Adaptive Optimistic Total-Order Protocols for Wide-Area Database Replication

Large-Scale Distributed Systems and Middleware (LADIS)

Luís Rodrigues  José Mocito  Nuno Carvalho

University of Lisbon

March 21, 2007
Road-map

1 Introduction
2 Optimistic TO
3 Optimal SETO Protocol
4 Switching Protocol
5 Conclusions and Future Work
6 Bibliography
The GORDA Project

- GORDA (Open Replication of Databases) is an IST STREP project.
- Replication as a means to offer:
  - High availability, disaster recovery, load balance (read transactions).
- Different replication strategies:
  - Many of these strategies rely on total order protocols.
Optimistic Total Order

• Total Order (TO) is a communication primitive that, usually, involves several communication steps.

• Optimistic total order:
  • Provides an early estimate of the final order.
  • Application steps may be executed in parallel with the communication steps of the TO algorithm.

• Issues:
  • Which is the best optimistic protocol for wide-area operation?
  • How to switch in run time between two protocols?
Road-map

1. Introduction
2. Optimistic TO
3. Optimal SETO Protocol
4. Switching Protocol
5. Conclusions and Future Work
6. Bibliography
Optimism in Total Order Deliveries

- Message is sent.
Optimism in Total Order Deliveries

- Message is sent.
- Message final delivery occurs.
Optimism in Total Order Deliveries

- Message is sent.
- Message final delivery occurs.
- Optimistic delivery occurs.
Timeline for Optimistic Deliveries

- Uniform delivery:
  - Final delivery for the database application (stronger guarantees).
  - If a process delivers all correct processes deliver.
Timeline for Optimistic Deliveries

- Spontaneous order:
  - The order by which messages are received from the network.
  - Inaccurate estimate of final order.
Timeline for Optimistic Deliveries

- Statistically Estimated Total Order (SETO):
  - An optimistic total order protocol for wide area networks was proposed by [Sousa et al. 2002].
  - Accurate in stable networks (but is it the best?).
Timeline for Optimistic Deliveries

- Regular Total Order:
  - Any “classical” total order mechanism (sequencer, token hybrid, etc).
  - Estimate only fails if crashes occur (accurate but slower).
Goal

- Use a SETO variant when the network is stable.
- Use a regular variant when the network is unstable.
- Dynamically switch among different approaches with minimum overhead.
Road-map

1. Introduction
2. Optimistic TO
3. Optimal SETO Protocol
4. Switching Protocol
5. Conclusions and Future Work
6. Bibliography
Simple Network

P1 \quad 10 \quad P2
Spontaneous Order
Delay Assignment 1

P1  P2

[0,10]  [10,0]
[+15,+5]  [+5,+15]
[15,15]  [15,15]

P1  10  P2
Delay Assignment 2

P1 → P2

[10, 10] → [10, 10]
[+10, 0] → [0, +10]
[0, 10] → [10, 0]
Sequencer Based Order

[0,10]  [10,20]

[0,10]  [10,0]

P1 -> 10 -> P2
Sequencer Based Order

[0,10] [10,20]
[0,0] [0,+20]
[0,10] [10,0]

P1 P2

10
Proposed by [Sousa et al. 2002].
Optimistic variant of the sequencer based algorithm.
  - The authoritative total order is the spontaneous order observed by the sequencer process.
Each node cleverly delays received messages to mimic the spontaneous order observed by the sequencer.
The delay calculation relies on a statistically estimation of the latency among nodes.
SETO Delay Assignment

\[ \begin{bmatrix} 5, 7, 7 \end{bmatrix} \quad \begin{bmatrix} 7, 9, 9 \end{bmatrix} \]

\[ \begin{bmatrix} 7, 9, 9 \end{bmatrix} \]
SETO Delay Assignment

$p_3$ (S)

$p_1$  5 ms  $p_2$

7 ms

5 ms

9 ms

[5, 7, 7]

[7, 9, 9]
Is SETO Optimal?

\[7, 9, 9\] 

\[5, 7, 7\] 

\[7, 9, 9\] 

\[3, 5, 7\] 

\[5, 7, 9\] 

\[7, 9, 11\]
Is SETO Optimal?

\begin{itemize}
\item \([7, 9, 9]\)
\item \([7, 9, 9]\)
\item \([7, 9, 11]\)
\item \([5, 7, 7]\)
\item \([7, 9, 9]\)
\item \([3, 5, 7]\)
\item \([5, 7, 9]\)
\end{itemize}
What we did

- Shown that optimal the optimal solution can be formulated as a linear programming problem.
- Designed a heuristic to approximate the optimal solution.
  - Executed when the use of a solver is not feasible or impractical.
- Designed an algorithm to compute and deploy the assignment.
Evaluation

- Tests performed in a simulated environment consisting of:
  - Network topologies
  - Transmission rates
  - Models that execute the three algorithms at stake
- Network topologies where generated in BRITE
- Networks of 30 nodes
Network Plane Size

Average Overall Latency (ms)

Network plane side (unit)
Sequencer Position

![Graph showing Average Overall Latency (ms) vs. Position of the sequencer (process id). The graph compares Optimal Assignment, Heuristic, and SETO methods. The x-axis represents the position of the sequencer (process id) ranging from 1 to 10, while the y-axis shows the average overall latency in milliseconds (ms), ranging from 0 to 800 ms. The graph indicates fluctuations in latency across different positions, with the Optimal Assignment showing the lowest variability.](image-url)
Road-map

1. Introduction
2. Optimistic TO
3. Optimal SETO Protocol
4. Switching Protocol
5. Conclusions and Future Work
6. Bibliography
Switching Protocols

- Allow for the execution of the most suitable total order algorithm, at any point in time

Previous approaches usually involve:
- the message flow to be explicitly stopped ([Renesse et al., 1998])
- the buffering of messages ([Liu & Renesse, 2000], [Chen et al., 2001])
- the implicit delay in the communication during the transition phase ([Rutti et al., 2006])
Our Approach

- The transition is non-disruptive regarding the communication flow
- Adaptation is transparent to the application
- The total order algorithms are executed without any modifications
  - A wrapping layer coordinates the transition between algorithms

| ADAPTIVE OPT   |  
|---------------|---
| SETO TO       | Regular TO |
Protocol

- Premise: switch from TO-A to TO-B
- Switching protocol:
  - Control message is broadcast to all processes
  - When received, the process starts transmitting using both algorithms at the same time
  - First message to be sent is *flagged*
  - Messages received from TO-A are delivered
  - Messages received from TO-B are buffered in order
  - When a *flagged* message is received from every node:
    - Discard all buffered messages already delivered by TO-A
    - Deliver the remaining messages in order
  - From this point forward:
    - Messages received by TO-A are simply discarded
    - Only TO-B is used to receive and send messages
Evaluation

- Simulations performed using SSFNet
- Simulation scenario:
  - Multiple nodes in a WAN (4 senders).
  - Periodic replacements (2ms) of the same total order algorithm
Switching Protocols Used

- Stoppage based protocol
  - Node coordinates traffic stoppage in all nodes
  - When all traffic is stopped all nodes start using the new TO algorithm
- Protocol by [Rutti et al., 2006]
  - Control message is broadcast in total order
  - When received, each node re-transmits all the messages it has not received yet, using the new algorithm
Transmission Rate

With and without replacement

Proposed protocol VS stoppage based approach
Message Transmission Latency

Proposed protocol

Protocol by [Rutti et al., 2006]
Message Inter-arrival Time

Proposed protocol

Protocol by [Rutti et al., 2006]
Road-map

1 Introduction
2 Optimistic TO
3 Optimal SETO Protocol
4 Switching Protocol
5 Conclusions and Future Work
6 Bibliography
Conclusions and Future Work

Conclusions

- New (optimized) SETO algorithm.
- New switching algorithm.

Future Work

- Integrate with the GORDA’s replication protocols.
