Improving Wide-area Replication Performance through Informed Leader Election and Overlay Construction

Syed Kewaan Ejaz, Diogo Behrens, Thomas Knauth, Christof Fetzer

TU Dresden, Germany
Data-centers go offline

Hurricane Sandy takes data centers offline with flooding, power outages
Hosting customers stranded as generators in NY data centers run out of fuel.

by Jon Brodkin - Oct 30 2012, 5:25pm CET

Flooding from Hurricane Sandy.
Clients are geographically distributed
Renewed interest in strong consistency

- easier to reason about
- may decrease availability
- main challenge is to achieve reasonable performance
How do WANs differ from LANs?

<table>
<thead>
<tr>
<th></th>
<th>LAN</th>
<th>WAN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>latency</strong></td>
<td>microseconds</td>
<td>milliseconds</td>
</tr>
<tr>
<td><strong>bandwidth</strong></td>
<td>10 GbE</td>
<td>&lt; 1 GbE</td>
</tr>
<tr>
<td><strong>jitter</strong></td>
<td>low</td>
<td>high</td>
</tr>
</tbody>
</table>
Strongly consistent replication basics

- Paxos for state machine based replication
- used by, for example, Google's MegaStore and Spanner systems
- leader orders the clients requests, collects majority of acceptance, then replies client
- leader involved in all decisions, i.e., it's important to pick a “good” leader
- communication: broad- and unicast
Sometimes poor leader is elected

- uninformed election
- site D is a bad leader, e.g., at the end of a high latency link
Broadcast can benefit from overlay

- site A must broadcast result
Relay messages to improve communication

- site B relays message to C and D

![Diagram showing the relay process between sites A, B, C, D, and E. Site A (replication leader) sends a message to site B, which relays it to sites C and D. Site E is also connected to site D.]
Election and Overlay Construction

- each participant collects network performance metrics
- metrics are collected by a coordinator
- coordinator aggregates local views into global view
- global view used to calculate optimal for overlay and elect replication leader
- info distributed to other participants
Optimization Functions

- request rate (election), latency (overlay), global bandwidth (overlay), quorum bandwidth (overlay)
- many possible more possible

- topology
- global optimization
- quorum optimization
Experimental setup

- YCSB benchmark with varying ratios of reads and writes
- Implemented strong consistency on top of Redis key-value store
  - only provides weak consistency semantics by default
- 5 server machines running Linux
- Emulated wide-area network with netem Linux module
Optimizing the overlay for bandwidth

A: 50R/50U  B: 95R/05U  D: 95R/05I  E: 95S/5I

Star topology
Optimizing the overlay for bandwidth

Bandwidth topology

A: 50R/50U B: 95R/05U D: 95R/05I E: 95S/5I

Bar chart showing total throughput (k.ops/s) for workloads A, B, D, and E.

- A: 50R/50U
- B: 95R/05U
- D: 95R/05I
- E: 95S/5I

Legend:
- Star
- Bandwidth
- Quorum
Optimizing the overlay for bandwidth

Quorum topology

- **Star**
- **Bandwidth**
- **Quorum**

<table>
<thead>
<tr>
<th>Workload</th>
<th>A: 50R/50U</th>
<th>B: 95R/05U</th>
<th>D: 95R/05I</th>
<th>E: 95S/5I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quorum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Total throughput (k.ops/s)**

- **majority**
- **full replication**
Optimizing for latency

![Bar chart showing 99th percentile latency for different workloads (A, B, D, E). The chart compares two categories: Star and Latency. Workload D has the highest latency, followed by E, while A and B have similar latencies.]
Summary

- Want to achieve wide-area replication with strong consistency and good performance
- Optimize replication leader and messaging w.r.t. observed network performance metrics
- **unmodified Paxos implementation**
- Examples with 1.5x up to 3x performance improvement (depends on topology)