HTAP技術：
トランザクション処理と
解析処理の壁を越えて

横田　治夫
東京工業大学　情報理工学院

２０２０年１０月２７日　劔ユーザ会

データベース処理の大別

- OLTP（On Line Transaction Processing）
  - 同時に多数で頻繁な更新を含むデータベース処理
    - 例：銀行ATM、コンビニPOS、座席予約システム、課金システム等々
    - 特徴：コンシステンシィとレスポンスタイムの制約
      - 基幹系の更新時に異常（Anomaly）が入ることは許されない
      - 端末の前の顧客を待たせることは許されない
- OLAP（On Line Analytical Processing）
  - 蓄積されたデータを解析するようなデータベース処理
    - 例：期間別売上集計、商品別販売動向調査、知識（ルール）発見等々
    - 特徴：大量のデータに効率よくアクセスする必要
      - データの量が意思決定に影響

古くて新しい話

- OLTPのデータをどうやってOLAPに反映させるか
  - 昔からの課題：基幹系と情報系の融合
    - 基幹系と情報系と呼ぶことが多かった
  - 反映の遅延は諦めていた
    - 週末、夜間にバッチ処理で、OLTP からOLAP用データを抽出
      - 先週のデータ、昨日のデータで解析
      - 古くは磁気テープに吸い出して違うシステムに移動
  - 要求：できるだけ素早く反映させた解析を行いたい！
- HTAP（Hybrid Transactional/Analytical Processing）
  - 近年、多くの国際会議で話題に
  - トランザクション処理と解析処理の壁を乗り越える
    - 背景：記憶デバイス、ネットワーク環境の進歩、メニーコア環境

最近の論文例（一部）

[1] A. Raza, P. Chrysogelos, A. C. Anadiotis, and A. Ailamaki. “Adaptive HTAP through Elastic Resource Scheduling”. In Proc. of SIGMOD'20, pp. 2043–2054, (2020)

[2] D. Huang, Q. Liu, Q. Cui, Z. Fang, X. Ma, F. Xu,…, W. Wei, “TiDB: a Raft-based HTAP database”. PVLDB, *13*(12), 3072-3084, (2020)

[3] D. Butterstein, D. Martin, K. Stolze, F. Beier, J. Zhong, Li. Wang, “Replication at the Speed of Change - a Fast, Scalable Replication Solution for Near Real-Time HTAP Processing” PVLDB, 13(12): 3245-3257 (2020)

[4] J. Yang, I. Rae, J. Xu, J. Shute, Z. Yuan, K. Lau, Q. Zeng, X. Zhao, J. Ma, Z. Chen, Y. Gao, Q. Dong, J. Zhou, J. Wood, G. Graefe, J. F. Naughton, J. Cieslewicz, “F1 Lightning: HTAP as a Service” PVLDB, 13(12): 3313-3325 (2020)

[5] C. Riegger, T. Vinçon, R. Gottstein, I. Petrov, “MV-PBT: Multi-Version Indexing for Large Datasets and HTAP Workloads”, in Proc of EDBT’20, pp. 217-228 (2020)

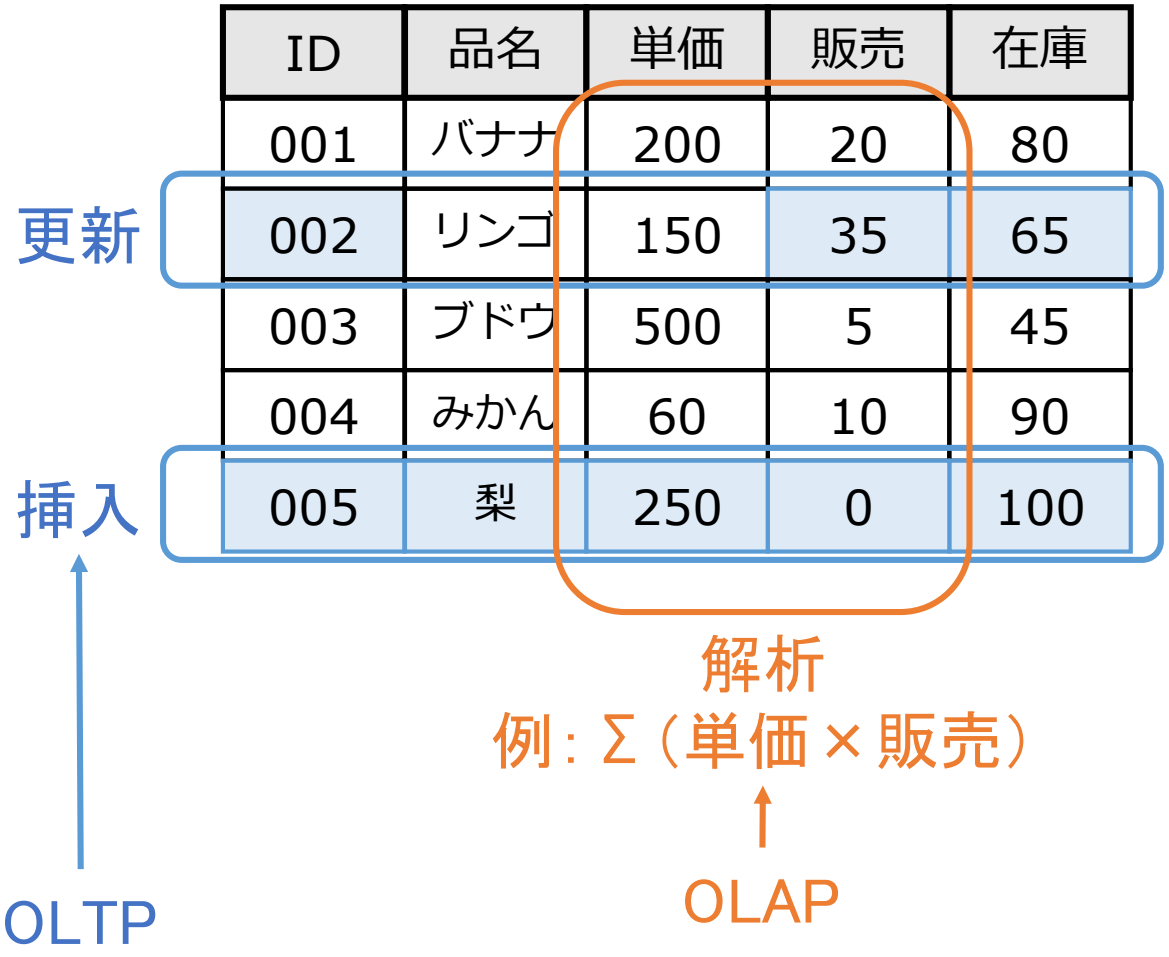
[6] C. Luo, P. Tözün, Y. Tian, R. Barber, V. Raman, R. Sidle, “Umzi: Unified Multi-Zone Indexing for Large-Scale HTAP” In Proc. of EDBT’19, pp. 1-12, (2019)

[7] M. Boissier, R. Schlosser, M. Uflacker, “Hybrid data layouts for tiered HTAP databases with pareto-optimal data placements” In Proc. of ICDE’18, pp. 209-220, (2018)

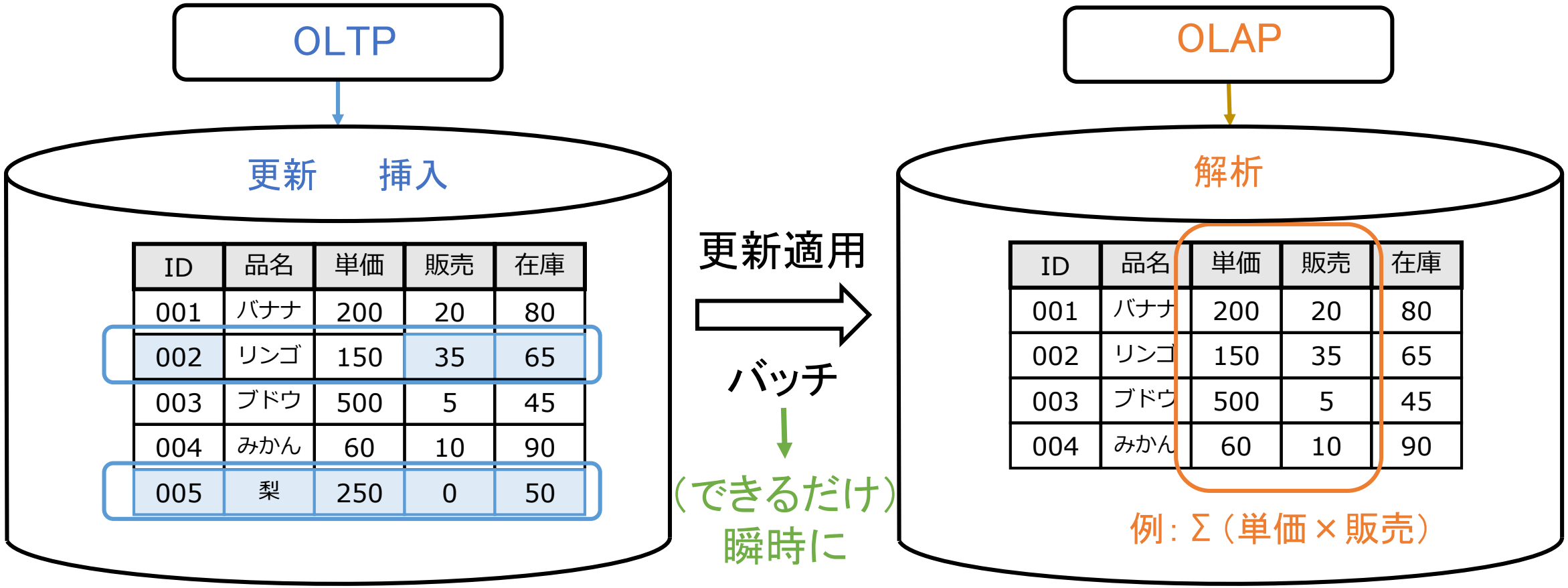
＊HTAPがタイトルに含まれているもの

そもそもなぜ壁があるのか

- 処理対象の違い
  - OLTP: 行単位
    - 一つの行中で複数個所を検索更新
    - 行の挿入
  - OLAP: 列（属性）単位
    - 全部の行の特定の属性を集計
- 同時実行制御上の問題
  - OLTP:
    - 小さい単位で短時間排他
  - OLAP:
    - 大きな単位で長時間共有



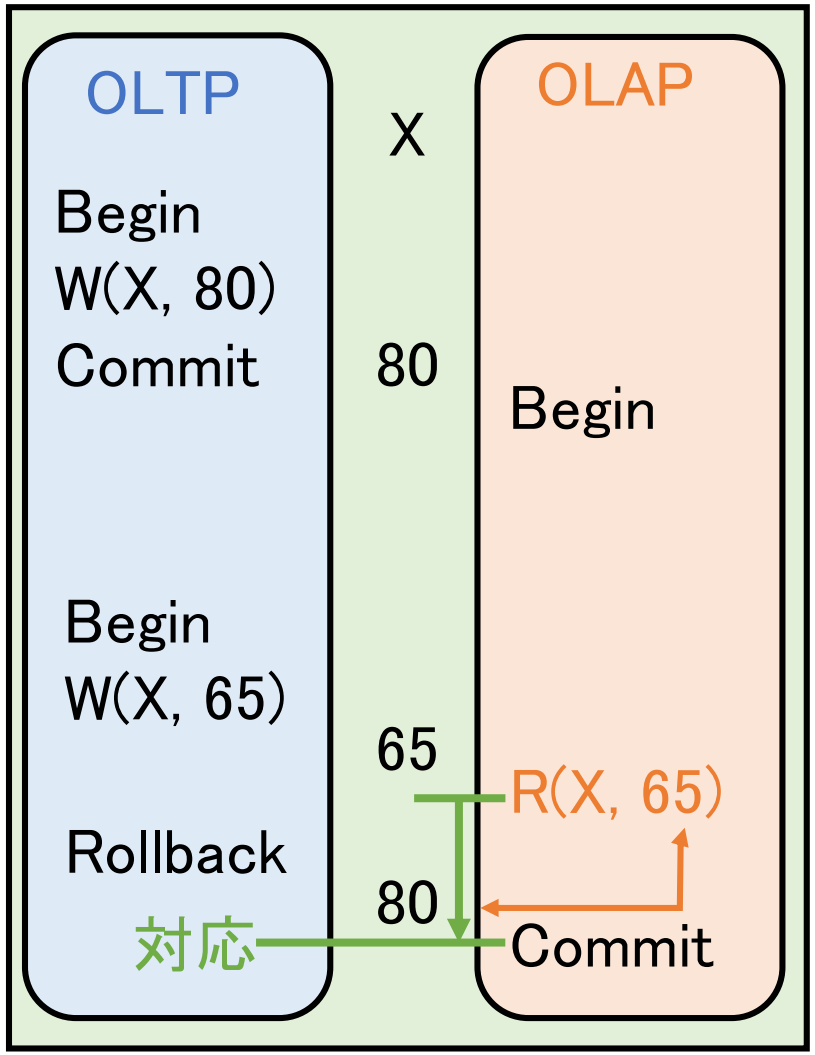
更新適用



- OLTPとOLAPを分けると、それぞれを高速に処理可能
  - 更新を適用：従来は週末、夜間　→　早期に解析結果が知りたい（瞬時？）

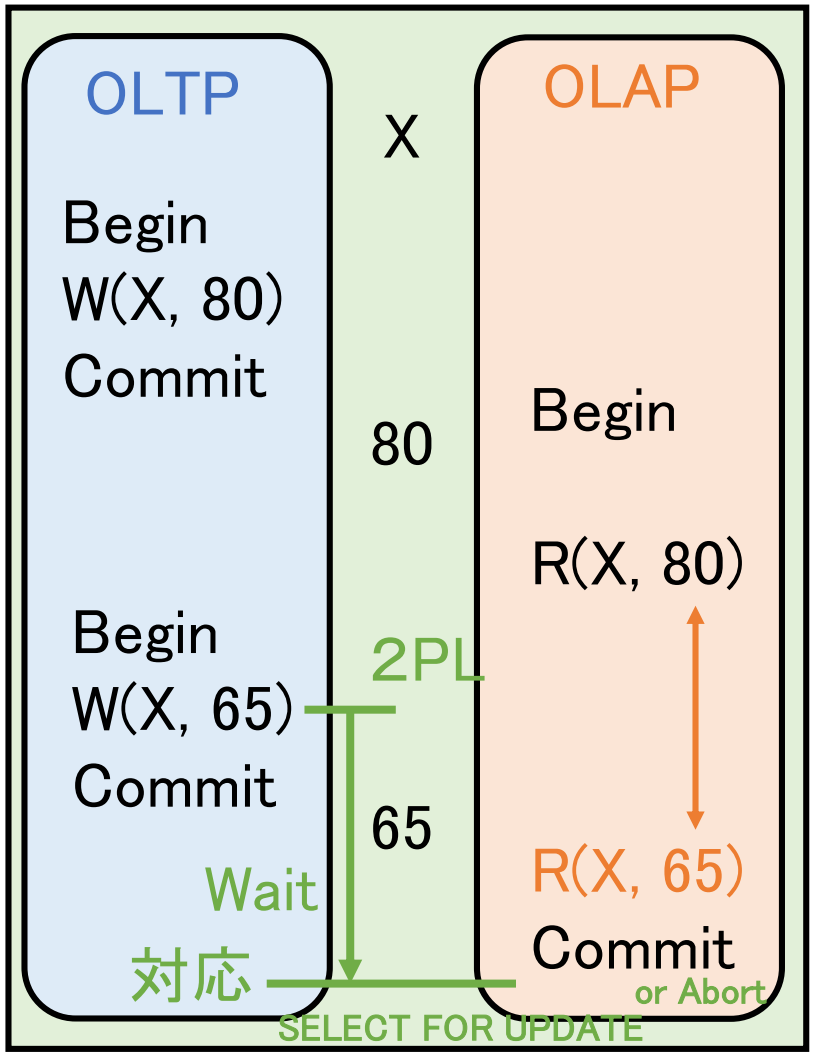
同時実行制御の問題

OLTPがOLAPで待たされるorアボート



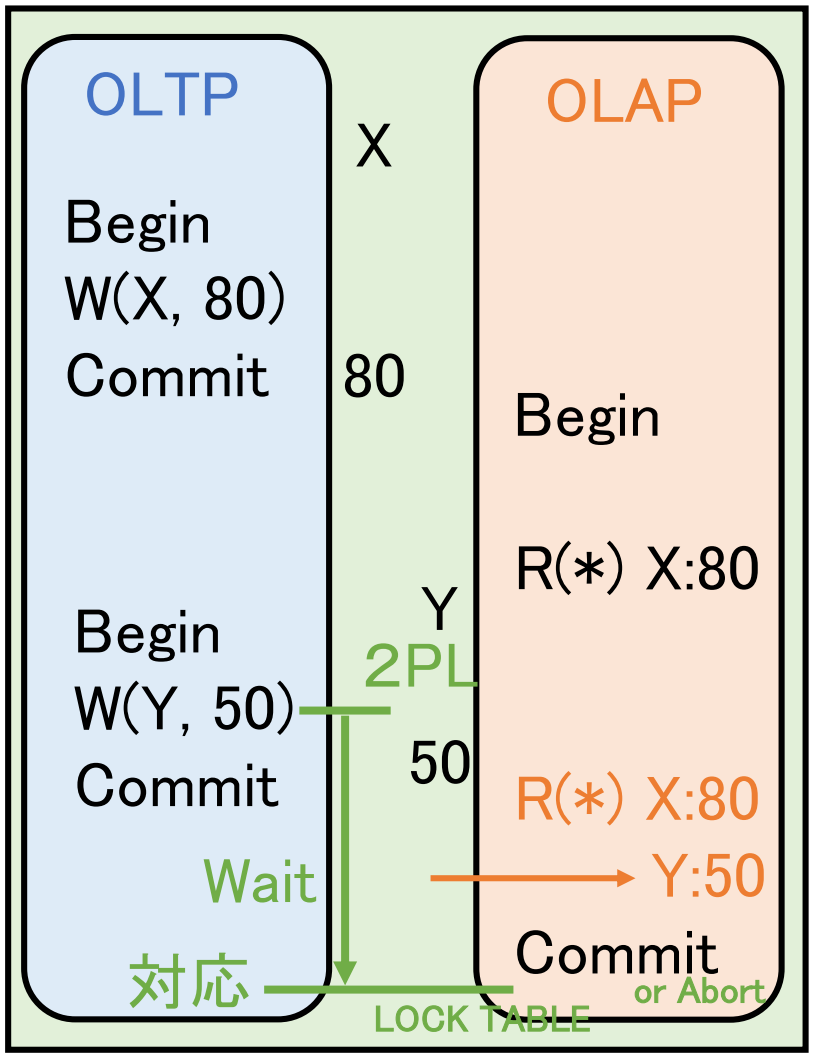
Read Uncommitted

Dirty Read



Read Committed

Non-Repeatable (Fuzzy) Read



Repeatable Read

Giant Lock

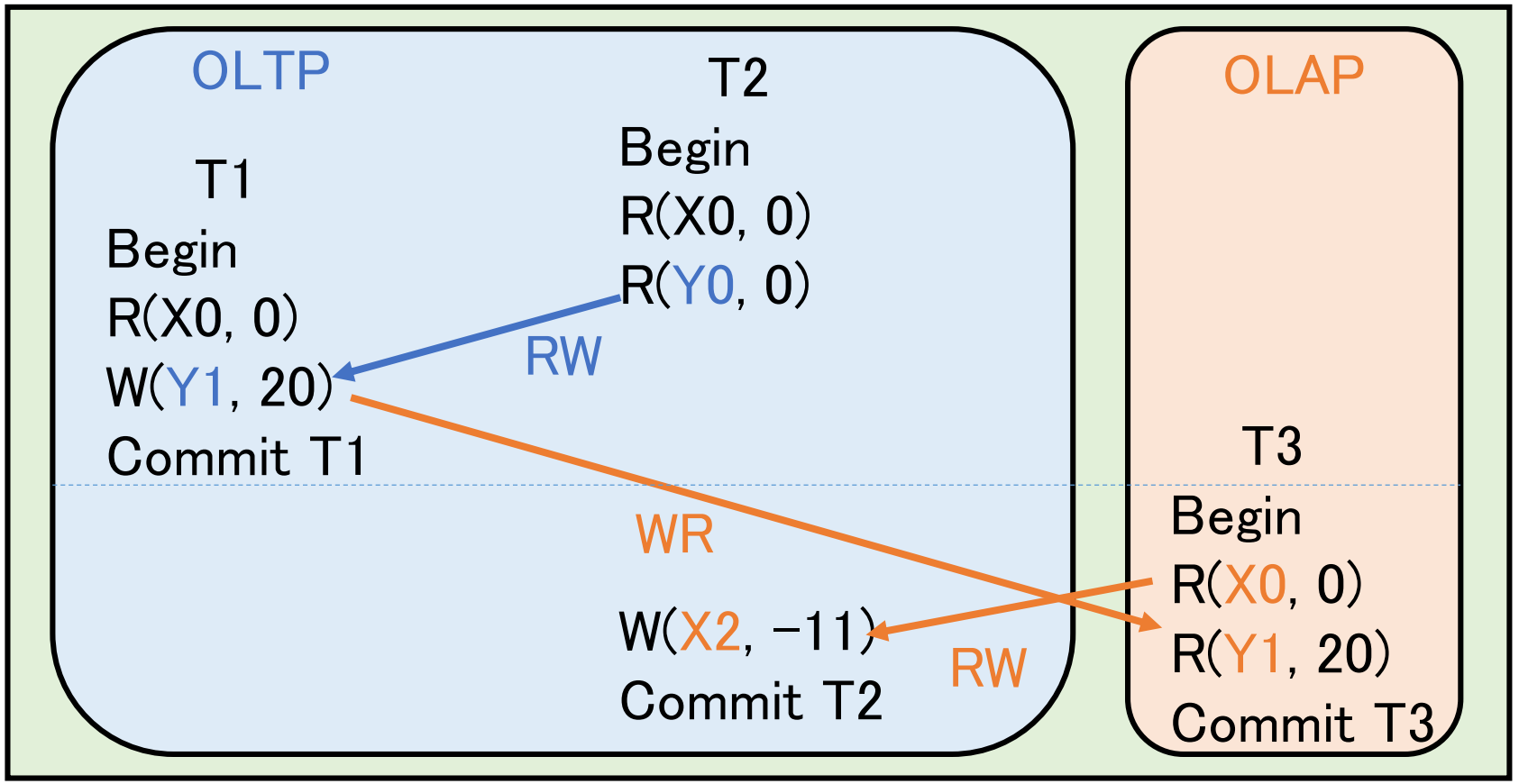
Phantom Read

Snapshot Isolation

- トランザクション開始時に必要なデータセットのスナップショット（バージョン）を処理の対象とする
  - 元データが変更されても、その前のバージョンが対象
    - 開始前にCommit されたデータセット
  - non-repeated read も phantom read も、対象のデータセットが変わらないので発生しない
- Serializable にならない（Write Skew Anomaly）
  - 同じスナップショットを使った条件付きの複数の更新が成立した後で条件が崩れるような場合
    - 条件X+Y>0の下で、X=70、Y=80の時、T1:X-100、T2:Y-100
  - RW-Dependency が連続して2つ以上あったらアボート：SSI

Read Only (Skew) Anomaly

- Snapshot を読むだけで発生する Anomaly
  - Commit したデータセットを読んでも並行して走っているトランザクションからの依存関係にループが発生



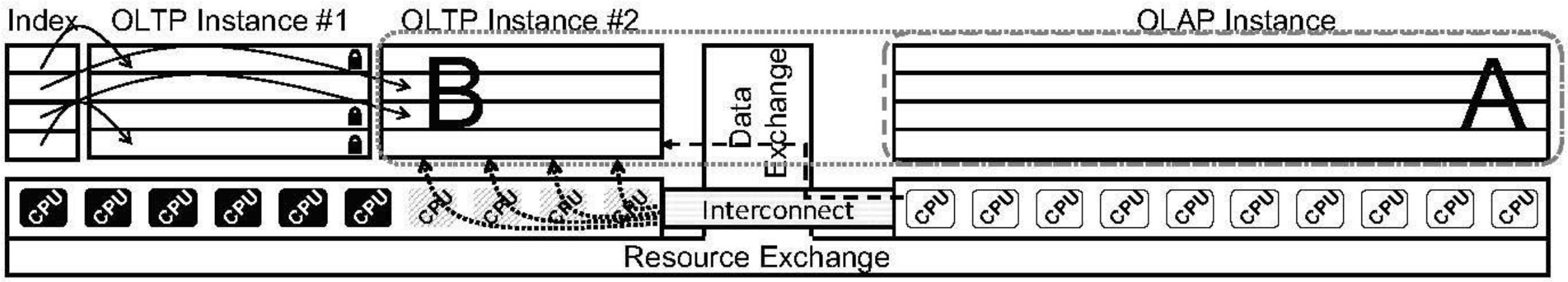
依存関係ループ
T2-T1-T3-T2

OLTPから
OLAPは見えない

本プロジェクト：
解決の目途

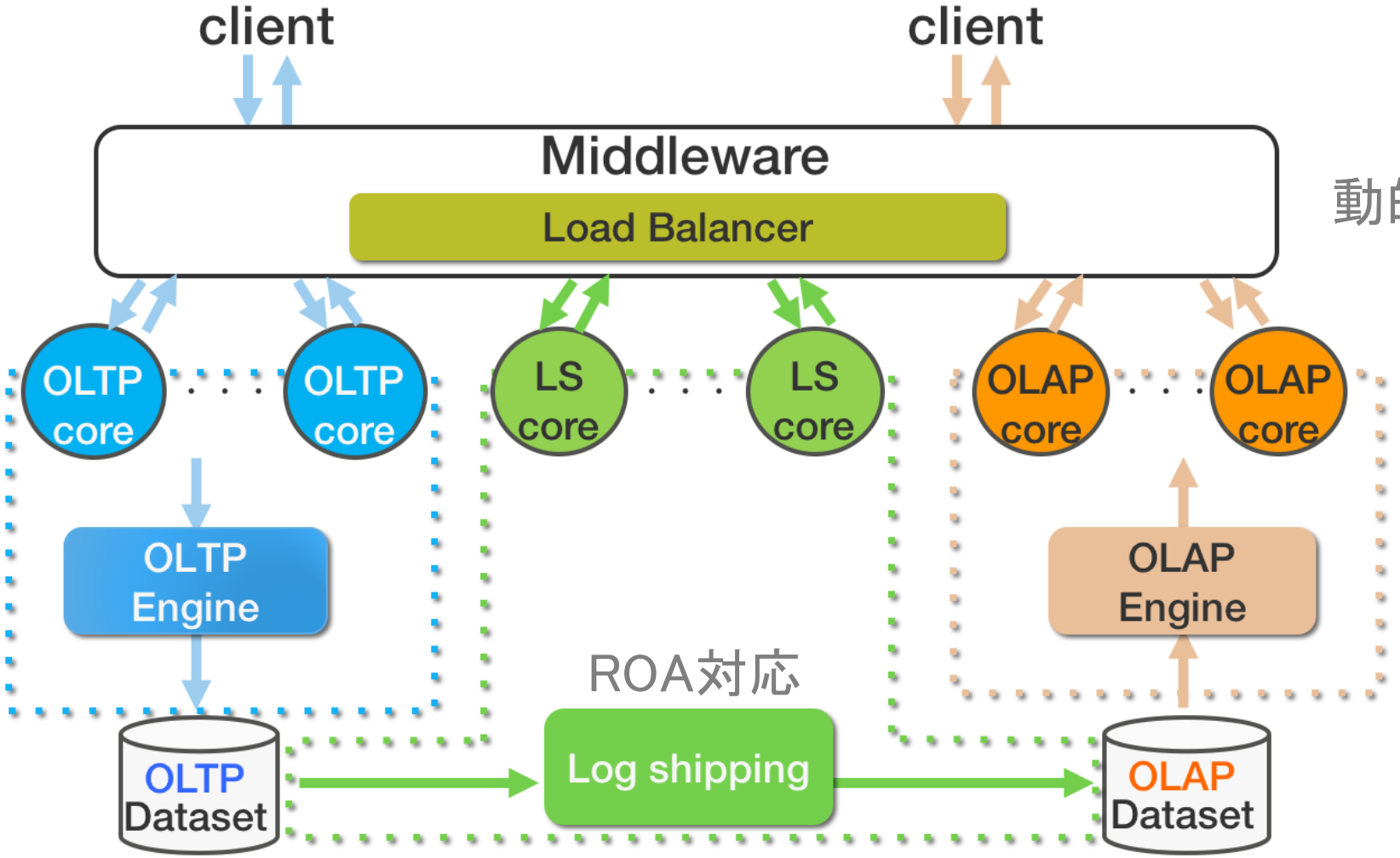
OLTP/OLAP リソース管理

- OLTP と OLAP のリソースを交換
  - [1] A. Raza, P. Chrysogelos, A. C. Anadiotis, and A. Ailamaki. “Adaptive HTAP through Elastic Resource Scheduling”. In Proc. of SIGMOD'20, pp. 2043–2054, (2020)
  - 更新適用のロードや ROA 等は考慮されていない



Many Core での動的リソース管理

- 更新適用での ROA、およびコアアサイン [DEIM2020]



関連文献

1. 諸岡 大輝, 塩井 隆円, 引田 諭之, Le Hieu Hanh, 横田 治夫. 複数コア環境におけるHTAPを想定した OLTP更新のOLAPへの適用手法の検討・評価. 第12回データ工学と情報マネジメントに関するフォーラム論文集, H2-2, 2020.3.
2. 塩井 隆円, 横田 治夫. OLAP用列指向データ構造へのOLTP用行指向データの変換方式の検討. 第11回データ工学と情報マネジメントに関するフォーラム論文集, J5-5, 2019.3.
3. 諸岡 大輝, 塩井 隆円, Le Hieu Hanh, 横田 治夫. 複数コア環境におけるアクセス範囲を考慮したOLTP/OLAP同時実行手法. 第11回データ工学と情報マネジメントに関するフォーラム論文集, H2-1, 2019.3.
4. 塩井 隆円, 横田 治夫. SCMを想定した行単位更新の列指向DBへの反映手法, 第１０回データ工学と情報マネジメントに関するフォーラム論文集, C6-1, 2018.3.
5. Takamitsu Shioi, Kenji Hatano, Haruo Yokota, "SCMAT: A Mechanism Presuming SCMs To Efficiently Enable both OLAP And OLTP", Proceeding of the 6th IEEE International Congress on Big Data (BigData Congress 2017), pp. 313-320, 2017.6.

Tokyo Tech

ご清聴ありがとうございました