

TF の高性能化とデータ一貫性の確保

荻原湧志, 川島英之, 大矢晃久, 萬礼応

概要

Robot Operating System(ROS) はロボットソフトウェア用のミドルウェアソフトウェアプラットフォームであり、近年多くの研究用ロボットで用いられている。TF ライブラリは ROS で頻繁に使用されるパッケージであり、ロボットシステム内の座標変換を追跡し、データを変換する標準的な方法を提供するために設計されたものである。ROS の開発初期には複数の座標変換の管理が開発者共通の悩みの種であると認識されていた。このタスクは複雑なために、開発者がデータに不適切な変換を適用した場合にバグが発生しやすい場所となっていた。また、この問題は異なる座標系同士の変換に関する情報が分散していることが多いことが課題となっていた。そこで、TF ライブラリは各座標系間の変換を有向森構造として管理し、効率的な座標変換情報の登録、座標変換の計算を可能にした。しかしながら、この有向森構造にはデータの暗黙的な線形補間による一貫性の欠落、及び非効率な並行性制御によりアクセスするスレッドが増えるに従ってパフォーマンスが低下するという問題があることがわかった。そこで、我々はデータベースのトランザクション技術における再粒度ロック法、及び並行性制御のアルゴリズムの一種である 2PL を応用することにより、この問題を解決した。提案手法では、スレッド数が 12 までスケールアップすることを示した。また、多くのアクセスパターンにおいて提案手法は既存手法より高いスループットを出すことを示した。

1 序論

1.1 背景

ロボットを使って作業を行う場合、ロボット自身はどこにいるのか、ロボットにはどこにどんなセンサーがついており、また周りの環境のどこにどんなものがあるかをシステムが把握することが重要である。例えば、図 1 のように部屋の中にロボットと、ロボットから観測できる二つの物体があるケースを考える。図中にてロボットは円形、物体は星形で表現される。ロボットが向いている方向は円の中心から円の弧へつながる直線の方で表している。途中で交わる二つの矢印は各座標系の位置と原点、姿勢を表す。ここでは、地図座標系、ロボットの座標系、二つの物体それぞれの座標系が示されている。

システムはロボットに搭載されたセンサーからのデータを元に各座標系間の位置関係を随時更新する。座標系間の位置関係は並行移動成分と回転成分で表現できる。例えば、自己位置推定プログラムは LiDAR から点群データが送られてくるたびにそれを地図データと比較して自己位置を計算し、ロボットが地図座標系にてどの座標に位置するか、ロボットがどの方向を向いているかといった、地図座標系からロボット座標系への位置関係を更新する。物体認識プログラムはカメラからの画像データが送られてくるたびに画像中の物体の位置を計算し、ロボット座標系から物体座標系への位置関係を更新する。

このように、各座標系間の位置関係の更新にはそれぞれ異なるセンサー、プログラムが使われる。各センサーの計測周期、及び各プログラムの制御周期は異なるため、各座標系間の位置関係の更新頻度も異なるものとなる。図 2 では、地図座標系からロボット座標系への位置関係データと、ロボット座標系から物体座標系への位置関係デー

タがそれぞれ異なるタイミングで登録されていることを示している。

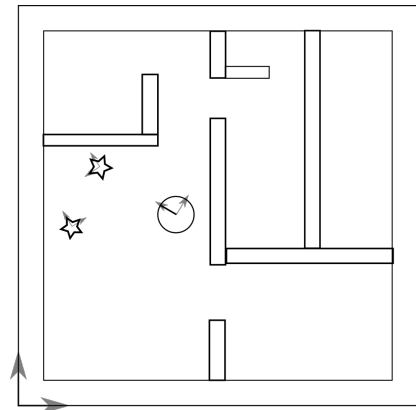


図 1: 部屋の中のロボット

ここで地図中での物体の位置を把握するために、地図座標系から物体座標系への位置関係を取得する方法について考える。地図座標系から物体座標系への位置関係は地図座標系からロボット座標系への変換とロボット座標系から物体座標系への変換を掛け合わせれば計算ができるが、図 2 のように各変換データは異なるタイミングで来るため、最新の変換データを取得するプログラムは複雑なものとなる。A の時刻で地図座標系から物体座標系への変換データを計算しようとするロボット座標系から物体座標系への最新の変換データを取得できるが、地図座標系からロボット座標系への変換データはまだ取得できない。このため、最新の変換データ θ を取得する、もしくは過去のデータを元にデータの補外をすると必要がある。B の時刻で地図座標系から物体座標系への変換データを計算しようとするロボット座標系から物体座標系への最新の変換データを取得できるが、ロボット座標系から物体座標系への変換データはその時間には提供され

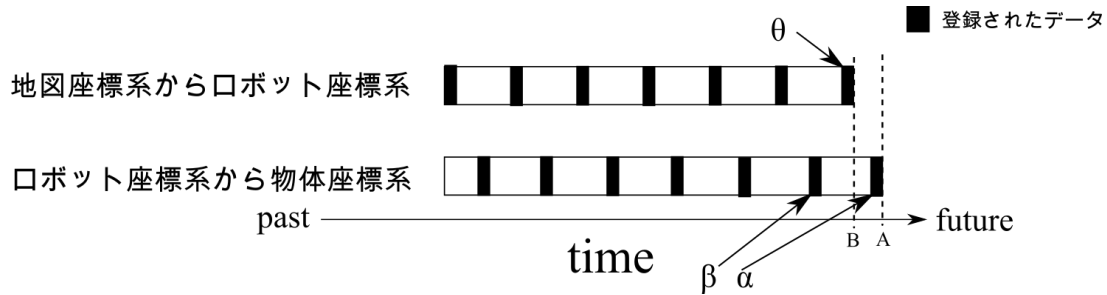


図 2: 位置関係の登録のタイムライン

ていない。このため、 α と β のデータから線形補間を行う、もしくは最新の変換データ β を取得する必要がある。また、地図座標系からロボット座標系への変換とロボット座標系から物体座標系への変換は別のプログラムで管理されており、座標系同士の変換に関する情報が分散している。

このように、ROS の開発初期には複数の座標変換の管理が開発者共通の悩みの種であると認識されていた。このタスクは複雑なために、開発者がデータに不適切な変換を適用した場合にバグが発生しやすい場所となっていた。また、この問題は異なる座標系同士の変換に関する情報が分散していることが多いことが課題となっていた。

そこで、TF ライブラリは各座標系間の変換を有向森構造として一元管理し、効率的な座標系間の変換情報の登録、座標系間の変換の計算を可能にした。まず、図 1 を表す木構造は図 3 で表現できる。木構造のノードが各座標系を表し、木構造のエッジは子ノードから親ノードへの変換データが存在することを表す。

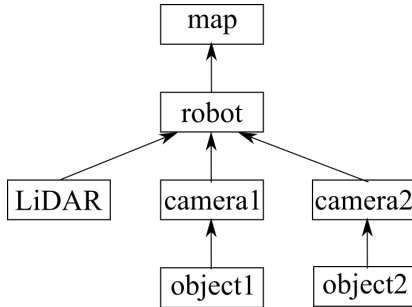


図 3: 図 1 に対応する木構造

各ノードは TF ではフレームと呼ばれ、ノード中の文字列は各座標系に対応するフレーム名が書かれている。図 3 では地図座標系のフレーム名は map、ロボット座標系のフレーム名は robot、物体 1 の座標系のフレーム名は object1 となる。木構造は子ノードから親ノードへポイントが貼られ、子ノードから親ノードを辿ることができる。このため、map から object1 への座標変換を計算するには object1 から map への座標変換の計算し、その逆数を取る必要がある。

子ノードから親ノードへの位置関係は子ノードが保持する。図中では示されていないが、木のルートノードのエッジは NULL ポインタを指している。

図 3 における位置関係登録のタイムラインが図 4 のよう

になっているとき、TF では object1 から map への最新の位置関係は次のように計算する。

まず、object1 から map へのパスを確認する。ここでは object1 から map へのパスは $object1 \rightarrow camera1$, $camera1 \rightarrow robot$, $robot \rightarrow map$ であることがわかる。

次に、全てのパスにおいて座標変換が提供できる時間で、最も最新の時間を取得する。図 4 を確認すると、 $object1 \rightarrow camera1$, $camera1 \rightarrow robot$, $robot \rightarrow map$ において最新のデータが最も古いのは $robot \rightarrow map$ である。このため、時刻 A がここでは要件を満たす。

最後に、時刻 A での各パスのデータを取得し、それらを掛け合わせる。 $robot \rightarrow map$ については登録されたデータを使えるが、 $object1 \rightarrow camera1$, $camera1 \rightarrow robot$ は時刻 A におけるデータは存在しない。このため、TF では時刻 A の前後のデータを元に時刻 A のデータを線形補間によって構築する。

このように TF はシステム内部の座標系間の位置関係を一元管理する機構を提供する。しかしながら、これには以下のような問題点が挙げられる。

問題 1：ジャイアント・ロック

まず、TF の森構造には複数のスレッドがアクセスするため並行性制御が必要となるが、一つのスレッドが森構造にアクセスしている際は他のスレッドは森構造にアクセスできないアルゴリズムとなっている。これは、マルチコアが常識となっている現代では大きな問題となる。

問題 2：データの一貫性

次に、TF が提供するインターフェイスでは座標変換の計算の際に最新のデータを使わず、線形補間されたデータを用いてしまう。図 4 において object1 から robot の座標変換を計算する際には時刻 A の $object \rightarrow$ のデータを使うが、ここでは線形補間によってデータが暗黙的に生成されており、さらに線形補間の際にも最新のデータは参照されていない。これは次のようなケースで問題となる。

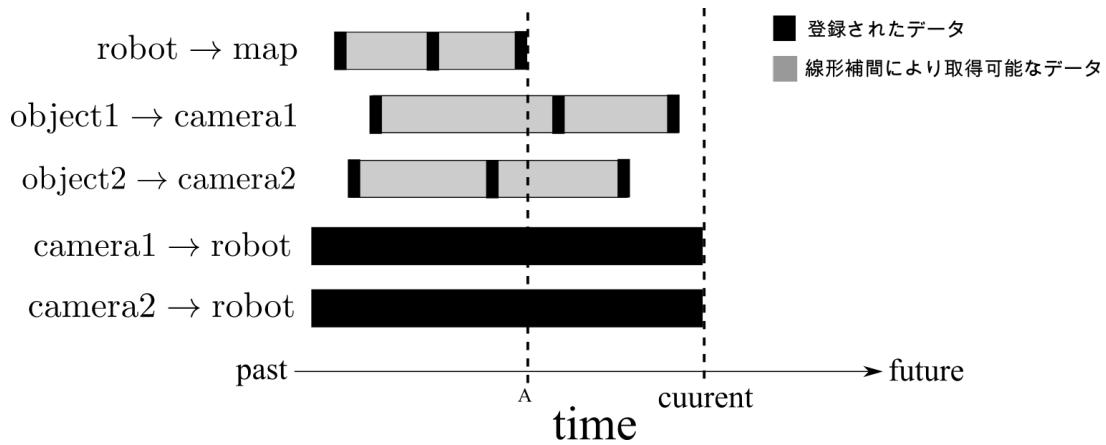


図 4: 図 3における位置関係登録のタイムライン

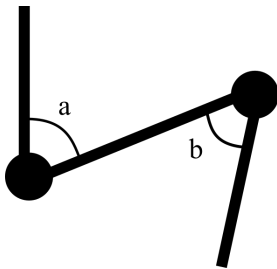


図 5: 関節 a と b の角度の合計が一定のアーム

図 5はロボットのアームを示しており、このアームは常に各関節の角度 a と b の和が一定値になるように制御される。このため、TF の森構造にこのアームの情報が登録され、アームの各関節の角度を読み取る制御プログラムは角度 a と b の和が一定値であることを想定する。TF によるデータの暗黙的な線形補間でこの一貫性が失われる可能性があり、制御プログラムは意図しないデータを読み込むことになり、例外処理の発生につながる。

このように高精度の制御が求められる場面では現状の TF の並行性制御アルゴリズム、及びインターフェイスには問題がある。

1.2 貢献

本研究ではデータベースのトランザクション技術における再粒度ロック法、及び並行性制アルゴリズムの一種である 2PL を応用することにより、問題 1 および問題 2 を解決した。

1.3 論文構成

本論文の構成は次の通りである。2 章では関連研究について述べる。3 章では既存の TF の森構造とその問題点について述べる。4 章では提案手法である森構造への再粒度ロックの導入とデータ一貫性のためのインターフェイスの提供について述べる。5 章では提案手法の評価結果を述べる。6 章では本研究の結論を述べる。7 章では今後の課題について述べる。

2 関連研究

TF のようにデータを時系列的に管理するものとして SSM が挙げられる。

データベースの技術をロボットに適用するという内容では GAIA[5] が挙げられる。ロボットというより自律システム

C++に宣言的にデータ変更時のルールを記述できる。これによって簡単にイベントベース trigger 付きの UDF みたいな？

これは RDB ベースで reactive な挙動を提供する。

プロダクトレベルのものを目指す ROS2[6] や Autoware[7] でも、liveliness などの指標 DDS が導入されたが、データベースの並行性制御の導入はない。

本研究のようなアプローチは存在しない。

2PL 以外の並行性制御アルゴリズムとしては Silo[8] が挙げられる。

3 既存の TF 森の構造とその問題点

TF 森の実態は tf2 パッケージ中にある BufferCore クラス [4] である。

3.1 構造

まずは TF 森とタイムテーブルから？TF は構造上各座標系同士の回転移動、並行移動で表現できる位置関係は TF 森で表現される

例えば図のようなマップ座標系、ロボット座標系、カメラ座標系、物体の座標系は図のような木構造に対応する。

また、図のフレーム A, B のように他の木とは分離された木が存在してもよい。このため、TF 森構造となる。

さて、座標変換の情報は変わらない場合と時刻とともに変わる場合がある

さて、座標変換の情報は主にセンサーからデータが送られてくるたびに計算され、TF 森に登録される。このため各フレーム間の座標変換の情報はセンサーの更新頻度

に依存し、それぞれ異なるタイミングで更新される。例えば、マップ座標系からロボット座標系の座標変換を計算するプログラムは LiDAR のデータ更新頻度に合わせて自己位置を計算し、マップ座標系からロボット座標系の座標変換を登録する。カメラ座標系から物体の座標系の座標変換を計算するプログラムはビデオカメラのデータ更新頻度に合わせて物体の位置を計算し、カメラ座標系から物体の座標系の座標変換を登録する。カメラの更新頻度と LiDAR の更新頻度が異なる場合、マップ座標系からロボット座標系の座標変換が登録されるタイミング、カメラ座標系から物体の座標系の座標変換が登録されるタイミングにズレが生じる。マップ座標系から物体の座標系を計算する際にデータ同期の問題が生じてしまう。これに対処するため、まず TF 森は各フレーム間の座標変換情報を過去一定期間保存する。これは図のように表現できる

図は各フレーム間の座標変換情報が提供される時刻を表す。横軸が時間軸を表し、左側が過去、右側が最新の時刻を表す。

灰色線は対応する時刻の座標変換情報が登録されたことを表す。図の点線の位置での map から apple への座標変換を計算する際、各フレーム間の座標変換情報が必要になる。robot から camera への座標変換情報は常にある物として扱われるが、map から robot、camera から apple の座標変換データは該当する時刻では登録されていない。そのため、TF は前後のデータから線形補間を行うことにより座標変換データを計算する。つまり、TF は該当する時刻の座標変換データが保存されている、もしくは前後の値を元に線形補間ができる場合にはその時刻の座標変換データを提供できる、とみなす。

まずは CompactFrameID 等の typedef は説明が面倒なので、展開してしまうか。

座標変換は TransformStorage で表現され、これは以下で構成される回転成分を表す rotation_。これは Quaternion で表現される並行移動成分を表す translation_。これは Vector3 で表現される座標変換の時刻を表す stamp。これは ros::Time 型で表現される座標変換の親フレームの frame id を表す frame_id。これは整数型で表現される座標変換の子フレームの frame id を表す child_frame_id。これは整数型で表現される

TimeCacheInterface は特定のフレームから親フレームへの座標変換を登録・管理する抽象クラスである。主に以下のようなメソッドがある bool getData(ros::Time time, TransformStorage & data_out, std::string * error_str = 0)

特定の時刻 time における TransformStorage を取得する。取得できた場合には true を返し、取得できなかった場合には false を返しエラーの内容を error_str に代入する。

bool insertData(const TransformStorage& new_data)

座標変換の情報を登録する。情報の登録に成功した場合には true を、そうでない場合には false を返す。

```
std::pair<ros::Time, CompactFrameID> getLatestTimeAndParent()
```

最新の座標変換情報のうちその時刻と、親フレームの id を返す。

TimeCache は TimeCacheInterface を継承し、あるフレームにおける親フレームへの座標変換を時系列的に管理する。これは TransformStorage の双方向キューで構成され、ユーザーが座標変換情報を登録する際には先頭に push し、保存していた座標変換の時刻が 10 秒以上過去のものとなれば deque から追い出す。getData のアルゴリズムは以下ようになる

StaticCache は TimeCacheInterface を継承し、先程の例の robot-_isensor 間の座標変換のように不変な座標変換情報を管理する。

BufferCore は TF 森を管理する。

文字列である frame は TF 森内部では整数型の id で管理している

TimeCacheInterface を継承する TimeCache と StaticCache はそれぞれあるフレームからみた親フレームの id、親フレームの座標変換を保持するため、これらは TF 森における子フレームから親フレームへのエッジ及び子フレームのノードを表す。このため TF 森は子フレームから親フレームへエッジが伸びる有効森構造となっている。

以下のようなインターフェイスがある。bool setTransform(const geometry_msgs::TransformStamped& transform, const std::string & authority, bool is_static = false)

座標変換情報を登録する。authority には座標変換情報を登録するプログラムの名前等を指定できる。is_static を true にすると、StaticCache にデータを登録できる。下記の更新、挿入において詳細を説明する

```
geometry_msgs::TransformStamped lookupTransform(const std::string& target_frame, const std::string& source_frame, const ros::Time& time) const;
```

時刻 time における target_frame から source_frame への座標変換を返す。下記の検索の節において詳細を説明する

```
template<typename F> int walkToTopParent(F& f, ros::Time time, CompactFrameID target_id, CompactFrameID source_id, std::string* error_string) const;
```

指定の時刻 time について target_id から source_id への座標変換を計算する。返り値は 0 以外の場合にはエラーコードとなり、エラーの原因が error_string に代入される。

テンプレートパラメータの F にはよく TransformAccum クラスが与えられ、これは通過したパスの座標変換情報を蓄積するオブジェクトである。

主に lookupTransform の中で用いられる。

```
int getLatestCommonTime(CompactFrameID target_frame, CompactFrameID source_frame, ros::Time& time, std::string* error_string) const;
```

target_frame から source_frame までのパスにおいて、座標変換のデータが提供できる時刻で最新のものを取得する。そのような時刻が存在する場合には time に代入され 0 を返す。もしそのような時刻がない場合にはエラーコードを返し、エラーの内容を error_string に代入する。

これは図のように説明できる。

図において map から robot へのパスにおいて必要なのは map から robot への座標変換、robot から sensor への座標変換、sensor から apple への座標変換である。この中で最新のデータが最も古い時刻に来ているのは map から robot への座標変換である。この時刻よりも前ならば map から apple への座標変換データは提供できるが、この時刻より後ではパス中の座標変換が提供できなくなる。このため、getLatestCommonTime ではこの時刻 A が選択される。

ここで、図中では lidar から object への座標変換が最も古い時刻に来ているが、map から robot への座標変換を計算する時にこの座標変換は使わないために時刻 D が選ばれないことに注意する。

(robot -i lidar -i とか別のもっと遅いものも追加！)

```
TimeCacheInterfacePtr BufferCore::getFrame(CompactFrameID
frame_id)
```

与えられた frame_id に対応する TimeCacheInterface へのポインタを取得する。該当するものがない場合には null を返す。

また、BufferCore には以下のようなフィールドがある

frames_: 各フレーム id に対応する TimeCacheInterface のポインタを管理する配列。TF 森はこの配列で表現される。

frame から frame id を検索するテーブル、及び frame id から frame を検索するテーブルが存在する frameIDs_, frameIDs_reverse

frame_mutex_: frames_, frameIDs_, frameIDs_reverse は複数のスレッドから操作される。データ競合を防ぐため、各スレッドは frame_mutex_ を用いて排他処理を行う。

特定のフレームに対応する TimeCacheInterface を取得するには、まず frameIDs_ から対応する frame id を取得し、続いて frames_ から対応する TimeCacheInterface へのポインタを取得する。この二回の検索処理の後、TF 森のフレームに該当する TimeCacheInterface にアクセスできる。

TransformAccum クラスは、lookupTransform にて木構造を辿って座標変換を計算する際に座標変換の計算を蓄積していくクラスである。

```
int gather(TimeCacheInterface* cache, Time time,
string * error_string)
```

cache から time の時の座標変換を取得し、それを TransformAccum インスタンス内の一時変数に記録する。cache から親のフレームの id を取得しそれを返す。

3.2 lookupTransform

擬似コード

Algorithm 1 lookupTransform

```
e = walkToTopParent
```

```
if  $n < 0$  then
```

```
     $X \leftarrow 1/x$ 
```

```
     $N \leftarrow -n$ 
```

```
end if
```

まずは running example を示す。なるべく C++ レベルでの説明は避けるべきか。とりあえず、途中で親が変わらないと仮定する。じゃないと説明が面倒

1. source から root への座標変換を計算 2. target から root への座標変換を計算 3. source から root への座標変換 * root から target への座標変換を計算

まずは普通のケースから。同じ親を持ち、まずはこの時間で見えていくまずは d から r への座標変換を計算次に e から r への座標変換を計算 r から e への座標変換は、単に e から r への逆変換をとれば良い。あとは d から r への変換 * r から e への変換で計算できる

このように target が直接の

ここら辺は説明省いても良い????

この例のように、source と target が同じ木の中にある場合、つまり同じ root を共有しない場合にはエラーとなる。

running example が十分であれば、擬似アルゴリズムでの説明を省けるかも。

Algorithm 2 walkToTopParent(time, source_id, target_id)

```
frame_id = source_id
top_parent_id = frame_id
// source frame から root へのパスをたどる
while frame_id ≠ 0 do
    cache = getFrame(frame_id)
    if cache = NULL then
        // 木構造の root に到達
        top_parent_id = frame_id
        break
    end if
    parent_id = cache から座標変換と親の id を取得
    if frame_id == target_id then
        // target frame は source frame の祖先なので早期リターン
        累積したデータから座標変換を計算
        return 0
    end if
    座標変換を蓄積
    top_parent_id = frame_id
    frame_id = parent_id
end while
// target_id から root へのパスをたどる
frame_id = target_id
while frame_id ≠ top_parent_id do
    cache = getFrame(frame_id)
    if cache = NULL then
        // 木構造の root に到達
        break
    end if
    parent_id = cache から座標変換と親の id を取得
    if frame_id == source_id then
        // source frame は target frame の祖先なので早期リターン
        累積したデータから座標変換を計算
        return 0
    end if
    座標変換を蓄積
    frame_id = parent_id
end while
if frame_id ≠ top_parent_id then
    // source frame と target frame は同じ木構造に属していない
    return エラーコード
end if
// source frame と target frame は祖先関係にはないが、同じ木に属している
source frame から target frame への座標変換の計算
return 0
```

3.3 setTransform

座標変換を登録する。特に説明の必要はない

3.4 問題点

ここより上の部分では、どの程度エラーケースを、どの程度正確な説明が必要か不明瞭。とりあえず以下を書き下す。

giant lock

最新のデータを返さない example を示そうこの場合には、b に引っ張られて a の最新のデータが取れない

線形補間により、一貫性がないケースがある。例えば関節間の制約やジンバルロック、特異点など制御プログラムが意図しない状態を見てしまうかもしれない

4 提案手法

まず、ginat lock をしない方法を導入読み書きを行う際には必要なノードのみ lock する

setTransform とは異なり、lookupTransform は各ノードの情報の読み込みのみ行う。

読み込み処理であれば複数のスレッドからの同時アクセスを行っても data race は発生しない。このため、

S lock は data race が発生しない read の時に、X lock は二つ以上のスレッドから write 操作されると data race が発生するのを防ぐために使われる。

(S-X はもう古い？ read と write の方が説明が適切？)

さらに、shared lock と exclusive lock を導入するこれは次のような表で表現できる表は列が既にかかっているロック (N は何もかかかっていない)、行はかけようとしているロックを表し、o であれば新たにかけようとしているロックの確保に成功し、x であればロックがかけられないことを表す例えば、S lock が既にかかられていても新たに S lock を他のスレッドが書けることが可能になる。対し、X lock が既にかかっている場合には S lock はかけられず、S lock がかかっている場合にも X lock はかけられない。X lock をかけることができるスレッドは常に一つだけである。

続いて、最新のデータを取らない問題と線形補間によりデータの一貫性がなくなる問題について

既存の lookupTransform では、最新のデータを取得しようとしても過去のデータを参照してしまう、また線形補間されてしまう

最新のデータをそれぞれ取ってくるという方法もサポートする。(しかしこれだけでは十分ではない。j- この導入はいまいち。) 冒頭で説明したように、ROS は分散アーキテクチャを採用する。TFMessage には各フレーム間の座標変換情報を複数登録できるが、TF は複数の座標変換を setTransform を複数回呼び出すことにより実現している。これにより、中間の状態を見てしまう。これは図の

ように説明できる。この図のように、giant lock を毎回の setTransform で取ってはいるが、その操作が終わり次の setTransforms を呼ぶまでの間は lock が外される。これにより、一貫性のない状態を見てしまう恐れがある。

そこで、我々は最新のデータを atomic に取得する lookupLatestTransform、及び複数の座標変換を一度に atomic に登録できる seTransforms を追加した。複数のデータに対する読み込み、書き込みを atomic に行うために、我々は 2PL[?] を実装した。

しかしながら 2PL には dead lock の問題がある。例えば次の例、これは木を登る方向と下る方向の両方があるからこうなる。

データを reorder できれば 2PL では deadlock は発生しない (DAG が構築できる)、が TF 木ではできそうにない。

そこで我々は deadlock prevent の方法として No-wait[Bern 1981] を採用した。これは write lock をかけようとして失敗したら最初からやり直す。

Wound wait, nonpreemptive (Concurrency Control in Distributed Database Systems PHILIP.A BERNSTEIN AND NATHAN GOODMAN P196)

transaction に priority を足すとかは？

5 評価

実験を行う

スレッド数を上げるとスループット伸びる joint を増やすと緩やかにスループットが落ちる iter を増やすと 1000 以降で急激にスループットが下がる。no wait による弊害か？lock の確保に失敗しているのかも read ratio は高くなるほどブロックされる可能性が減る read len は安定しているように見える。なぜ trn で read len が小さいところなるのか。。。2PL でロックする箇所が変わるから？

write len を上げるとやはりブロックされる部分が増える

6 結論

再粒度ロックの導入により、パフォーマンスを上げることができた特に、スレッド数の増加とともにスループッ

トが下がる問題を解決できた

データの一貫性は必須、というデータも示したい。

7 今後の課題

ここでは取り上げなかった TF 木の問題点として、部分的に座標情報の登録に失敗するケース。rollback など maybe transactional に行う必要

また、insert/delete が大量に発生するケースも考える。

参考文献

- [1] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, “Ros: an open-source robot operating system,” in ICRA Workshop on Open Source Software, 2009.
- [2] T. Foote, “tf: The transform library,” 2013 IEEE Conference on Technologies for Practical Robot Applications (TePRA), 2013, pp. 1-6, doi: 10.1109/TePRA.2013.6556373.
- [3] Philip A. Bernstein and Nathan Goodman, “Concurrency Control in Distributed Database Systems” in ACM Computing Surveys, 1981, pp. 185-221
- [4] “BufferCore.h”, https://github.com/ros/geometry2/blob/noetic-devel/tf2/include/tf2/buffer_core.h
- [5] “GAIA platform”, <https://www.gaiaplatform.io>
- [6] “ROS2”, <https://docs.ros.org/en/rolling/>
- [7] “Autoware”, <https://tier4.jp/en/autoware/>
- [8] Stephen Tu, Wenting Zheng, Eddie Kohler †, Barbara Liskov, and Samuel Madden. Speedy transactions in multicore in-memory databases. In SOSP, pages 18–32. ACM, 2013