PRACTI Replication
Nalini Belaramani, Mike Dahlin, Lei Gao, Amol Nayate, Arun Venkataramani, Praveen Yalagandula, Jiandan Zheng

presented by Chenjin Liang

March 6, 2013
Taxonomy and challenges
PR-AC-TI
Why is PRACTI hard?

Design
Basic idea
Separation of invalidations and bodies
Partial replication of invalidations

Evaluation
Partial replication
Topology independence
Arbitrary consistency
Taxonomy and challenges
PR-AC-TI
Why is PRACTI hard?

Design
Basic idea
Separation of invalidations and bodies
Partial replication of invalidations

Evaluation
Partial replication
Topology independence
Arbitrary consistency
Taxonomy and challenges

PR-AC-TI

Why is PRACTI hard?

Design

Basic idea
Separation of invalidations and bodies
Partial replication of invalidations

Evaluation

Partial replication
Topology independence
Arbitrary consistency
PR-AC-TI

- Partial Replication: A system can place any subset of data and metadata on any node.
Partial Replication: A system can place any subset of data and metadata on any node.

Arbitrary Consistency: A system provides flexible semantic guarantees, including the ability to selectively enforce both consistency and coherence guarantee.
PR-AC-TI

- **Partial Replication**: A system can place any subset of data and metadata on any node.
- **Arbitrary Consistency**: A system provides flexible semantic guarantees, including the ability to selectively enforce both consistency and coherence guarantee.
- **Topology Independence**: Any node can exchange updates with any other node.
Taxonomy and challenges

PR-AC-TI

Why is PRACTI hard?

Design

Basic idea
Separation of invalidations and bodies
Partial replication of invalidations

Evaluation

Partial replication
Topology independence
Arbitrary consistency
Naive addition of PR to an AC-TI log exchange protocol
Challenge

When a node receives updates, it must not only consistently order updates to the data it cares about but also ensure that it has enough information to order updates for the data of interest to all nodes with which it might communicate in the future.
Taxonomy and challenges

PR-AC-TI

Why is PRACTI hard?

Design

Basic idea
Separation of invalidations and bodies
Partial replication of invalidations

Evaluation

Partial replication
Topology independence
Arbitrary consistency
Taxonomy and challenges

PR-AC-TI

Why is PRACTI hard?

Design

Basic idea

Separation of invalidations and bodies

Partial replication of invalidations

Evaluation

Partial replication

Topology independence

Arbitrary consistency
Basic idea

- *Log* of updates and random-access *Checkpoint*: Log of updates and random-access Checkpoint: modifications are appended to the log and then update the checkpoint, and reads access the checkpoint.

- Ensures that each node's log and checkpoint always reflect a causally consistent view of the system's data.
Basic idea(Cont.)

- An enterprise file system with 10,000 users and 10,000 files per user, each palmtop were required to store 100 bytes of per-object metadata, an average user issued just 100 updates per day.
Basic idea (Cont.)

- An enterprise file system with 10,000 users and 10,000 files per user, each palmtop were required to store 100 bytes of per-object metadata, an average user issued just 100 updates per day.
- 10GB of storage per node.
An enterprise file system with 10,000 users and 10,000 files per user, each palmtop were required to store 100 bytes of per-object metadata, an average user issued just 100 updates per day.

- 10GB of storage per node.
- 100MB/day of bandwidth per node.
Basic idea (Cont.)

- Separation of invalidations and bodies.
Basic idea (Cont.)

- Separation of invalidations and bodies.
- Partial replication of invalidations.
Taxonomy and challenges
PR-AC-TI
Why is PRACTI hard?

Design
Basic idea
Separation of invalidations and bodies
Partial replication of invalidations

Evaluation
Partial replication
Topology independence
Arbitrary consistency
Separation of invalidations and bodies

- PRACTI invalidation streams notify a receiver that writes have occurred, but separate body messages contain the contents of the writes.
Separation of invalidations and bodies

- PRACTI invalidation streams notify a receiver that writes have occurred, but separate body messages contain the contents of the writes.
- A receiver of an invalidation inserts the invalidation into its sorted log and updates its checkpoint.
Separation of invalidations and bodies

- PRACTI invalidation streams notify a receiver that writes have occurred, but separate body messages contain the contents of the writes.
- A receiver of an invalidation inserts the invalidation into its sorted log and updates its checkpoint.
- A node can send any body from its checkpoint to any other node at any time.
Taxonomy and challenges
   PR-AC-TI
   Why is PRACTI hard?

Design
   Basic idea
   Separation of invalidations and bodies
   Partial replication of invalidations

Evaluation
   Partial replication
   Topology independence
   Arbitrary consistency
Partial replication of invalidations

- *Imprecise invalidation*: One or more objects in `targetSet` were updated between `start time` and `end time`.
Partial replication of invalidations

- **Imprecise invalidation**: One or more objects in targetSet were updated between start time and end time.
- Each node maintains a global variable currentVV.
Partial replication of invalidations

- **Imprecise invalidation**: One or more objects in `targetSet` were updated between `start time` and `end time`.
- Each node maintains a global variable `currentVV`.
- Each node maintains for each interest set `IS` the variable `IS.lastPreciseVV`
Partial replication of invalidations (Cont.)

Precise Invalidations

Imprecise Invalidation

Fig. 4: Example imprecise invalidation.
Partial replication of invalidations (Cont.)

1. **Initial State**  
   IS is PRECISE

   - Global State: \( \text{currentVV[node1]} = 100 \)
   - Per-IS State: \( \text{lastPreciseVV[node1]} = 100 \)
   - Per-Obj State:  
     - A VALID 99@node1
     - B VALID 99@node1
     - C VALID 100@node1

2. **Imprecise Inval Arrives**  

   \( I = \{\text{target = \{A,B,C\}, start=101@node1, end=103@node1}\} \)

3. **IS is now IMPRECISE**

   - Global State: \( \text{currentVV[node1]} = 103 \)
   - Per-IS State: \( \text{lastPreciseVV[node1]} = 100 \)
   - Per-Obj State:  
     - A VALID 98@node1
     - B VALID 99@node1
     - C VALID 100@node1

4. **Missing Precise Invals Arrive**  

   \( PI1 = \{A, 101@node1\}, PI2 = \{B, 103@node1\} \)

5. **Final State**  
   IS is PRECISE

   - Global State: \( \text{currentVV[node1]} = 103 \)
   - Per-IS State: \( \text{lastPreciseVV[node1]} = 103 \)
   - Per-Obj State:  
     - A INVALID 101@node1
     - B INVALID 103@node1
     - C VALID 100@node1
Taxonomy and challenges
PR-AC-TI
Why is PRACTI hard?

Design
Basic idea
Separation of invalidations and bodies
Partial replication of invalidations

Evaluation
Partial replication
Topology independence
Arbitrary consistency
Taxonomy and challenges
PR-AC-TI
Why is PRACTI hard?

Design
Basic idea
Separation of invalidations and bodies
Partial replication of invalidations

Evaluation
Partial replication
Topology independence
Arbitrary consistency
Partial replication

Fig. 7: Impact of locality on replication cost.
Partial replication (Cont.)

Fig. 8: Read response time available bandwidth varies for full replication, demand reads, and self-tuning replication.
Taxonomy and challenges
PR-AC-TI
Why is PRACTI hard?

Design
Basic idea
Separation of invalidations and bodies
Partial replication of invalidations

Evaluation
Partial replication
Topology independence
Arbitrary consistency
Topology independence

Fig. 10: Synchronization time among devices for different network topologies and protocols.
Taxonomy and challenges

PR-AC-TI

Why is PRACTI hard?

Design

Basic idea
Separation of invalidations and bodies
Partial replication of invalidations

Evaluation

Partial replication
Topology independence

Arbitrary consistency
Arbitrary consistency

(a) Best consistency (order error) achievable for a given bandwidth cost.
(b) Best unavailability achievable while meeting a required order error of 100.
(c) Bandwidth cost of distributing consistency information.

Fig. 12: Consistency trade-offs (a-b) and costs (c).