
DB
Master
DB
DB

Log
Log
Master
Log
Log
Log

Multi-Paxos
Multi-Paxos

Master Log

Replica Log

Recovered

Failed

New Op
Multi-Paxos

accept <1, 14, op>

Master Log

1 2 ... 13 14

Replica Log

1 2 ... 14

Replica Log

1 2 ... 13 14

Replica Log

1 2 ... 13

Replica Log

1 2 ... 9 14
Multi-Paxos
Multi-Paxos
Multi-Paxos

New Master

Old Master

Replica Log

propose <2, (10-14, 15+)>
Multi-Paxos

- New Master
  - promise <2, 16+>
  - promise <1, 10, op>
  - ... promise <1, 15, op>

- Replica Log
  - 1 2 ... 14 15

- Old Master
  - 1 2 ... 14

- Replica Log
  - 1 2 ... 13

- Replica Log
  - 1 2 ... 14 15
Multi-Paxos

New Master

Replica Log

Old Master

accept
acknowledge
commit

accept
acknowledge
commit

accept
acknowledge
commit
Implementing Paxos

Algorithmic Challenges

Software Engineering

Unexpected Failures
Algorithmic Challenges

- Disk Corruption
- Master Leases
- Group Membership
Software Engineering

- Compiler Support
- Runtime Checking
- Testing
Unexpected Failures

• OS Bugs
• Script Bugs
• Rollback Errors
• System Upgrade
# Measurements

<table>
<thead>
<tr>
<th>Test</th>
<th># workers</th>
<th>file size (bytes)</th>
<th>Paxos-Chubby (100MB DB)</th>
<th>3DB-Chubby (small database)</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ops/s Throughput</td>
<td>1</td>
<td>5</td>
<td>91 ops/sec</td>
<td>75 ops/sec</td>
<td>1.2x</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>5</td>
<td>490 ops/sec</td>
<td>134 ops/sec</td>
<td>3.7x</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>5</td>
<td>640 ops/sec</td>
<td>178 ops/sec</td>
<td>3.6x</td>
</tr>
<tr>
<td>MB/s Throughput</td>
<td>1</td>
<td>8 KB</td>
<td>345 KB/s</td>
<td>172 KB/s</td>
<td>2x</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>8 KB</td>
<td>777 - 949 KB/s</td>
<td>217 KB/s</td>
<td>3.6 - 4.4x</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>32 KB</td>
<td>672 - 822 KB/s</td>
<td>338 KB/s</td>
<td>2.0 - 2.4x</td>
</tr>
</tbody>
</table>
Reference

Thank you.