Distributed Sagas

McCaffrey, Caitie
Sporty Tights, Inc

Kingsbury, Kyle
The SF Eagle

Narula, Neha
That's DOCTOR Narula to you!

May 20, 2015

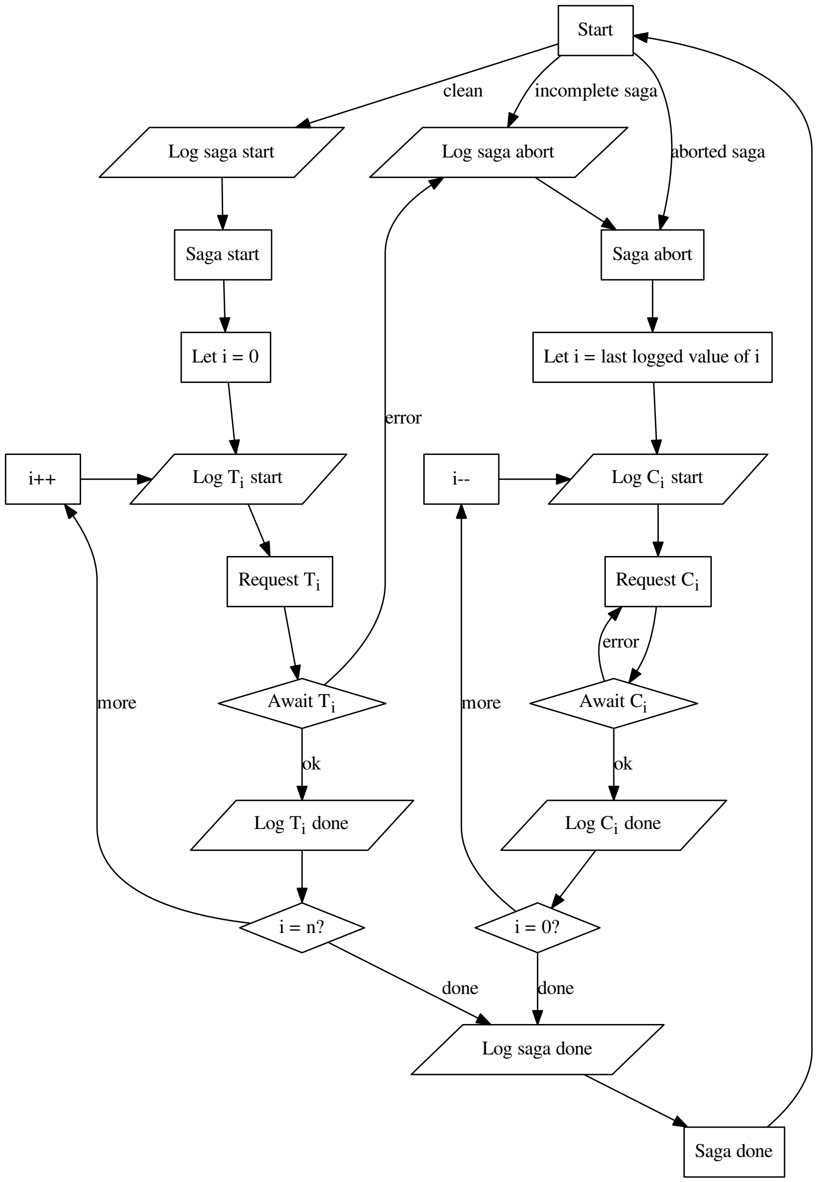
1 Introduction

The saga paper outlines a technique for long-lived transactions which provide atomicity and durability without isolation (what about consistency? Preserved outside saga scope, not within, right?). In this work, we generalize sagas to a distributed system, where processes communicate via an asynchronous network, and discover new constraints on saga sub-transactions.

We are especially interested in the problem of writing sagas which interact with *third-party services*, where we control the Saga Execution Coordinator (SEC) and its storage, but not the downstream Transaction Execution Coordinators (TECs) themselves. Communication between the SEC and TEC(s) takes place over an asynchronous network (e.g. TCP) which is allowed to drop, delay, or reorder messages, but not to duplicate them.

We assume a high-availability SEC service running on multiple nodes for fault-tolerance, where multiple SECs may run concurrently. They coordinate their actions through a linearizable data store, which ensures saga transactions proceed sequentially.

2 The Saga Execution Coordinator



3 Both Rollback and Roll-forward

Lemma 3.1. *If T_i is received by a TEC, then $T_0, T_1, ...T_{i-1}$ have already been acknowledged by a TEC, where $0 < i \leq n$.*

Proof. In order for T_i to be received by a TEC, it must have been requested by an SEC. In a roll-forward SEC, this could be a retry of a failed attempt to execute T_i, but regardless of whether the SEC is roll-back or roll-forward, entering that part of the algorithm requires the SEC to journal its intent to start T_i.

There are only two paths to that journaling operation. The first case, $i = 0$, falls outside our constraint $0 < i \leq n$. Therefore the SEC *must* have taken the other path: incrementing i before beginning a new transaction.

That path depends on $i - 1 \neq n$ being false, which holds since we are considering $i \leq n$. That in turn depends on journaling T_{i-1}'s completion, which depends on a successful response from a TEC for T_{i-1}. Therefore some TEC acknowledged T_i. That in turn requires that TEC to have received T_i.

So, the receipt of T_i implies both the receipt and acknowledgement of T_{i-1}. By induction, receiving T_i implies *all* transactions $T_0, T_1, ...T_{i-1}$ have been acknowledged.

□

Corollary 3.1.1. *The first transaction to be received and acknowledged is T_0.*

Proof. Assume the first transaction to be processed is not T_0, but rather, some $T_i \mid 0 < i \leq n$. By 3.1, T_{i-1} must have been received and acknowledged by a TEC already. T_i is therefore *not* the first transaction: a contradiction.

□

Lemma 3.2. *If C_i is received by a TEC, then T_{i-1} must have been acknowledged by some TEC, where $0 < i \leq n$.*

Proof. Receipt of C_i by a TEC implies the request of C_i by some SEC. An SEC can only request C_i if it logs its intent to start C_i, which can occur by two paths: either the completion of C_{i+1}, or by the initialization of i to its last logged value. Both branches imply the SEC read i, or some higher value, from storage.

i is only incremented by an SEC which has successfully completed T_{i-1}. Since i is nonzero, it was incremented, and T_{i-1} was acknowledged by some TEC.

□

Lemma 3.3. *If C_i is requested, T_i may or may not have been requested.*

Proof. We know from 3.2 that C_i implies the acknowledgement of all T_j where $0 \leq j < i$. But what of that final transaction, T_i? Can we guarantee its completion?

The answer is no. All that is necessary for C_i to occur is for an SEC to write T_i's start. If the SEC crashes just after journaling, it will never request T_i. If it does not crash, T_i will be requested.

□

Lemma 3.4. *If C_i is the highest compensating transaction requested, no T_j will ever have been requested, for all $i < j$.*

Proof. Assume some T_j subsequent to T_i *is* requested. Then some SEC must have written j to storage prior to that request. In order to reach C_i, an SEC must have received acknowledgement for C_j first, which implies C_i is not the highest compensating transaction requested: a contradiction.

□

Lemma 3.5. *If T_i is the highest transaction requested, no C_j will ever have been requested, for all $i + 1 < j$.*

Proof. Assume some C_j *is* eventually requested. Then some SEC must have written j to disk, which implies T_{j-1} was acknowledged. Since T_{j-1} was requested, and $i < j - 1$, T_i cannot have been the highest transaction requested: a contradiction.

□

Lemma 3.6. *If a saga completes successfully, every transaction T_i will have been acknowledged at least once, for $0 \leq i \leq n$.*

Proof. A saga can complete successfully iff the highest transaction T_n has been acknowledged. By 3.1, every T_i must *also* have completed, where $0 \leq i < n$.

□

Lemma 3.7. *If a saga completes the abort process, and T_i was received by a TEC, C_i was also acknowledged by a TEC.*

Proof. Let C_m be the highest compensating transaction acknowledged. Assume C_i was not received: $m < i$. By 3.4, no transaction T_i with $m < i$ can ever occur, so $i \leq m$—which contradicts $m < i$. C_i must have been acknowledged.

□

4 Rollback

Lemma 4.1. *Transactions are requested and received at most once.*

Proof. In order for an SEC to request a transaction T_i, it has to record its intent to execute T_i in shared SEC storage. Since that storage is linearizable, any other SEC recording an intent to execute T_i would be visible to the requesting SEC.

Case 1 Another SEC has already recorded its intent to request T_i. The given SEC chooses to crash instead of requesting T_i.

Case 2 No other SEC has recorded its intent to request T_i. The given SEC requests T_i once.

In both cases, T_i is requested at most once, across all SECs, depending on whether or not the successfully-recording SEC crashes before making its request.

Because the network does not duplicate requests, the number of times T_i can arrive at a TEC is less than or equal to the number of requests any SEC makes for T_i. Since that number is at most one, T_i is received at most once.

□

Lemma 4.2. *Transactions are seen by TECs in sequential order:* $T_0, T_1, \ldots, T_j$, *where* $0 \leq j \leq n$.

Proof. Consider a sequential history $S = (T_0, ..., T_i)$ followed by T_j. Is $(T_0, ...T_i, T_j)$ sequential? We must show $i + 1 = j$.

Case 1 Assume $j \leq i$. Then T_j is a duplicate of some transaction already in S, which violates 4.1: a contradiction.

Case 2 Assume $i + 1 < j$. By 3.1, T_{i+1} must appear before T_j—but S cannot contain T_{i+1}, since it only ranges from 0 to T_i.

Case 3 Assume $i < j \leq i + 1$. Then $i + 1 = j$.

Since cases 1 and 2 are impossible, *any* history comprised of a transaction following a sequential history of at least one element must be sequential as well.

Now, consider histories of one element or fewer:

Case 1 No transactions occur. The history is trivially sequential.

Case 2 Exactly one transaction occurs. By 3.1.1, that transaction must be T_0. This history is trivially sequential.

So any history of one element or fewer is sequential, and any transaction *appended* to that history will also form a sequential history, and so on. By induction, all transactions in a rollback saga system occur sequentially.

□