ログ転送グループ化による 外部整合的トランザクション処理の高性能化

堀江 悠樹[1,a)]　梶原 顕伍[1]　川島 英之[2]　建部 修見[1]

概要：分散トランザクション処理を実行する際、外部整合性（external consistency, strong 1SR）が求められる。これを実現するために single master 方式では master node でトランザクション処理を行い、その結果を backup node に複製する方式が主である。この方式ではトランザクションを 1 つずつ実行するために性能が犠牲になる。そこで本研究ではログ転送グループ化を提案する。提案手法により複数のクライアントから到着するトランザクションを一括転送可能により、効率的な処理が可能になる。提案手法を分散合意プロトコル Raft と並行性制御法 S2PL を用いて設計、実装、評価した。

Yuki Horie[1,a)]　Kengo Kajiwara[1]　Hideyuki Kawashima[2]　Osamu Tatebe[1]

Abstract: When executing distributed transaction processing, external consistency or strong 1SR is required. In order to achieve this, the single master system mainly performs transaction processing on only a single master node and replicates the result to the backup nodes. This method sacrifices performance because the transactions are executed one by one sequentially. In this paper, we propose the log transfer grouping method. With the proposed method, transactions arriving from multiple clients can be buffered on the master node and they are executed in a batch, thereby enabling efficient processing. The proposed method is designed, implemented and evaluated using the distributed consensus protocol Raft and the concurrency control protocol S2PL.

1. はじめに

システムを高信頼化するために、複数のノードでデータの複製を保有する分散データベースシステムが広く使われている。このような分散データベースにおいてデータベースの一貫性を保ちながらデータアイテムの書き換えを行うには、分散トランザクション処理が必要になる。分散トランザクションにおいて要求される性質は、consistency と isolation である。Consistency とは、あるデータアイテムを複数のプロセスが観測した際の見え方に関する基準である。最も堅牢な consistency は linearizabile と言われる。Isolation とは複数のトランザクションがデータベースにアクセスした際の、その値の見え方に関する基準である。最も堅牢な isolation は serializable と言われる。本研究では linearizable かつ serializable である、すなわち、外部整合的 (external consistent) である分散トランザクションを効率的に実現する方式に関して述べる。

[1] 筑波大学
[2] 慶應義塾大学
[a)] horie@hpcs.cs.tsukuba.ac.jp

2. 研究課題

外部整合性を実現する方式には、大別して single master 方式と multi-master 方式の 2 つがある。前者には Google Spanner[1] があり、後者には SLOG[2], OceanVista[3] などがある。後者の実現には、ユーザが発行するトランザクションに deterministic 性が要求されるなど制約が厳しい一方、前者の制約は緩い。本研究では single master 方式を取り上げる。システム構成としては、linearizability を保証するために Raft[4] を用い、serializability を保証するために S2PL[5] を用いる。この構成において性能ボトルネックとなるのは Raft における合意形成処理である。通常の合意形成処理においては、クライアントからの要求を逐次的に実行する必要がある。この場合、データベースに対して並行アクセスができないため、トランザクション処理スループットは比較的低くなる。この逐次実行処理を回避することができれば、性能を向上させられるだろうが、我々の知る限り、それを実現した例は存在しない。

3. 提案

逐次実行処理問題を解決するために、本研究ではログ転送グループ化法 [6] を提案する。この方式ではクライアントから到着する複数のトランザクションを leader ノードにおいてバッチ的に処理することにより、データベースへの並行アクセスを実現する。さらに、leader ノードから follower ノードへのログ転送もグループ化することで通信帯域を有効活用する。ログ転送グループ化を実現するには、S2PL のロッキング範囲を拡大することが必要になる。プロトコルは次のようになる。

(1) クライアントから leader ノードはトランザクションを受け取る
(2) データベースオブジェクトをロックする
(3) データベースオブジェクトを更新する
(4) leader ノードはログを生成し、follower へ転送して合意形成を行う
(5) 合意形成に成功したならば、ロックを解放する

このロック獲得時間中に、Raft の log index が変更されないことが肝要である。

4. 評価

提案システムを C++言語により設計・実装し、5 台のノード (1 台の leader と 4 台の follower) により評価を行った。通信環境には ethernet を用いた。マシンのスペックを表 1 に示す。実験結果を図 1 に示す。一つのトランザクションには 2 つの read-modify-write を含めた。この結果より、クライアント数を増やすことでスループットが増大していることがわかる。ここで注意すべきは、leader ノードは 1 つであり、Spanner のように zone 毎に異なる master を置いている訳ではない点である。

表 1 実験環境

#Nodes	5
OS	CentOS release 6.10
CPU	Intel(R) Xeon(R) CPU E5620 @ 2.40GHz
#Cores	24
RAM	24GB
Network	GigabitEthernet

5. 結論

外部整合性を保証する分散トランザクションシステムを効率的に実現するために、本研究では Raft と S2PL に基づくログ転送グループ化方式を提案した。実験の結果、提案システムはクライアント数に対してスケールすることが観察された。今後の課題はさらなる性能向上である。

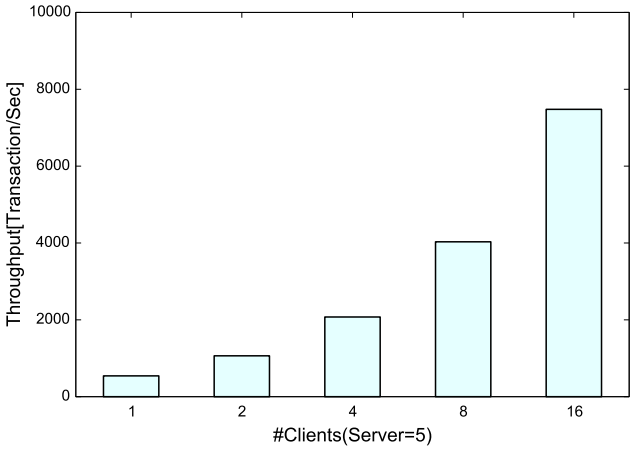

図 1 実験結果:スループット

参考文献

[1] JamesC.Corbett,JeffreyDean,MichaelEpstein,AndrewFikes, ChristopherFrost,JJFurman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Se- bastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford. Spanner: Google's globally-distributed database. In 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12), pp. 261264, Hollywood, CA, 2012. USENIX Association.

[2] Ren, Kun, Dennis Li, and Daniel J. Abadi. "SLOG: serializable, low-latency, geo-replicated transactions." Proceedings of the VLDB Endowment 12.11 (2019): 1747-1761.

[3] Fan, Hua, and Wojciech Golab. "Ocean vista: gossip-based visibility control for speedy geo-distributed transactions." Proceedings of the VLDB Endowment 12.11 (2019): 1471-1484.

[4] Diego Ongaro and John K Ousterhout. In search of an understandable consensus algorithm. In USENIX Annual Technical Conference, pp. 305319, 2014.

[5] Weikum, Gerhard, and Gottfried Vossen. Transactional information systems: theory, algorithms, and the practice of concurrency control and recovery. Elsevier, 2001.

[6] 梶原 顕伍, 川島 英之, 建部 修見. Raft に基づく分散データベースにおけるデータ分割. 研究報告システムソフトウェアとオペレーティング・システム (OS), Vol. 2017, No. 20, pp. 16, 2017.