Time Stamped Anti-Entropy protocol

Authors: Joan Manuel Marquès, Antonio González, David Mor, Joan-Antoni Vilaseca.

Distributed Systems course

Spring 2022

1

2 Assignment Outline

The aim of this practical assignment is to implement and evaluate a weak-consistency protocol for data dissemination.

The project consists on:

- Implementing the Time Stamped Anti-Entropy (TSAE) protocol [1] into an application that stores cooking recipes in a set of replicated servers.
- Add a remove operation on the recipes application.
- Evaluate how TSAE behaves under different conditions.

1 Time Stamped Anti-Entropy (TSAE) protocol

Time Stamped Anti-Entropy (TSAE) [1] protocol is a weak-consistency protocol that provides reliable and eventual delivery of issued operations.

To get an overview and main ideas about TSAE read from [2]:

1 Section *1. Introduction* (excluding subsection 1.1).

2 Section *2.2. Timestamped anti-entropy* (also interesting to read: *2.1. Kinds of consistency*).

Then, to get all details about the protocol, read *Chapter 5. Weak-consistency communication* from [1]. More precisely, read:

- *5.1.1. Data structures for timestamped anti-entropy*.
- *5.1.2. The timestamped anti-entropy protocol*.
- (If you implement purge) *5.3. Purging the message log*. Instead of using vector acks (loosely-synchronized clocks) as described in this section, you should use unsynchronized clocks, described in section *5.4.4. Anti-entropy with unsynchronized clocks*.

1 Groups

Phase 1: should be done **individually**.

You are strongly advised to do the **phases 2 to 4 in groups of two students (from the same classroom)**, even though it is also possible to do it individually.

2 To Deliver

Two deliverables (more details in each phase):

1 Phase 1 (theoretical exercises and practice).

2 Phases 2 to 4 or second theoretical exercise.

Deadlines are in the course schedule.

1 D1. First Deliverable

The work of this delivery must be done **individually**.

1 1. Phase 1: Theoretical exercise of TSAE protocol

1 1.1. TSAE protocol exercise (no purged log)

Let's suppose that there are 3 hosts (A, B, C) that use TSAE protocol to exchange operations. At the initial time (t0) all hosts have the same state and their logs and summary vectors are as follow:

Summary A = A2, B1, C2 Log A = A1, A2, B1, C1, C2

Summary B = A1, B1, C1 Log B = A1, B1, C1, C2

Summary C = A1, B1, C2 Log C = A1, B1, C1, C2

A1 stands for first operation from host A, B2 stands for the second operation from host B, and so on.

Let’s assume that each anti-entropy session starts and ends at the same instant.

1 1.1.1. Exercise 1

For the following sequence:

Time	Operation
1	Host A executes operation A3
2	Host C executes operation C3
3	Host B executes operation B2
4	Host C does an anti-entropy session with host A
5	Host A does an anti-entropy session with host B
6	Host B executes operation B3
7	Host C executes operation C4
8	Host A executes operation A4
9	Host B does an anti-entropy session with C
10	Host C does an anti-entropy session with A

We want you to provide us with:

1.1 The data structures and operations exchanged during the anti-entropy sessions.

1.2 The data structures (log and summary) at each host after each anti-entropy session.

1.3 Indicate if the final state is consistent, i.e. all hosts have received the same operations. In case it is not consistent, indicate which sessions should be done to reach a consistent state.

(Note: log is not purged)

T1:

Summary A = A3, B1, C2 Log A = A1, A2, A3, B1, C1, C2

Summary B = A1, B1, C1 Log B = A1, B1, C1
Summary C = A1, B1, C2 Log C = A1, B1, C1, C2

T2:
Summary A = A3, B1, C2 Log A = A1, A2, A3, B1, C1, C2
Summary B = A1, B1, C1 Log B = A1, B1, C1
Summary C = A1, B1, C3 Log C = A1, B1, C1, C2, C3

T3:
Summary A = A3, B1, C2 Log A = A1, A2, A3, B1, C1, C2
Summary B = A1, B2, C1 Log B = A1, B1, B2, C1
Summary C = A1, B1, C3 Log C = A1, B1, C1, C2, C3

T4:
Host C sends to host A: C3
Host A sends to host C: A2, A3
Summary A = A3, B1, C3 Log A = A1, A2, A3, B1, C1, C2, C3
Summary B = A1, B2, C1 Log B = A1, B1, B2, C1
Summary C = A3, B1, C3 Log C = A1, A2, A3, B1, C1, C2, C3

T5:
Host A sends to host B: A2, A3, C2, C3
Host B sends to host A: B2
Summary A = A3, B2, C3 Log A = A1, A2, A3, B1, B2, C1, C2, C3
Summary B = A3, B2, C3 Log B = A1, A2, A3, B1, B2, C1, C2, C3
Summary C = A3, B1, C3 Log C = A1, A2, A3, B1, C1, C2, C3

T6:
Summary A = A3, B2, C3 Log A = A1, A2, A3, B1, B2, C1, C2, C3
Summary B = A3, B3, C3 Log B = A1, A2, A3, B1, B2, B3, C1, C2, C3
Summary C = A3, B1, C3 Log C = A1, A2, A3, B1, C1, C2, C3

T7:
Summary A = A3, B2, C3 Log A = A1, A2, A3, B1, B2, C1, C2, C3
Summary B = A3, B3, C3 Log B = A1, A2, A3, B1, B2, B3, C1, C2, C3

Summary C = A3, B1, C4 Log C = A1, A2, A3, B1, C1, C2, C3, C4

T8:

Summary A = A4, B2, C3 Log A = A1, A2, A3, A4, B1, B2, C1, C2, C3

Summary B = A3, B3, C3 Log B = A1, A2, A3, B1, B2, B3, C1, C2, C3

Summary C = A3, B1, C4 Log C = A1, A2, A3, B1, C1, C2, C3, C4

T9:

Host B sends to host C: B2, B3

Host C sends to host B: C4

Summary A = A4, B2, C3 Log A = A1, A2, A3, A4, B1, B2, C1, C2, C3

Summary B = A3, B3, C4 Log B = A1, A2, A3, B1, B2, B3, C1, C2, C3, C4

Summary C = A3, B3, C4 Log C = A1, A2, A3, B1, B2, B3, C1, C2, C3, C4

T10:

Host C sends to host A: B3, C4

Host A sends to host C: A4

Summary A = A4, B3, C4 Log A = A1, A2, A3, A4, B1, B2, B3, C1, C2, C3, C4

Summary B = A3, B3, C4 Log B = A1, A2, A3, B1, B2, B3, C1, C2, C3, C4

Summary C = A4, B3, C4 Log C = A1, A2, A3, A4, B1, B2, B3, C1, C2, C3, C4

Hosts B has pending operations to receive from host A or host C. **The final state is not consistent**.

1 1.1.2. Exercise 2

For the following sequence:

Time	Operation
1	Host A executes operation A3
2	Host A executes operation A4
3	Host B executes operation B2
4	Host C executes operation C3
5	Host B executes operation B3
6	Host B does anti-entropy session with host C
7	Host C does anti-entropy session with host A
8	Host A does anti-entropy session with host B

We want you to provide us with:

1.1 The data structures and operations exchanged during the anti-entropy sessions.

1.2 The data structures (log and summary) at each host after each anti-entropy session.

1.3 Indicate if the final state is consistent, i.e. all hosts have received the same operations. In case it is not consistent, indicate which sessions should be done to reach a consistent state.

(Note: log is not purged)

LSimLogger.log

T1:

Summary A = A3, B1, C2 Log A = A1, A2, A3, B1, C1, C2

Summary B = A1, B1, C1 Log B = A1, B1, C1

Summary C = A1, B1, C2 Log C = A1, B1, C1, C2

T2:

Summary A = A4, B1, C2 Log A = A1, A2, A3, A4, B1, C1, C2

Summary B = A1, B1, C1 Log B = A1, B1, C1

Summary C = A1, B1, C2 Log C = A1, B1, C1, C2

T3:

Summary A = A4, B1, C2 Log A = A1, A2, A3, A4, B1, C1, C2

Summary B = A1, B2, C1 Log B = A1, B1, B2, C1

Summary C = A1, B1, C2 Log C = A1, B1, C1, C2

T4:

Summary A = A4, B1, C2 Log A = A1, A2, A3, A4, B1, C1, C2

Summary B = A1, B2, C1 Log B = A1, B1, B2, C1

Summary C = A1, B1, C3 Log C = A1, B1, C1, C2, C3

T5:

Summary A = A4, B1, C2 Log A = A1, A2, A3, A4, B1, C1, C2

Summary B = A1, B3, C1 Log B = A1, B1, B2, B3, C1

Summary C = A1, B1, C3 Log C = A1, B1, C1, C2, C3

T6:

Host B sends to host C: B2, B3

Host C sends to host B: C2, C3

Summary A = A4, B1, C2 Log A = A1, A2, A3, A4, B1, C1, C2

Summary B = A1, B3, C3 Log B = A1, B1, B2, B3, C1, C2, C3

Summary C = A1, B3, C3 Log C = A1, B1, B2, B3, C1, C2, C3

T7:

Host C sends to host A: B2, B3, C3

Host A sends to host C: A2, A3, A4

Summary A = A4, B3, C3 Log A = A1, A2, A3, A4, B1, B2, B3, C1, C2, C3

Summary B = A1, B3, C3 Log B = A1, B1, B2, B3, C1, C2, C3

Summary C = A4, B3, C3 Log C = A1, A2, A3, A4, B1, B2, B3, C1, C2, C3

T8:

Host A sends to host B: A2, A3, A4

Host B sends to host A: -

Summary A = A4, B3, C3 Log A = A1, A2, A3, A4, B1, B2, B3, C1, C2, C3

Summary B = A4, B3, C3 Log B = A1, A2, A3, A4, B1, B2, B3, C1, C2, C3

Summary C = A4, B3, C3 Log C = A1, A2, A3, A4, B1, B2, B3, C1, C2, C3

All hosts have received the same operations. **The final state is consistent**.

1 1.2. TSAE protocol exercise (purged log)

Imagine a situation with 5 hosts with unsynchronized clocks and log purging where Hosts B and E have the following AckSummaries:

AckSummary B

\	**A**	**B**	**C**	**D**	**E**
A	A 3	B 2	C 1	D 1	E 2
B	A 3	B 4	C 2	D 2	E 4
C	A 1	B 1	C 1	D 1	E 1
D	A 2	B 4	C 1	D 2	E 2
E	A 3	B 4	C 1	D 2	E 3

AckSummary E

\	**A**	**B**	**C**	**D**	**E**
A	A 3	B 1	C 2	D 1	E 4
B	A 2	B 2	C 2	D 1	E 4
C	A 2	B 2	C 3	D 1	E 4
D	A 2	B 2	C 2	D 1	E 4
E	A 3	B 4	C 3	D 2	E 4

1.1 Which is the content of Log in host B?

A2, A3, B2, B3, B4, C2, D2, E2, E3, E4

1.2 Which is the content of Log in host E?

A3, B2, B3, B4, C3, D2

B and E do an anti-entropy session. During the session both know who each other is (this is different from the algorithm in the Golding thesis).

1.3 Which operations are exchanged during the anti-entropy session?

Operations sent by B = -

Operations sent by E = C3

1.4 Which AckSummary and log have each host after ending the session?

Final state:

AckSummary B & E

\	**A**	**B**	**C**	**D**	**E**
A	A3	B2	C2	D1	E4
B	A3	B4	C3	D2	E4
C	A2	B2	C3	D1	E4
D	A2	B4	C2	D2	E4

E	A3	B4	C3	D2	E4

Log B & E = A3, B3, B4, C3, D2

1 1.3. TSAE protocol exercise

1 In TSAE, what is the function of the message delivery? Explain with your own words the question.

The Golding Thesis marks the message delivery component as “The message delivery component can mask out transient network and host failures by delaying messages and resending them after the fault is repaired. It also allows messages to be batched together from transmission, which is often more efficient than transmitting each message singly. Both of these features are especially important for mobile systems that can be disconnected from the Internet for extended periods.” This is not an open answer.

2 What are the extensions to the TSAE algorithm proposed in Golding's doctoral thesis? Explain in detail at least one of the possible answers.

The TSAE protocol as presented requires loosely-synchronized clocks so that each principal can acknowledge messages using a single timestamp (Figure 5.5). If clocks are not synchronized, the clock at one principal may be much greater than the clock at another. If the minimum timestamp were selected to summarize the messages a principal has received, messages from the principal with the fast clock might never be acknowledged.

A principal’s summary vector is a more general and exact measure of the messages that have been received. If the entire summary vector is used as an acknowledgment, then clock values from different hosts need never be compared.

To use summary vectors for acknowledgment, each principal must maintain a two-dimensional acknowledgment matrix of timestamps, as shown in Figure 5.12. The summary vector is part of the acknowledgment matrix: the th column in The matrix is the summary vector for the local principal. Other columns are old copies of the summary vectors from other principals.

The unsynchronized-clock version of the TSAE protocol is little different from the synchronized clock version. During anti-entropy sessions, principals exchange the entire matrix and update the entire matrix using an element wise maximum at the end of a session.

The only other difference arises when the message ordering component is called upon to determine whether a message has been acknowledged by every principal. Consider a message sent from principal p at time t. A principal q knows that every other principal has observed the message when every timestamp in the p row of the message vector q at is greater than t.

This could be a correct and extended answer for the question.

1 2. Phase 1: Implementation and testing of Log and TimestampVector data structures

Phase 1 will consist of implementing the methods from `Log` and `TimestampVector` data structures.

In this phase only *add* operations are issued. Don't implement the functionality to purge the log.

Annex A includes details about the timestamps used in this practical assignment.

NOTE: be **careful with concurrent access to data structures**. Two actions issued by different threads may interleave. **Use** some **synchronization mechanism to avoid interference**.

Implement `Log` and `TimestampVector` data structures according to the TSAE protocol described in section *Time Stamped Anti-Entropy (TSAE) protocol*.

1 Environment

Requires Java 7.

We recommend you to use Eclipse as an IDE. We will provide you an Eclipse project that contains the implementation of the cooking recipes application except the parts related to TSAE protocol.

All scripts for running local tests are prepared for Ubuntu-linux but other OS can be used. In that case, you will be responsible for adapting the scripts to your OS.

2 2.1. Test locally

To test locally `Log` and `TimestampVector` data structures run:

```
$  java -cp ../bin:../lib/* recipes_service.Server --phase1
```

(run it from `scripts` folder)

Introduce a recipe and check if `Log` and `TimestampVector` data structures are correct.

```
$  ./start.sh 20004 --phase1
```

(run it from `scripts` folder)

$1: listening port of `Phase1TestServer`, the server that tests the correctness of your solution. In case that port `20004` is already in use by another application you can change it to any other unused port.

Executes `Log` and `TimestampVector` with a predefined set of users and operations and compares it with a `Log` and `TimestampVector` previously calculated.

1 2.2. Formal evaluation of phase 1

It is compulsory to **use DSLab to assess the phase 1** of the practical assignment. Therefore, once your application runs in local, upload `Log` and `TimestampVector` classes into DSLab (https://sd.uoc.edu/dslab/) and assess them:

1 Create a *Project* and upload `Log` and `TimestampVector` classes into this project.

2 Create an *Experiment* that executes the project created in step 1.

3 Check the result of the execution of step 2.

1 2.3. Things to deliver

A **file** according to the following convention:

Deliverable1-Year-FamilyName1_Name1.zip

This file should **include**:

- solution of the theoretical exercise.
- phase1.pdf: a (short) report detailing and explaining all decisions taken. A proposal of report template is included in the practical assignment distribution (*/doc* folder).

  In addition, you should detail the portions of your source code you consider are most significant.
- Source code: `Log` and `TimestampVector` classes.

The zip file should have a single directory named like the zip file with the following structure (use same structure and names):

Subdirectory	Content
`/doc`	Solution theoretical exercise. Report in pdf.
`/src`	`Log` and `TimestampVector` classes.

1 Annex A. Source code and documentation

`Papers` folder contains referenced paper and thesis that explain the TSAE protocol.

`TSAE` folder contains all packages and classes required for the practical assignment.

Important: **do not implement new classes**. **Modify only the classes indicated in each phase**.

(except if you modify the basic modeling of parameters in phase 4)

Classes that you should modify:

1 package `recipes_service.tsae.datastructures`:

- `Log`: class that logs operations.
- `TimestampVector`: class to maintain the summary.
- `TimestampMatrix`: class to maintain the acknowledgment matrix of timestamps

2 package `recipes_service.tsae.sessions`:

- `TSAESessionOriginatorSide`: Originator protocol for TSAE.
- `TSAESessionPartnerSide`: Partner's protocol for TSAE.

3 package `recipes_service`:

- `ServerData`: contains `Server`'s data structures required by the TSAE protocol (`log`, `summary`, `ack`) and the application (`recipes`). You can add any required method to allow `TSAESessionOriginatorSide` and `TSAESessionPartnerSide` to manipulate these data structures.
  - `addRecipe` method: adds a new recipe.
  - `removeRecipe` method: removes a recipe.

Classes that you should use but NOT modify:

1 package `recipes_service.tsae.data_structures`:

- `Timestamp`: a timestamp. A timestamp allows the ordering of operations issued from the same host. It is a tuple <hostId, sequenceNumber>. Sequence number is a number that grows monotonically. The first valid timestamp issued by a host will have an initial value of 0. A negative sequence number means that the host hasn't issued yet any operation. Timestamps can not be used to order operations issued in different hosts. Next timestamp is obtained by calling the method `nextTimestamp()` from class `ServerData` (package `recipes_service`).

1 package `recipes_service.data`:

- `AddOperation`: an add operation (operations are logged in the `Log` and exchanged with other partners).
- `RemoveOperation`: a remove operation (operations are logged in the `Log` and exchanged with other partners).

- `Recipe`: a recipe.
- `Recipes`: class that contains all recipes.

1 package `recipes_service.communication`)

- `MessageAErequest`: message sent to request an anti-entropy session.
- `MessageOperation`: message sent for each operation exchanged during an anti-entropy session.
- `MessageEndTSAE`: message sent to finish an anti-entropy session.

1 package `communication`:

- `ObjectInputStream_DS`: class that implements a modification of the `ObjectInputStream` to simulate failures. Use the method `readObject`().
- `ObjectOutputStream_DS`: class that implements a modification of the `ObjectOutputStream` to simulate failures. Use the method `writeObject`().

1 package `recipes_service.activitysimulation`: Only in case that in phase 4 you want to better understand or modify the modeling of activity and dynamicity.

1 Annex B. Activity simulation and dynamicity

Simulation of communication failures

`ActivitySimulation` class (`package recipes_service.activitysimulation`) decides when to simulate the disconnection (or network failure) of a host and when to reconnect it.

To simulate communication failures, a disconnected host abruptly closes all its Input and Output streams, which results in an exception in the two partners that are using the stream.

2 References

[1] **Richard A. Golding** (1992, December). Weak-consistency group communication and membership. Ph.D. thesis, published as technical report UCSC-CRL-92-52. Computer and Information Sciences Board, University of California. Santa Cruz. **(chapter 5)**

[2] **R. Golding; D. D. E. Long**. The performance of weak-consistency replication protocols, UCSC Technical Report UCSC-CRL-92-30, July 1992.