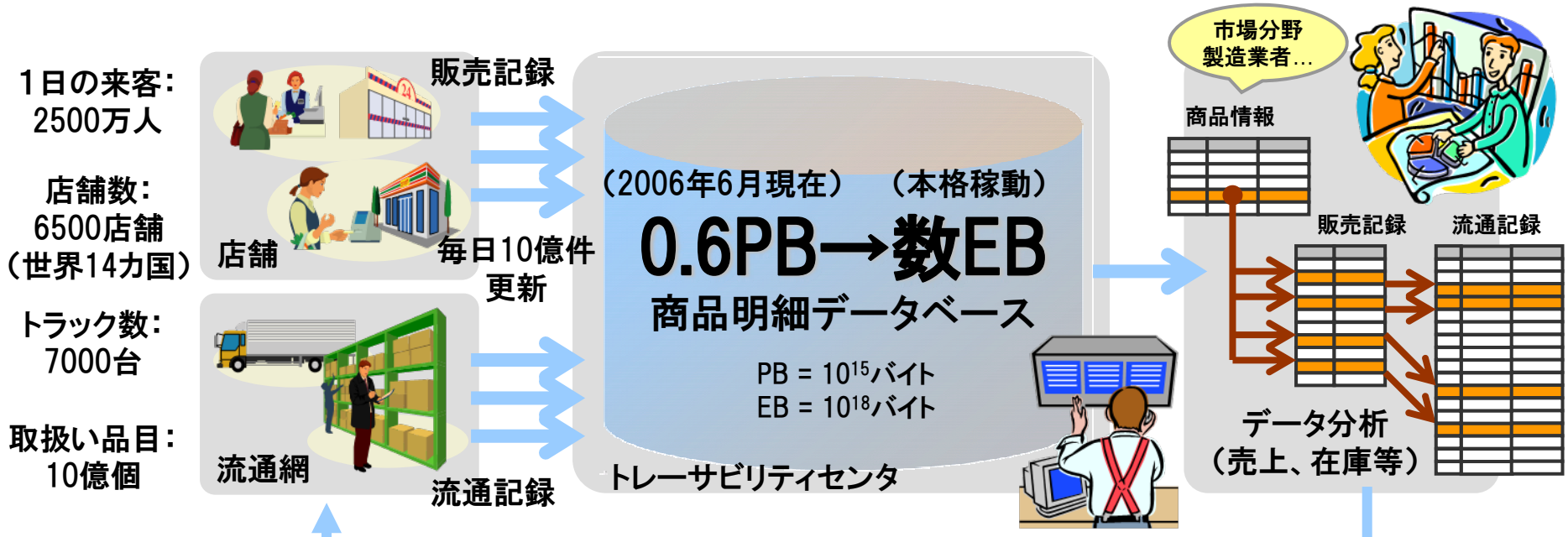
非順序型実行原理に基づく超高性能データベースエンジンの開発

東京大学 生産技術研究所
戦略情報融合国際研究センター
喜連川 優

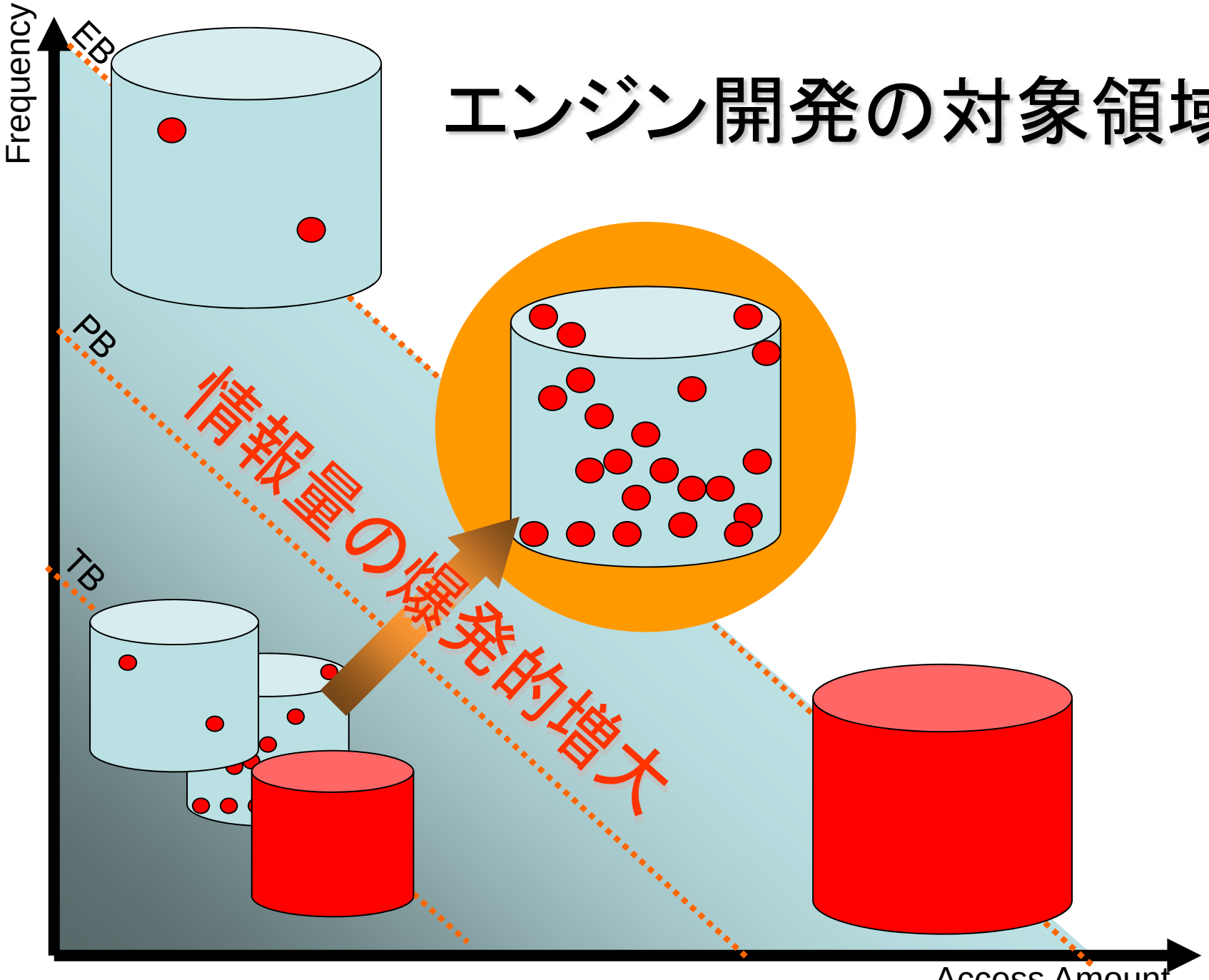
RFID Solution:
WalMartの超巨大データベース

- 米国最大手小売業WalMart社（年間売上高3450億ドル）
  - 2005年より部分的な電子タグ導入試験
    - 全商品の0.021%の荷台に電子タグを付与
    - 販売・流通の効率化、新規事業の創出に活用を予定
  - 2006年6月現在、0.6ペタバイト（PB）の商品明細データベースを保有
    - 本格的な電子タグ適用により少なくともエクサバイト（EB）級に拡大
    - 全個別商品への電子タグ適用により7.7EB／日のデータ生成を予測

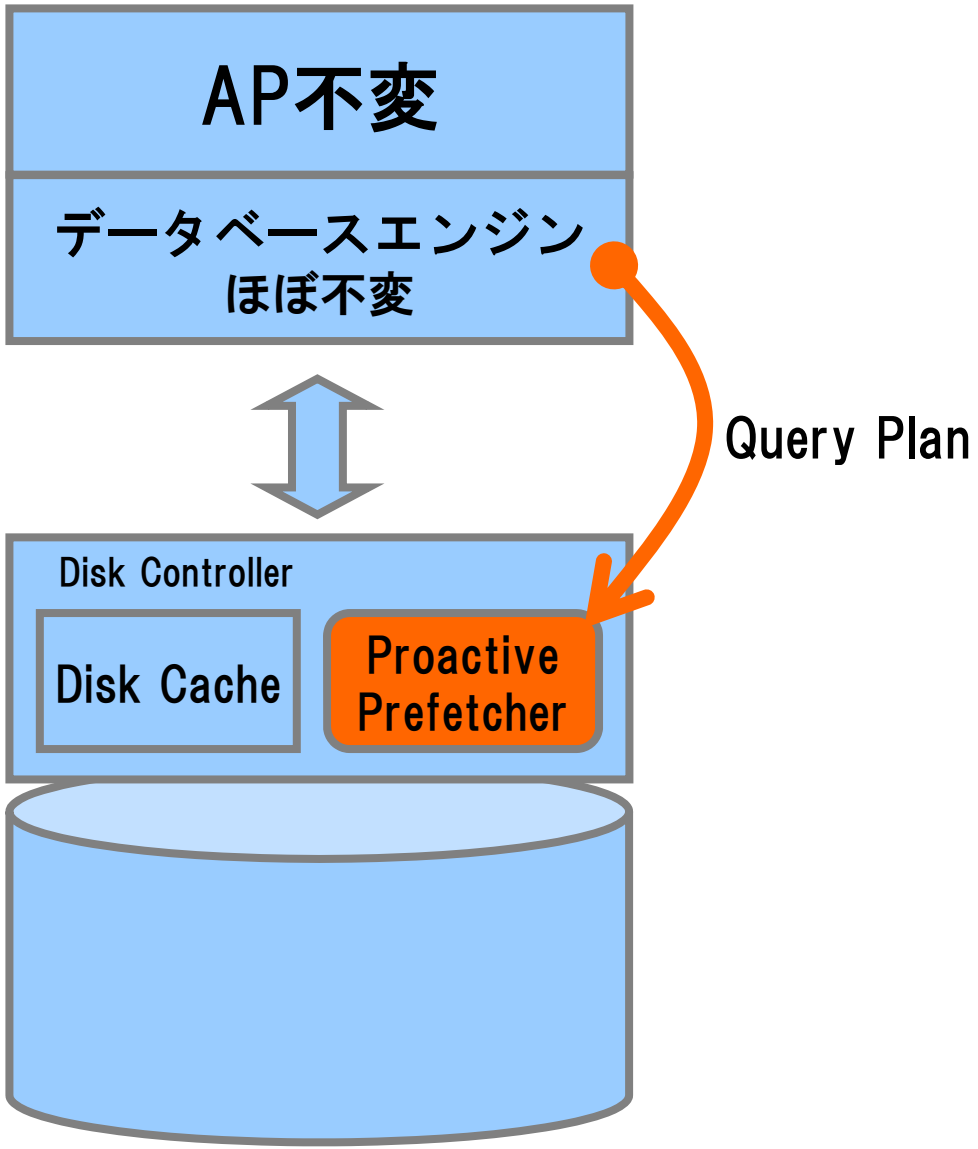


アットマークアイティー電子版、及びInformationWeek電子版の報道記事を元に作成

エンジン開発の対象領域



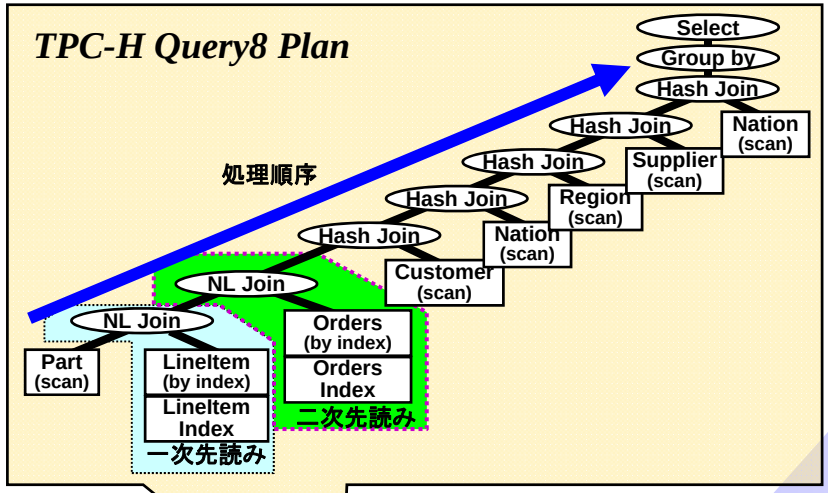
文部科学省　リーディングプロジェクト
「ストレージによるプロアクティブキャッシング」

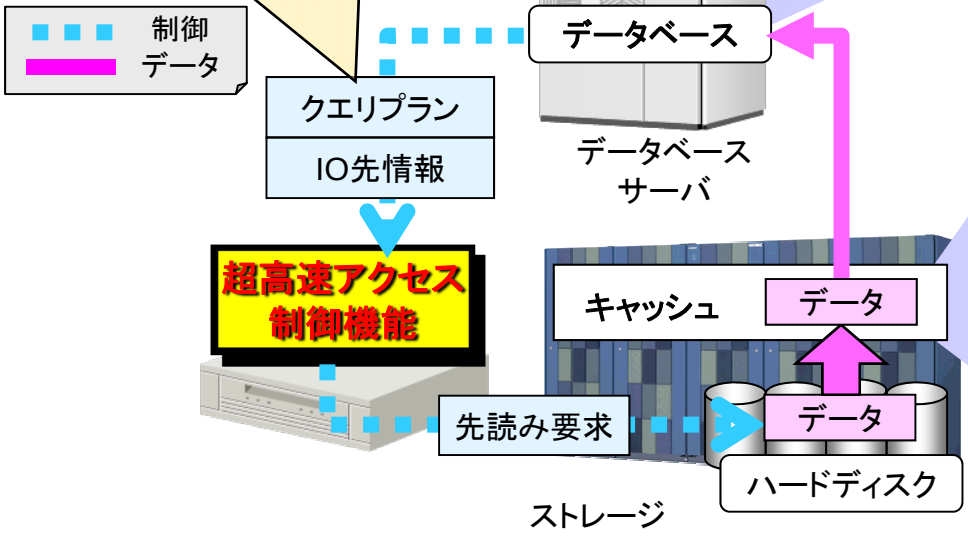


クエリプラン利用
プロアクティブキャッシング

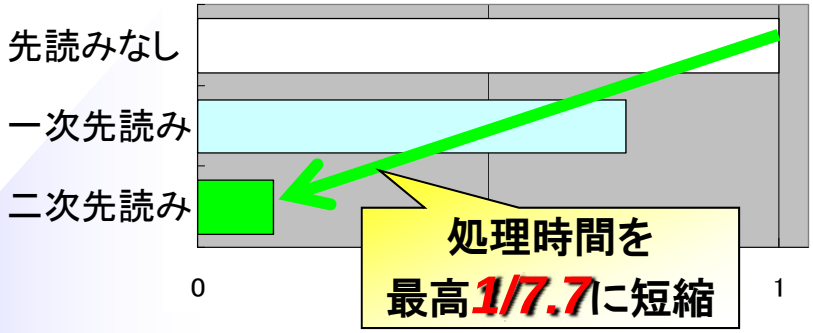
e-Society 文部科学省リーディングプロジェクト

最大7.7倍の性能向上を達成

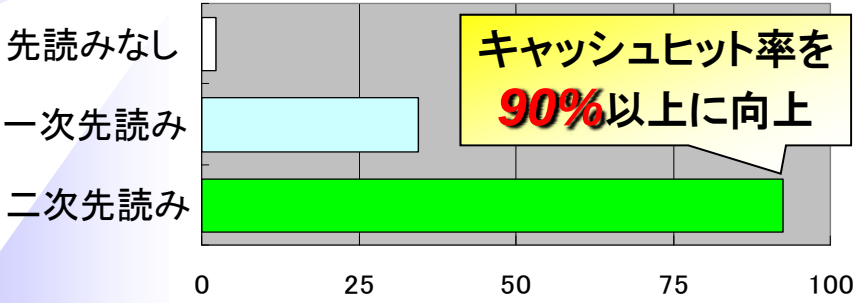




相対処理時間



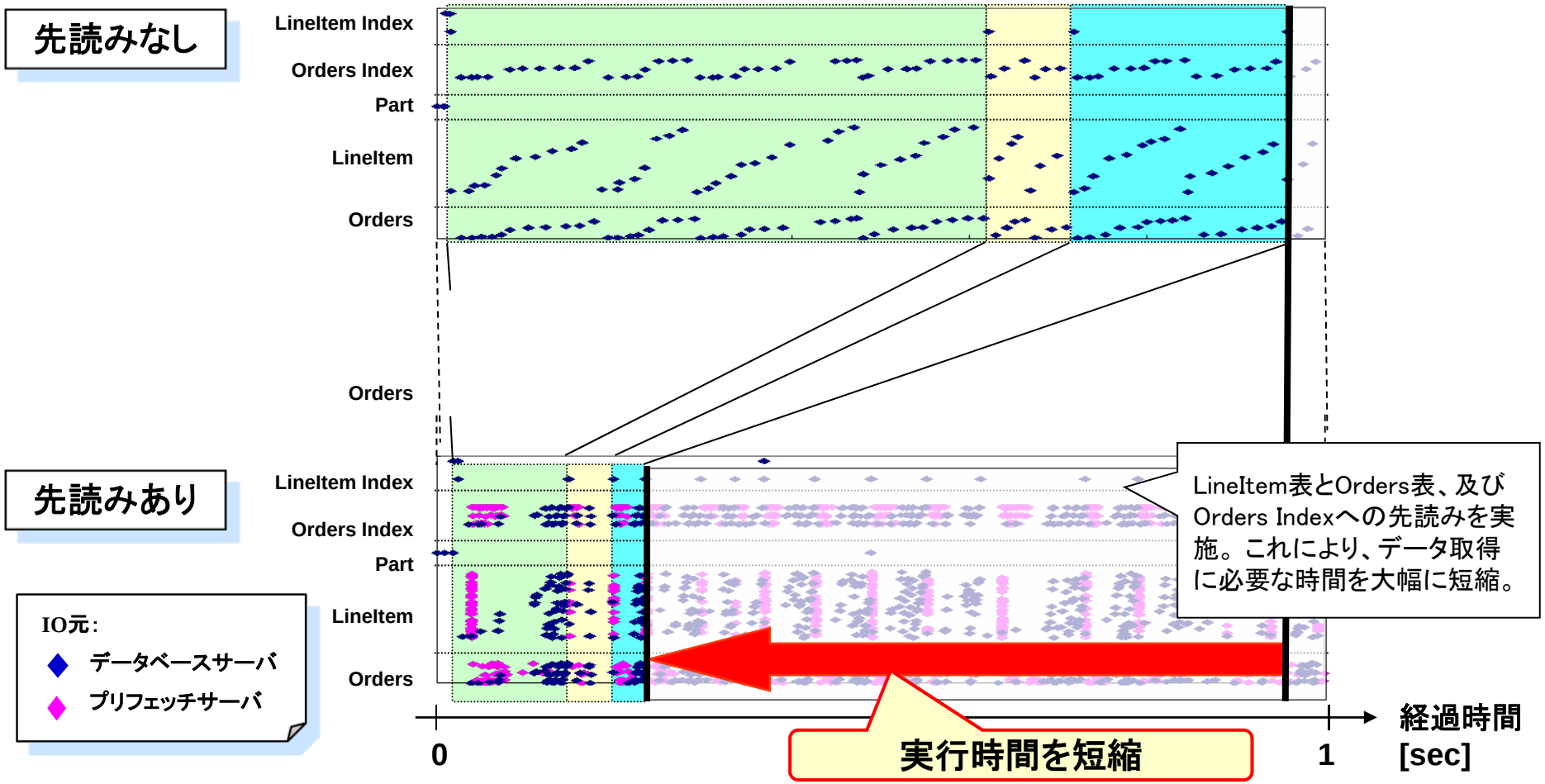
キャッシュヒット率（サイズ：1024MB）



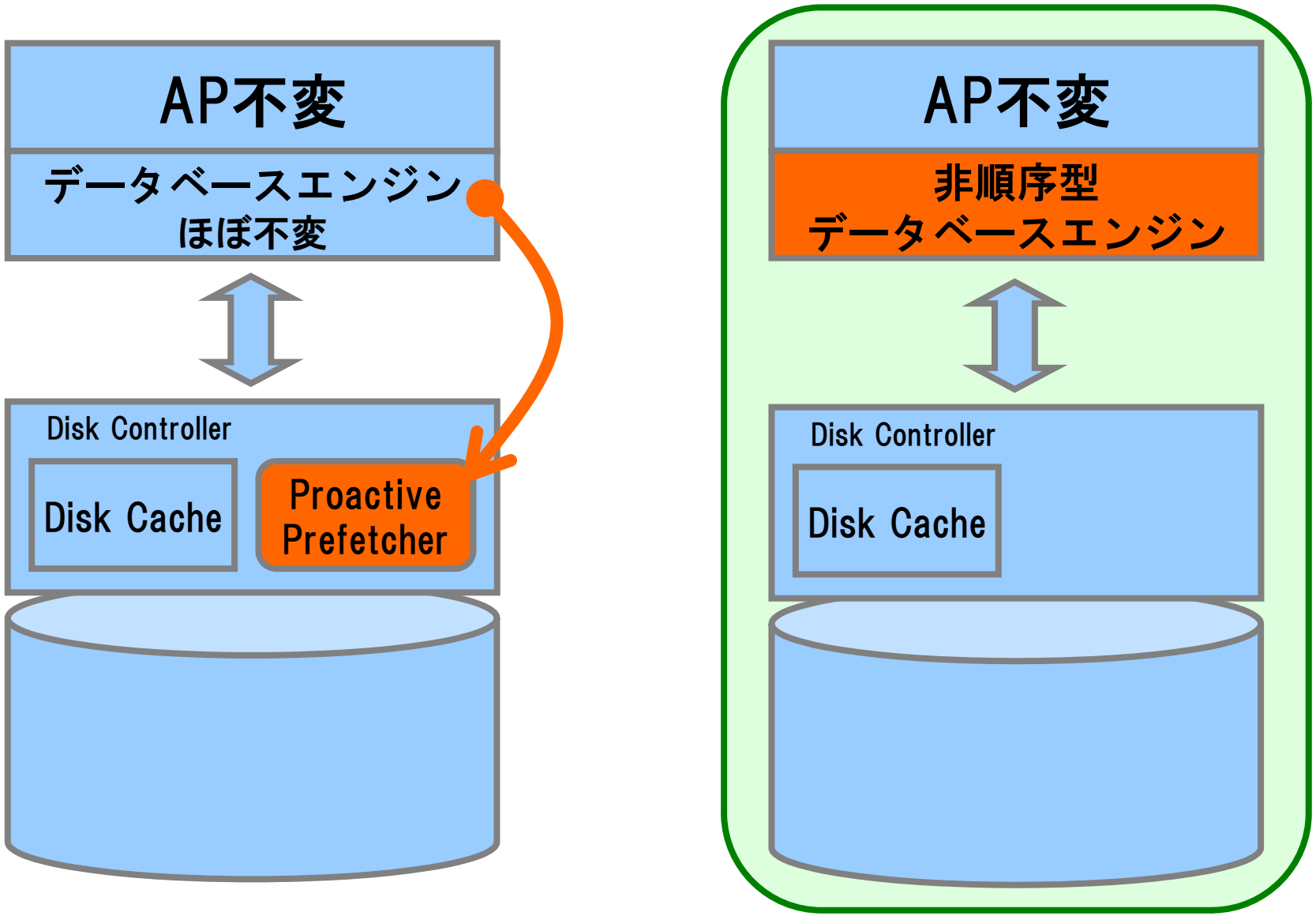
キャッシュサイズ依存性（二次先読み）

クエリプラン利用 プロアクティブキャッシング

■ プロトタイプシステム評価 –IO発行状況–



本プロジェクトの目指すところ



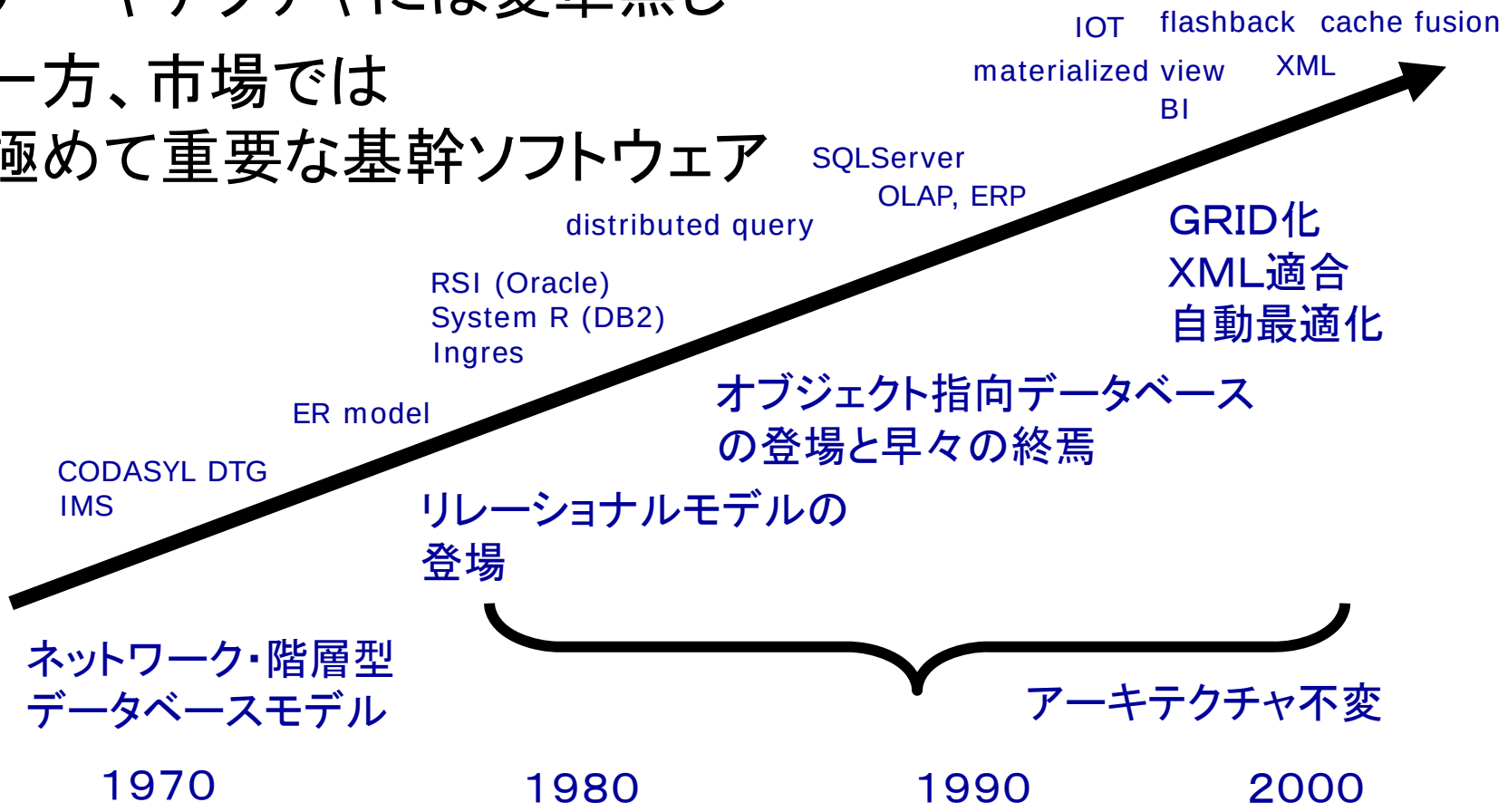
OoODE (Out-of-Order Database Engine):
非順序型データベースエンジン

- 非順序型実行原理に基づく
  超高性能データベースエンジン

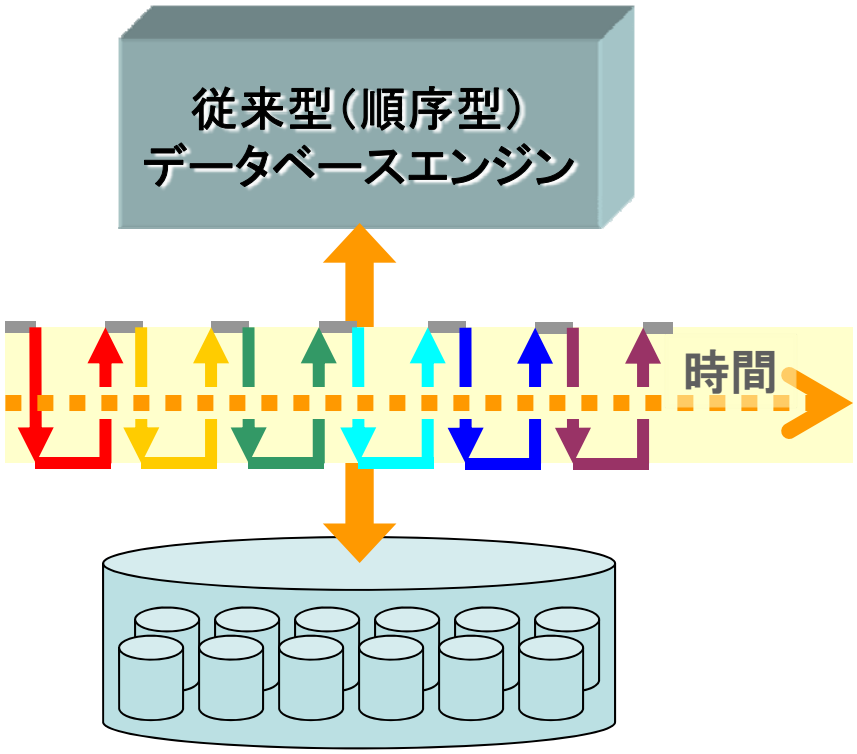
- 特徴
  - 超大量非同期IO発行機構
  - ストレージ駆動型アウトオブオーダ実行機構
  - 実行時動的IOスケジュール機構

新しい実行原理に基づく独自のデータベースエンジンの創出

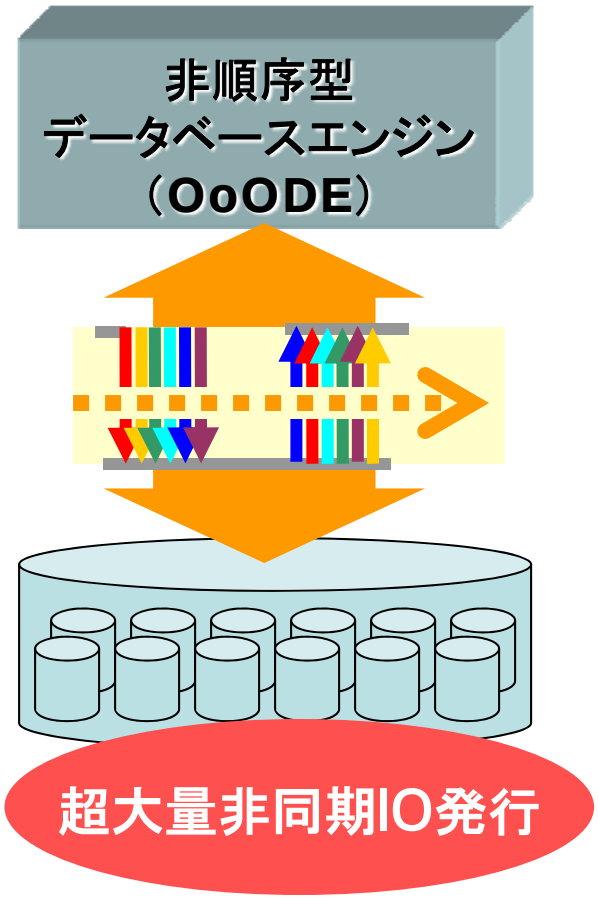
- 20年以上、リレーショナルデータベースエンジンのアーキテクチャには変革無し
- 一方、市場では極めて重要な基幹ソフトウェア



従来型データベースエンジンの問題と非順序型データベースエンジン

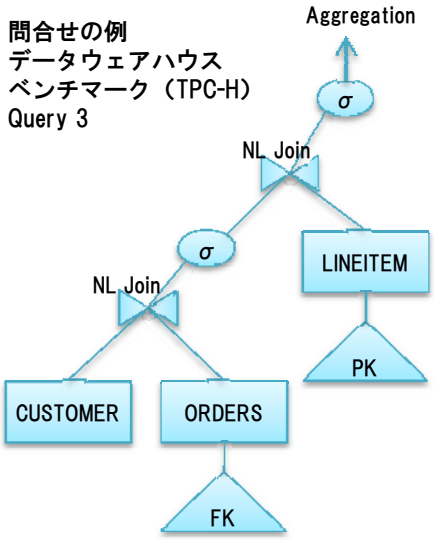


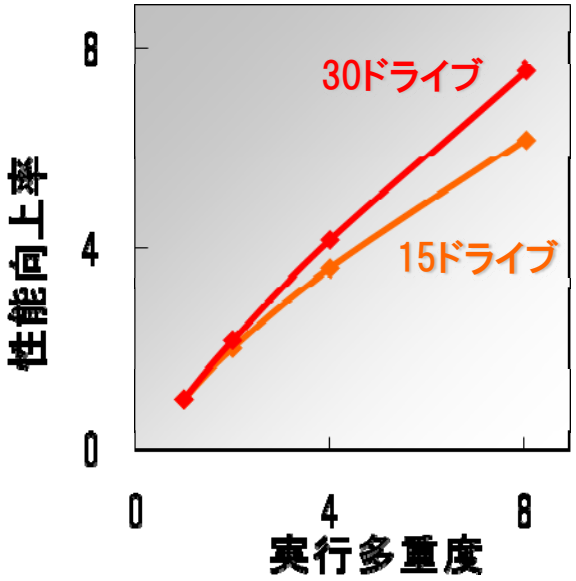
同期IO発行

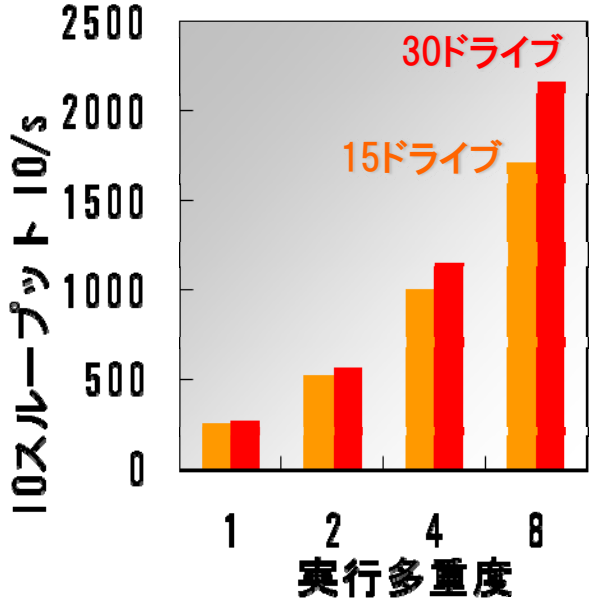


OoODE 実験結果

・Nested Loop 結合演算における性能向上効果を確認
（1段分の処理並列性を抽出）





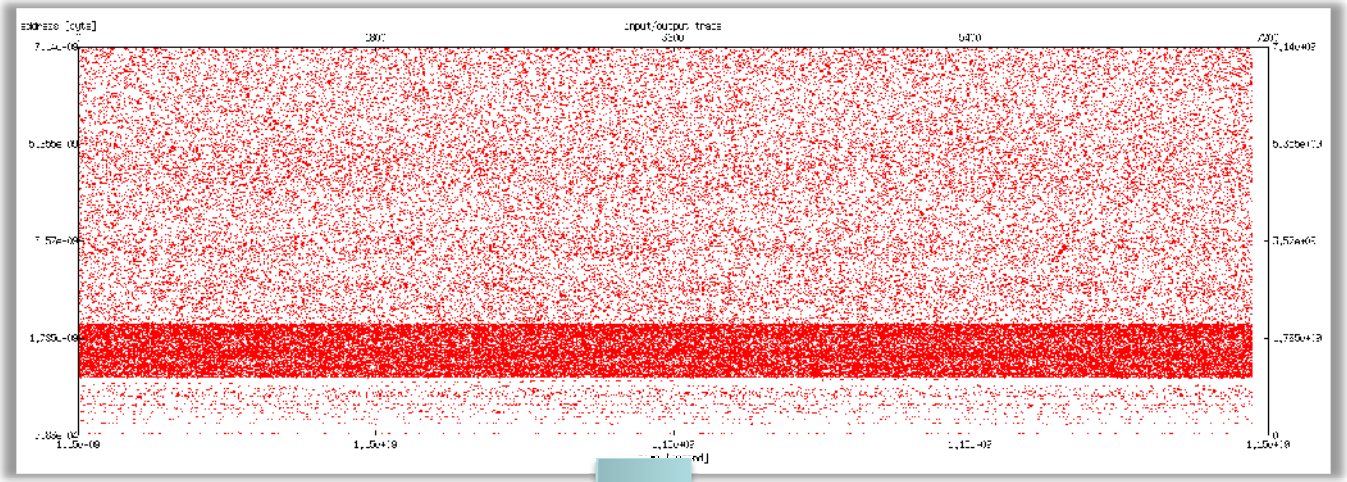


TPC-Hデータセット(SF=4.0; 4.0GB) Q.3
Itanium2 1.67GHz * 8, 15/30 HDDs, RedHat Linux AS3

超大量IO発行による大幅な性能向上

従来型（順序型）データベースエンジン

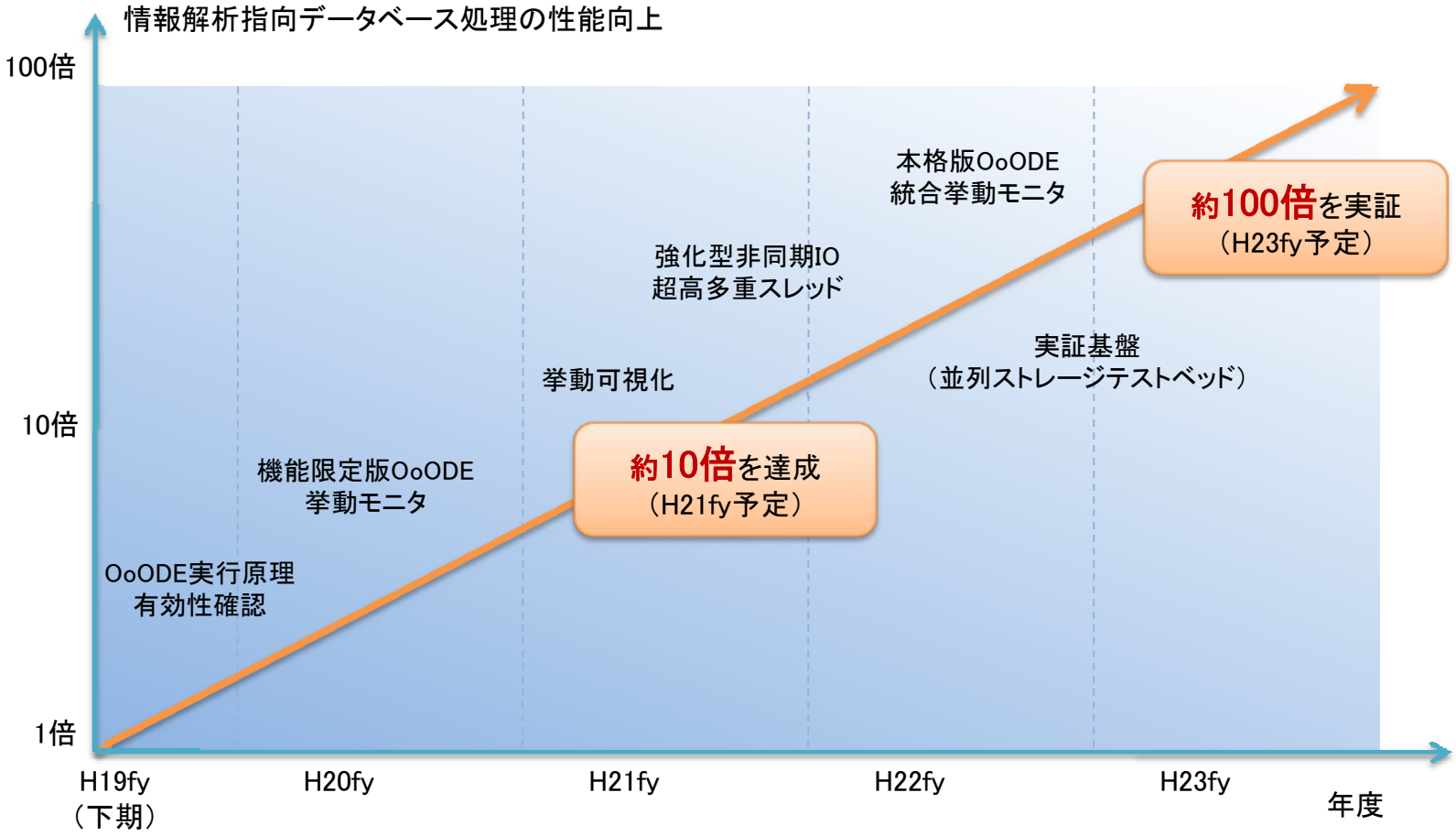
TPC-H ベンチマーク Q3 MySQL



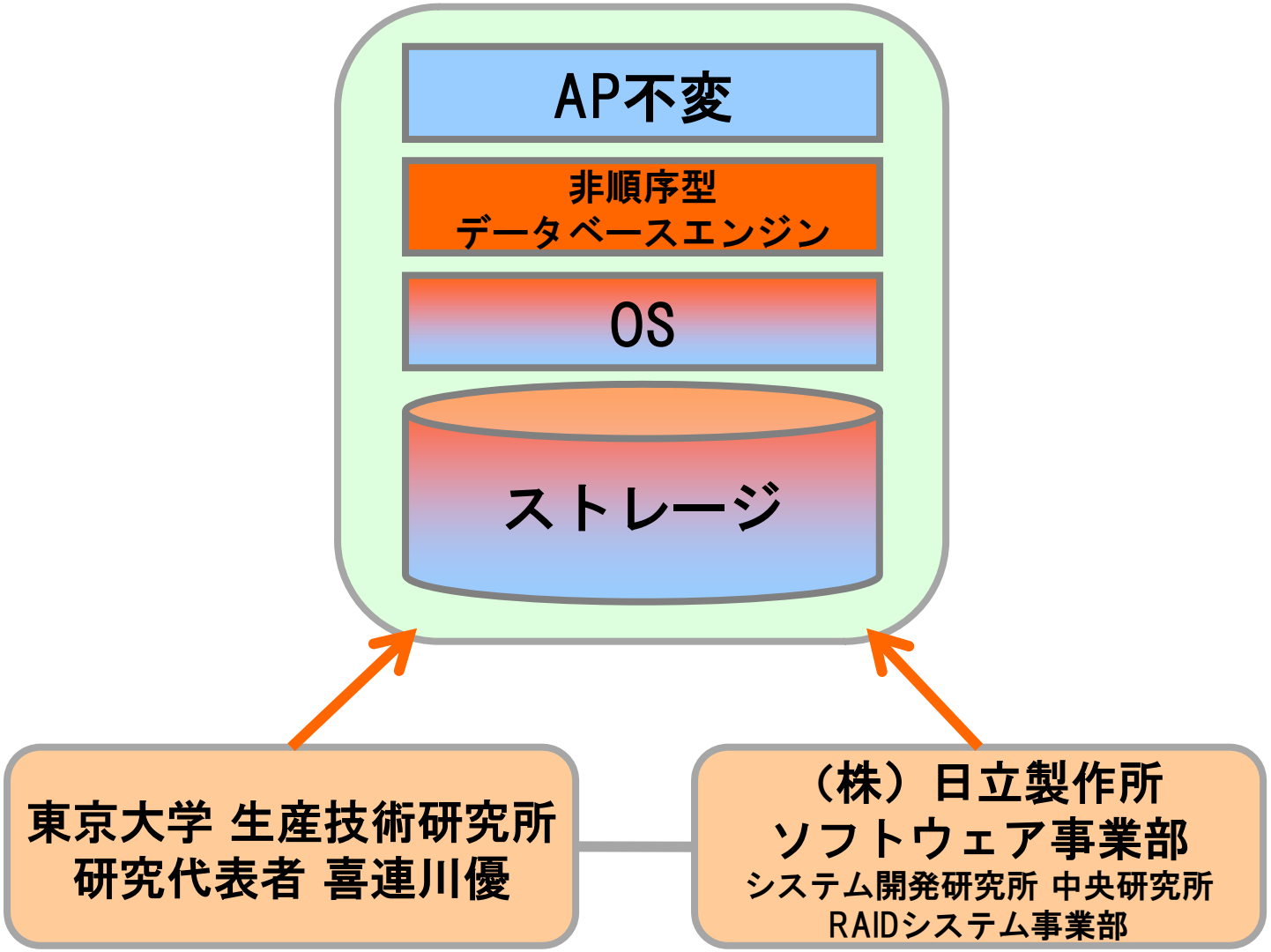
非順序型データベースエンジン

ロードマップ

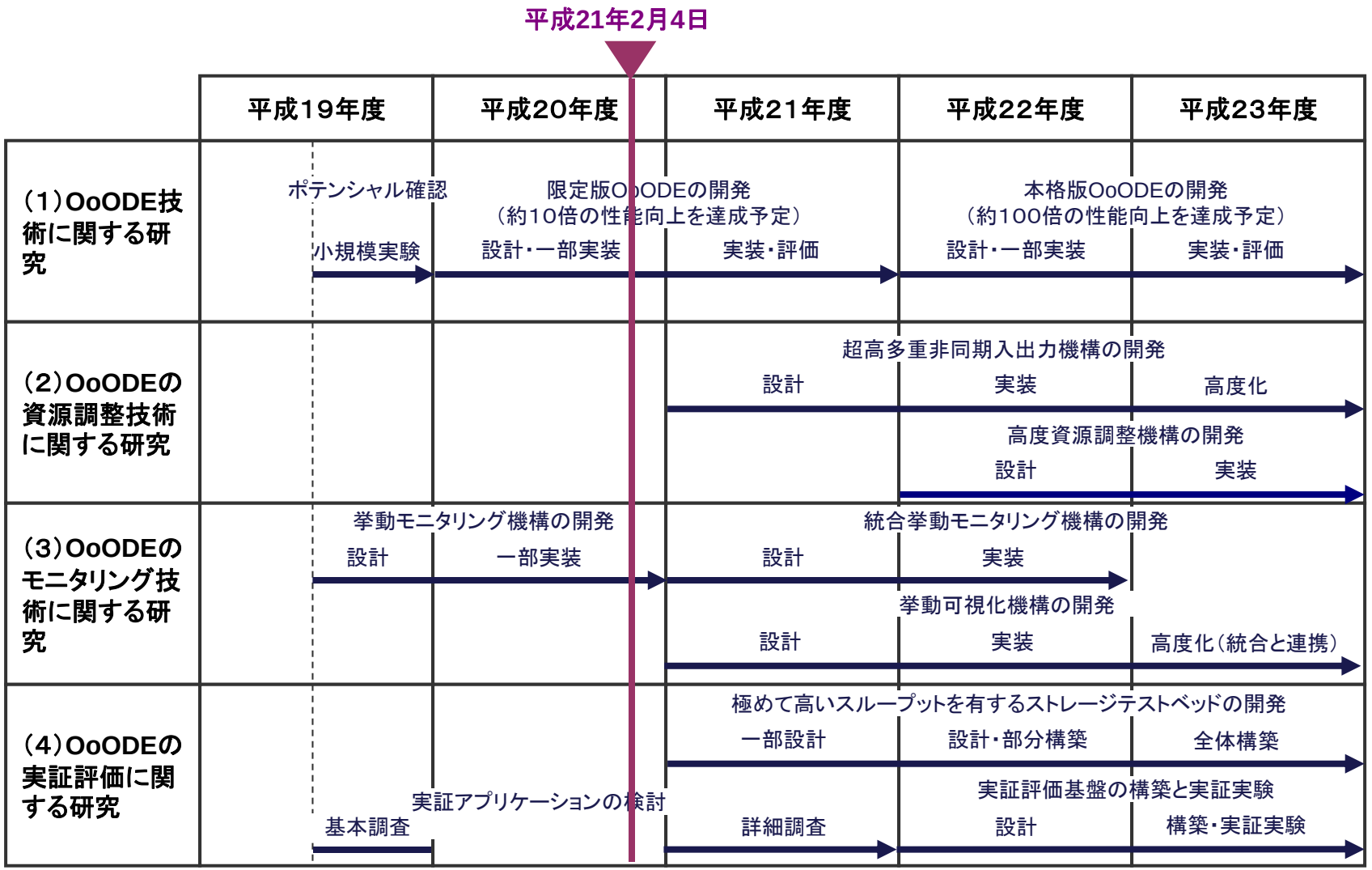
データベース処理性能ブレークスルーへの挑戦
（世界に先駆けた超巨大情報活用技術の創出による競争力の強化）



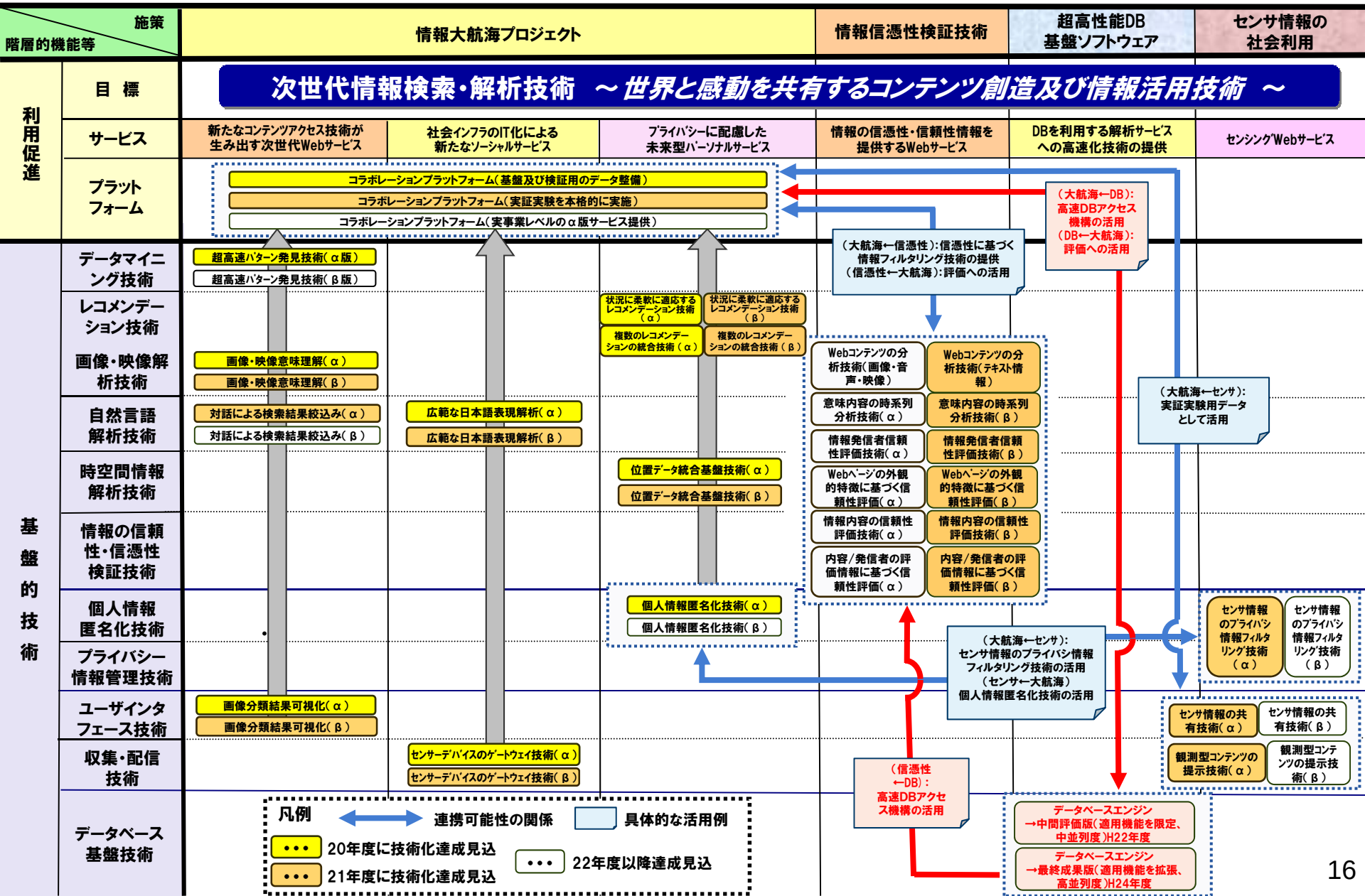
本プロジェクトの開発部位と実施体制



研究開発計画


平成21年2月4日
平成19年度
平成20年度
平成21年度
平成22年度
平成23年度
(1)OoODE技術に関する研究
ポテンシャル確認
小規模実験
限定版OoODEの開発
(約10倍の性能向上を達成予定)
設計・一部実装
実装・評価
本格版OoODEの開発
(約100倍の性能向上を達成予定)
設計・一部実装
実装・評価
(2)OoODEの資源調整技術に関する研究
超高多重非同期入出力機構の開発
設計
実装
高度化
高度資源調整機構の開発
設計
実装
(3)OoODEのモニタリング技術に関する研究
挙動モニタリング機構の開発
設計
一部実装
統合挙動モニタリング機構の開発
設計
実装
挙動可視化機構の開発
設計
実装
高度化(統合と連携)
(4)OoODEの実証評価に関する研究
極めて高いスループットを有するストレージテストベッドの開発
一部設計
設計・部分構築
全体構築
実証アプリケーションの検討
基本調査
詳細調査
実証評価基盤の構築と実証実験
設計
構築・実証実験

他プロジェクトとの連携



非順序型実行原理に基づく超高性能データベースエンジンの開発

日立の取り組みと実現方式検討・評価結果

株式会社 日立製作所
ソフトウェア事業部

河村　信男

1-1. はじめに

- 日立の取り組み方針
  - 非順序型データベースエンジン技術開発における日立の強み
    - MFからオープン系の自社開発のRDBMS製品を有する
    - Linuxカーネル開発コミュニティへの参画実績を有する
    - 自社開発のストレージ製品を有する

非順序型データベースエンジン技術の迅速な確立・実用化に貢献

・自社開発RDBMS製品の実装技術を基にした、非順序型データベースエンジンの実装技術開発の迅速化に貢献

・日立が有する技術を活用し、データベースエンジン・OS・ストレージからなるシステム全体の挙動を一元的に把握する挙動モニタリング機能を実現し、開発迅速化に貢献

2-1. 関連用語

用語	説明
実行Context	非順序型実行原理のもとで行われるデータベース処理のうち、並列に実行することが可能な処理単位。
タスク	処理系において、実行Contextを並列に処理するために設ける論理的な手続き単位。 実行Contextの対象となるデータと処理の管理情報の組合せとして実体化される。
スレッド	タスクをCPU上で実行する際の物理的な手続き単位、もしくはそれを制御する機構。 ユーザ空間で実装されプロセス内でスレッド切り替えが可能なユーザスレッド機構や、ユーザOSカーネルが提供するカーネルスレッド機構などがある。

2-2. 非順序型データベースエンジン実現への課題と取り組み

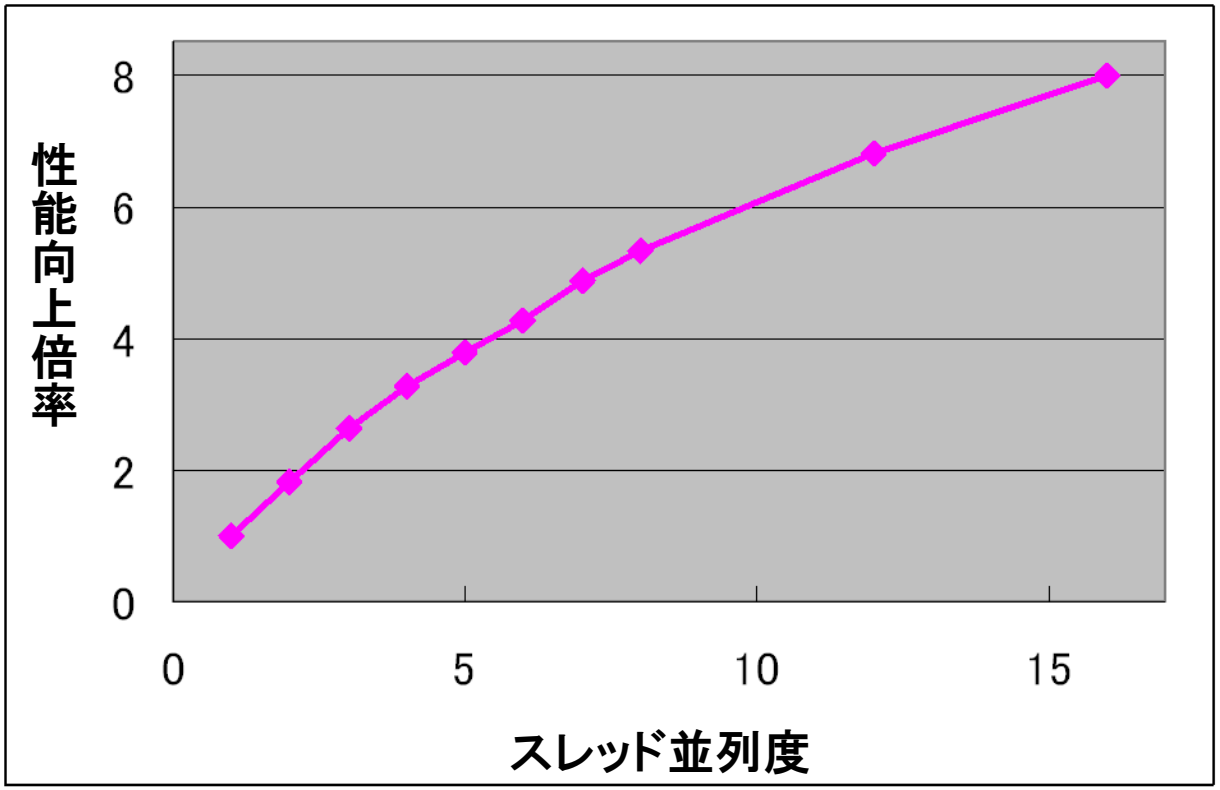
- 研究開発技術と課題の対応関係

<table>
<tr><th>レイヤ</th><th>研究開発技術</th><th>平成19・20年度の研究開発</th><th>課題</th></tr>
<tr><td rowspan="3">データベースエンジン</td><td rowspan="2">非順序型データベースエンジン技術</td><td>中並列度の機能限定版非順序型データベースエンジンの実現方式検討・部分実装評価</td><td>レコードレベルの処理細分化</td></tr>
<tr><td rowspan="2">非順序型データベースエンジンの内部挙動モニタリング機構の実現方式検討・部分実装評価</td><td rowspan="2">超高並列タスク処理</td></tr>
<tr><td rowspan="3">非順序型データベースエンジンのモニタリング技術</td></tr>
<tr><td>OS</td><td rowspan="2">OS・ストレージ入出力処理挙動モニタリング機構の実現方式検討・部分実装評価</td><td rowspan="2">超高多重IO処理</td></tr>
<tr><td>ストレージ</td></tr>
</table>

3　機能限定版非順序型データベースエンジン

Nested-Loop結合処理において、約8倍の性能向上を実現

機能限定版非順序型データベースエンジンの性能評価結果



CPU: Xeon 2.40 GHz
メモリ: 1GB
HDD: 10台 （10 krpm）
データ: TPC-Hデータセット
（SF=30、約30GB）
処理: TPC-H　Query8相当

※TPC-H:
意思決定支援システムを模した
業界標準RDBMSベンチマーク

4-1. 要素挙動モニタリング機構：概要

目的：
複雑かつ従来とは異なる処理特性を有する非順序型実行原理による処理の挙動を把握することにより、以下を実現

①開発効率の向上
②最高性能を引き出すシステム構成の明確化

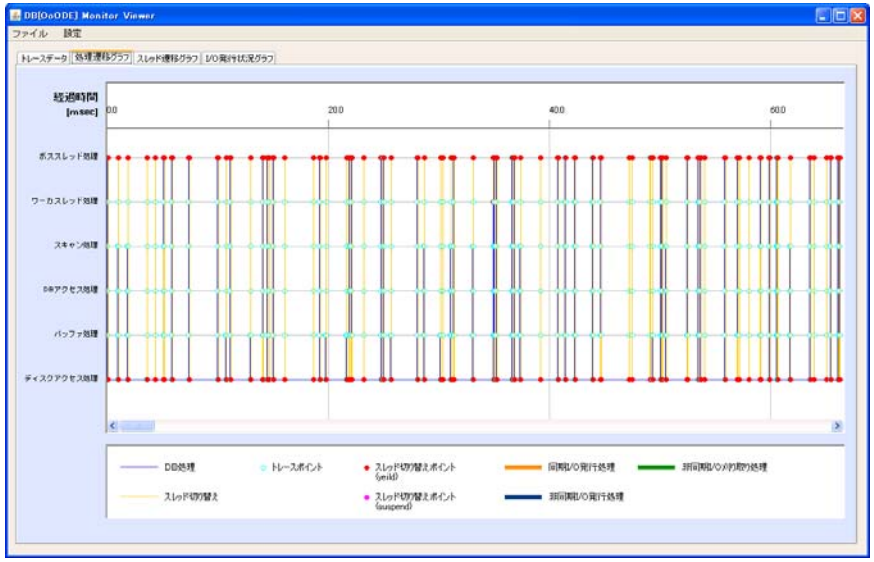
- 非順序型データベースエンジンのモニタリング技術が解決すべき課題
  - 非順序型データベースエンジンの内部挙動
    - 非順序型データベースエンジンの***超高並列タスク処理***における複雑な挙動の把握
  - OS・ストレージ挙動
    - 今後実現する、従来システムにおける想定以上での***超高多重IO処理***による処理オーバヘッド・システム挙動を把握

4-2. 要素挙動モニタリング機構：データベースエンジン

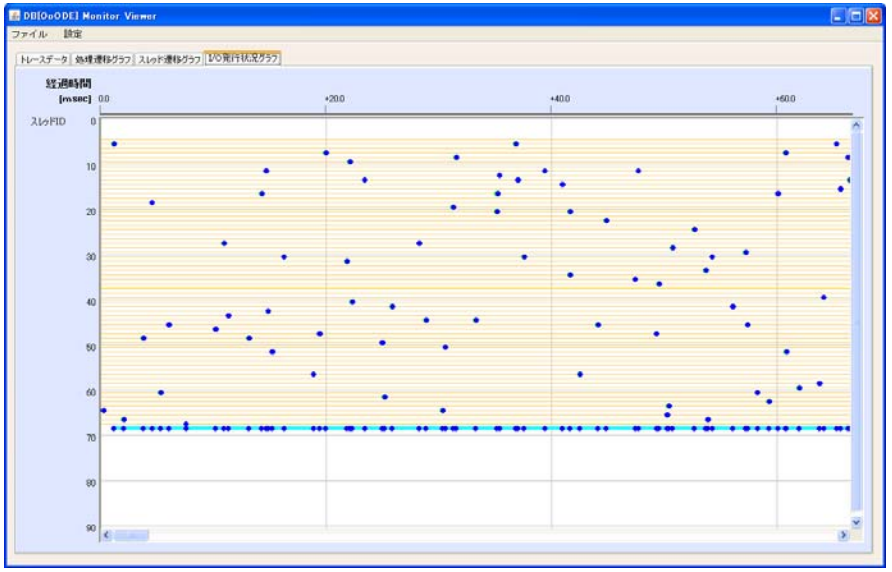
挙動モニタリングにより、複雑な高並列処理の開発（デバッグ）効率を向上

- データベースエンジン内 処理遷移のモニタリング
  並列タスク処理の処理遷移のモニタリングにより、処理挙動を把握
- データベースエンジン IO発行状況のモニタリング
  並列実行タスクからのIO発行状況のモニタリングにより、IO挙動を把握

内部処理遷移のモニタリング



IO発行状況のモニタリング

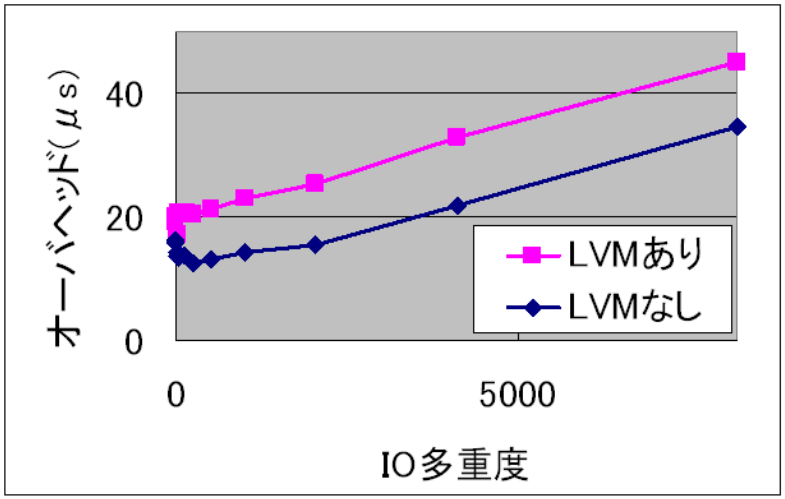


4-3. 要素挙動モニタリング機構：OS・ストレージ

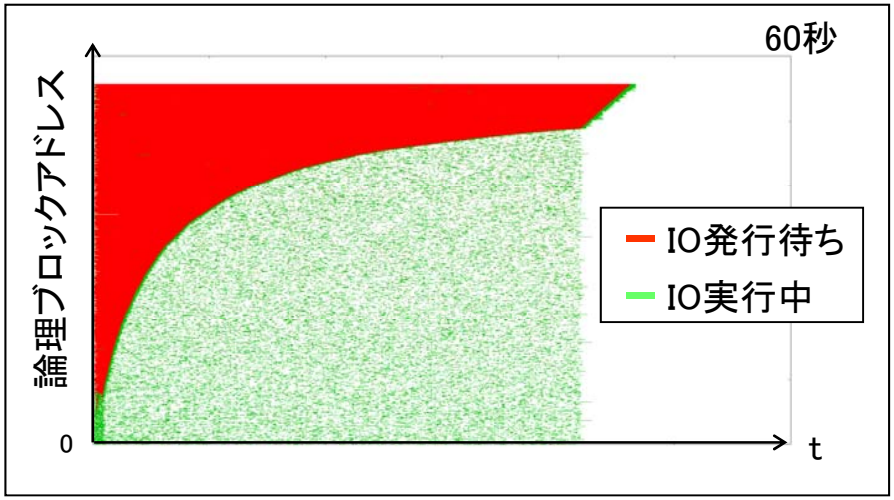
今後実現する超高多重IO処理における処理オーバヘッド計測・挙動明確化

- 超高多重IO処理のオーバヘッド解析
  LVM(OSのストレージ記憶領域動的変更機能)の処理オーバヘッドを把握
- 超高多重IO処理におけるOS内IO処理挙動モニタリング
  超高多重IO処理におけるIO処理挙動を把握、IO処理不公平性を確認

超高多重IO処理のオーバヘッド解析
(IO開始処理, RHEL4)



超高多重IO処理におけるOS内IO処理挙動のモニタリング （cfq IOスケジューラ, RHEL4 ）



※RHEL4: Red Hat Enterprise Linux v.4

5-1. 今後の取り組み方針

- 今後の研究開発
  - 非順序型データベースエンジン技術：
    適用演算を拡大し並列度を高めた本格版非順序型データベースエンジンの実現による更なる高性能化
  - 非順序型データベースエンジンのモニタリング技術：
    データベースエンジン・OS・ストレージ挙動モニタリング機構の統合によるシステム全体の挙動把握の容易化、開発の加速
- 実用化に向けた取り組み
  - 情報解析指向の超巨大データ活用アプリケーションを想定した実証評価：
    - 既存データ解析システムの大規模化と細粒度解析化
      - データ統合による顧客個別マーケティング向けデータ解析
    - 新技術・新ビジネスにより発生した大量データの解析
      - RFIDによるトレーサビリティデータの解析
      - 電子マネートランザクションの特定ユーザ解析（不正発見、等）
  - ニーズが顕在化していない用途（新市場）の調査・開拓