Brief Announcement: Selective Anti-Entropy

Weibin Zhao and Henning Schulzrinne
Columbia University
New York, NY 10027
{zwb,hgs}@cs.columbia.edu

Anti-entropy [1] is an important mechanism to achieve eventual consistency among a set of replicas. We refer to traditional anti-entropy [2, 3] as *complete* anti-entropy, where two replicas reconcile all inconsistent states between them in a two-way session or two one-way sessions. Complete anti-entropy works well in full replications, but may cause a summary problem (a replica cannot summarize its received updates using a summary vector) in partial replications (where the state set is divided into subsets referred to as *scopes*). To avoid the summary problem, we propose to use *selective* anti-entropy, which is to selectively reconcile inconsistent states between two replicas in a session, i.e., a replica specifies the subsets of updates it solicits, and thus controls what it receives from another replica.

In selective anti-entropy, all updates that are initially received by the same replica R_x belong to the same subset $\overline{R_x}$. The subset of a state is determined by the last update applied to it. A selective anti-entropy session where R_x requests new updates in $\overline{R_{x_1}}$, $\overline{R_{x_2}}$, ..., $\overline{R_{x_k}}$ from R_z is denoted as $\Theta(R_x, \{\overline{R_{x_1}}, \overline{R_{x_2}}, ..., \overline{R_{x_k}}\}, R_z)$. We distinguish four types of selective anti-entropy sessions (listed in Figure 1) based on the number of subsets requested and whether the session is direct or indirect. We assume that a replica R_x does not request updates in $\overline{R_x}$ from other replicas. A session is direct if it only requests updates in $\overline{R_z}$ from R_z (e.g., $\Theta(R_x, \{\overline{R_z}\}, R_z)$), otherwise the session is indirect (e.g., $\Theta(R_x, \{\overline{R_y}\}, R_z)$) or mixed (e.g., $\Theta(R_x, \{\overline{R_y}, \overline{R_z}\}, R_z)$). A select-multiple or select-all session can be viewed as comprising k ($k \geq 2$) select-one sessions, in which at most one is select-one-direct session, and the rest are select-one-indirect sessions. A select-all session is equivalent to a complete anti-entropy session.

An anti-entropy session is *safe* if it will not cause the summary problem. A replica can always choose safe sessions by using selective anti-entropy. Any select-one-direct anti-entropy session, such as $\Theta(R_x, \{\overline{R_z}\}, R_z)$, is safe. A select-one-indirect anti-entropy session $\Theta(R_x, \{\overline{R_y}\}, R_z)$ is safe if $S(R_x) \cap S(R_y) \subseteq S(R_z)$, where $S(R_x)$, $S(R_y)$ and $S(R_z)$ denote the scopes of R_x, R_y and R_z, respectively. In other words, if $S(R_x) \cap S(R_y) \supset S(R_z)$, $\Theta(R_x, \{\overline{R_y}\}, R_z)$ may cause the summary problem. In Figure 2, for example, $\Theta(R_3, \{\overline{R_1}\}, R_6)$ is safe because $S(R_3) \cap S(R_1) = \{S_1, S_2\} = S(R_6)$, but $\Theta(R_3, \{\overline{R_6}\}, R_2)$ may not be safe since $S(R_3) \cap S(R_6) = \{S_1, S_2\} \supset S(R_2) = \{S_1\}$. A select-multiple or select-all anti-entropy session is safe if each of its select-one-indirect sessions is safe.

A replica needs to perform complete anti-entropy sessions sequentially, but it can perform multiple selective anti-entropy sessions in parallel, such as all select-one-direct sessions. In general, R_x can perform k selective anti-entropy sessions $(s_1, s_2, ..., s_k)$ in parallel if and only if the requested update subsets $(u_1, u_2, ..., u_k)$ in these sessions do not overlap, i.e., $u_1 \cap u_2 \cap ... \cap u_k = \phi$. For instance, assume $k = 2$, $u_1 = \{\overline{R_2}, \overline{R_4}\}$, and $u_2 = \{\overline{R_3}, \overline{R_5}\}$, then session s_1 and s_2 can be performed in parallel since $u_1 \cap u_2 = \phi$. Running anti-entropy sessions in parallel at a replica can improve performance since the replica does not need to wait until a session is finished before it can start another session.

To use selective anti-entropy, a replica needs to have membership information about its replica set.

Type	Number of Subsets	Direct/Indirect
select-one-direct	1	direct
select-one-indirect	1	indirect
select-multiple	≥ 2 but $< r - 1$	indirect/mixed
select-all	$r - 1$	mixed

Figure 1: Types of selective anti-entropy sessions, where r is the number of replicas in the replica set.

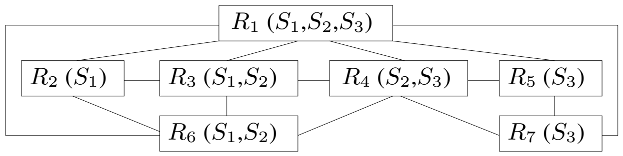

Figure 2: A partial replication system: replicas are R_1 to R_7, scopes are S_1 to S_3, and an edge between two replicas means that they share scopes.

1. REFERENCES

[1] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry. Epidemic algorithms for replicated database maintenance. In *The 6th ACM symposium on principles of distributed computing*, August 1987.

[2] R. Golding. *Weak-consistency group communication and membership*. PhD thesis, Computer Science, University of California at Santa Cruz, December 1992.

[3] K. Petersen, M. Spreizer, D. Terry, M. Theimer, and A. Demers. Flexible update propagation for weakly consistent replication. In *The 16th ACM symposium on operating systems principles*, October 1997.

PODC 2002, July 21-24, 2002, Monterey, California, USA.