分散データベースJungleに関する研究

大城信康 (並列信頼研究室)

研究概要

・スマートフォンやタブレット端末の普及によりウェブサービスの利用者が増加し、負荷が高まる
・ウェブサービスにはデータベースが必須であり、負荷に対するためデータベースにはスケーラビリティが求められる
・本研究ではスケーラビリティのある分散CMS用データベースとして非破壊的木構造データベースJungleにデータ分散の実装を行った
・Cassandraとの性能比較を行い、分散環境下においては10倍以上速くなる結果も確認した

スケーラビリティとは

・システムの負荷の増加に対し、汎用的なマシンを追加することで柔軟に拡張して対処できる性質(スケールアウトともいう)

CMS(コンテンツマネジメントシステム)

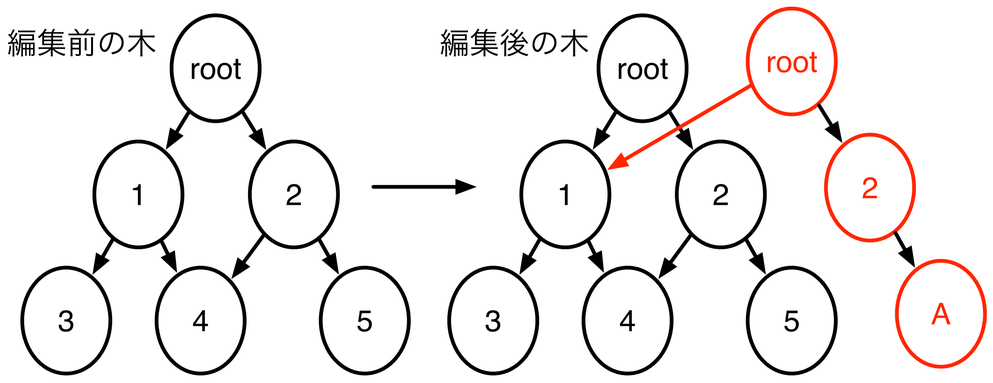
・Webコンテンツを構成するテキストや画像などのデジタルコンテンツを管理し配信するシステム
例：ブログツール、Wiki

分散CMS

・Webコンテンツを分散して管理する能力が必要
・スケーラビリティの性質を持つ
・データの整合性に多少の遅延がある結果整合性でもよい
・読み込みや書き込みを優先するデータベースが求められる

非破壊的木構造

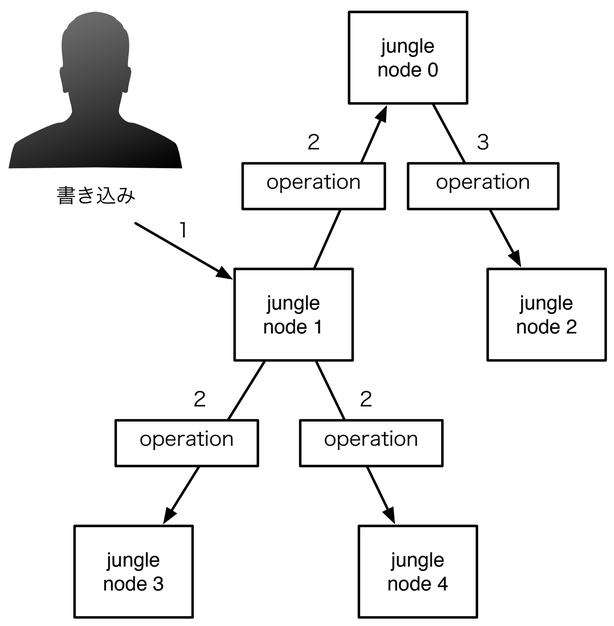
・一度作成したデータは変更しない
・新しい木構造を作成することでデータの編集を行う



通常の木構造と異なり並列に読み書きが可能である
ロックが必要になるのは新しいルートノードを登録するのみのため、
通常の木構造よりロックが少ない

Jungleの分散実装

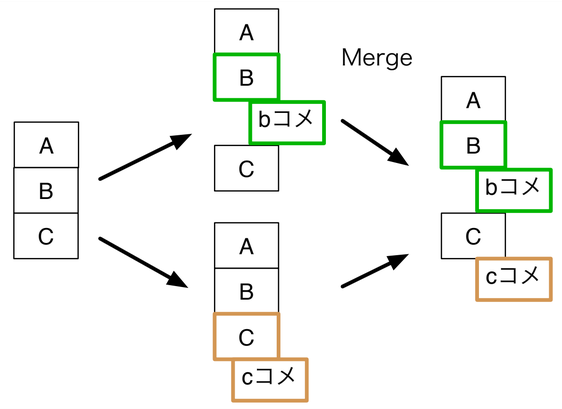
・Jungleの分散実装には当研究室で開発している並列分散フレームワークであるAliceを用いる
・Aliceはネットワークトポロジー形成と、サーバノード間のデータ送受信の機構を提供する
・Aliceを用い、書き込みにより行われたOperationをそのまま他のノードに流すことでデータの分散を行う
・データ更新の衝突に対してはMergeにより解決を測る



ツリートポロジーの形成

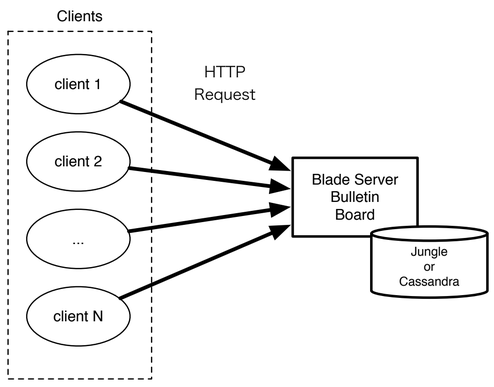
Mergeの実装

・2つの書き込みの結果から1つの書き込みを作る
・掲示板はcommutative(可換)な為、いつ書き込んでも良い
・そのため、Mergeを動的におこなうことができる

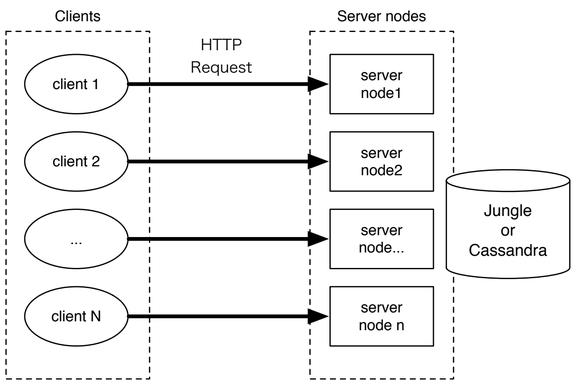


JungleとCassandraの比較

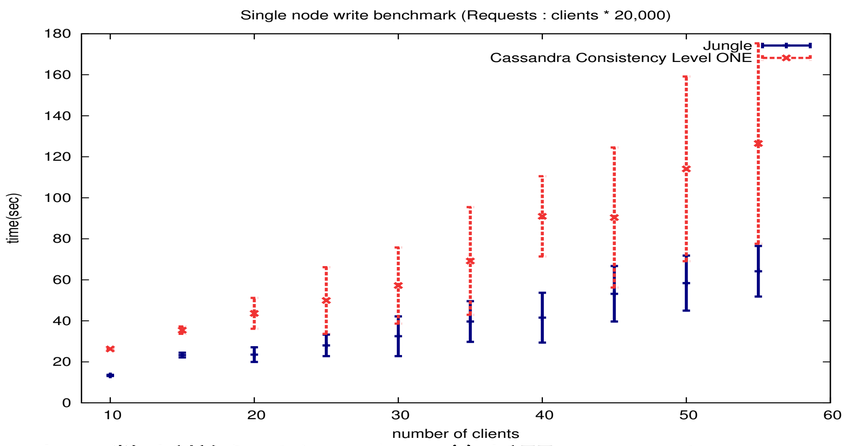
・Jungle、Cassandraを使用した簡易掲示板を作成し、並列に書き込みの負荷をかけ、アクセスの平均時間と標準偏差を測る
実験は以下の2つを行う
・実験1：複数のクライアントから単体のサーバへの負荷
・実験2：複数のクライアントから複数のサーバへの負荷(クライアント数とサーバ数が同じ)



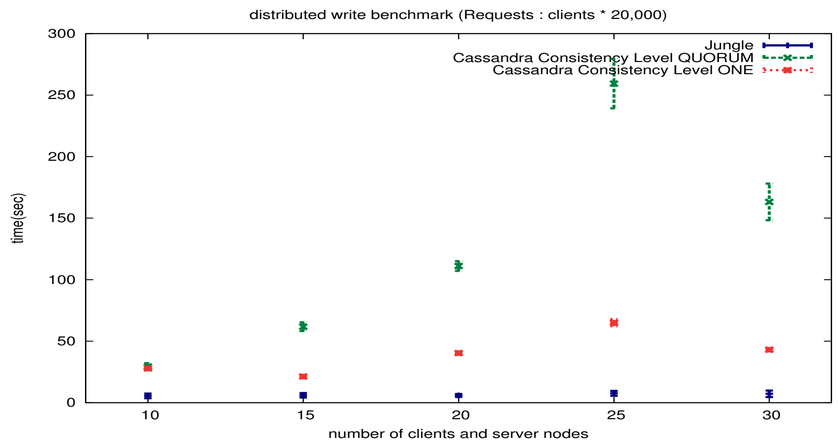
実験1 単体サーバへの負荷



実験2 複数のサーバへの負荷



・クライアントの数が増えるにつれて差が開いている
・平均時間だけをみるとクライアントが55台の時に倍以上Jungleが早い
・これはJungleではロックが少ないことが要因としてあげられる



・Jungleのグラフが横ばいになっていることに注目したい。Jungleはリクエストに対し手元のデータを返すため、サーバノードの数が増えてもレスポンスの早さを維持できる
・ただしJungleは全て非同期でデータの伝搬を行っているため、データ全体の整合性はおちる

今後の課題

・Jungleは多くのメモリを使用するため、ある程度の単位で過去のデータを掃除する必要がある
・アプリケーション毎のMergeアルゴリズムの設計を考えなければならない