次世代IT基盤構築のための研究開発

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

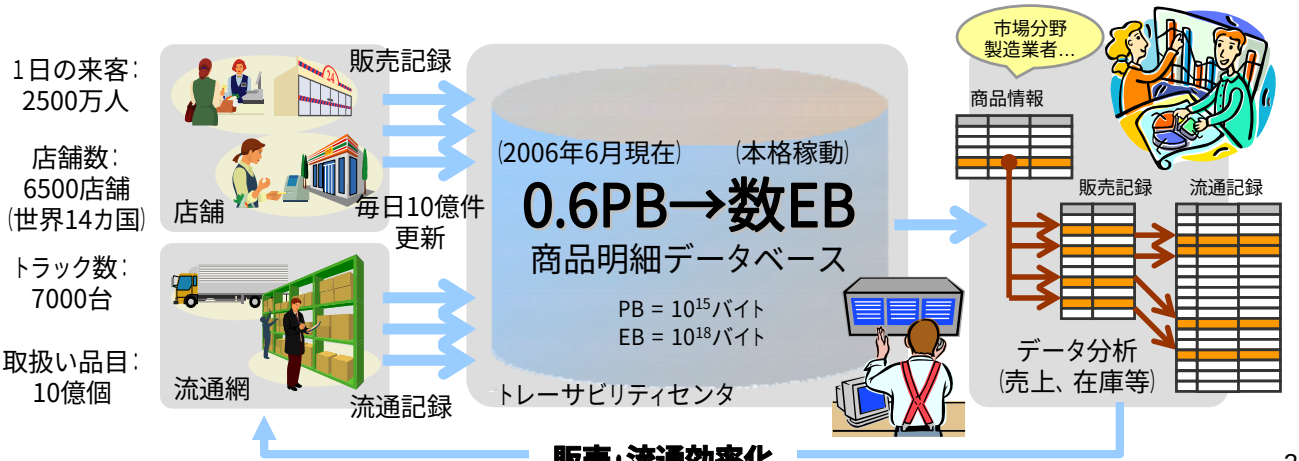
研究代表者 喜連川 優
(東京大学 生産技術研究所)

研究分担企業 日立製作所

RFID Solution: WalMartの超巨大データベース

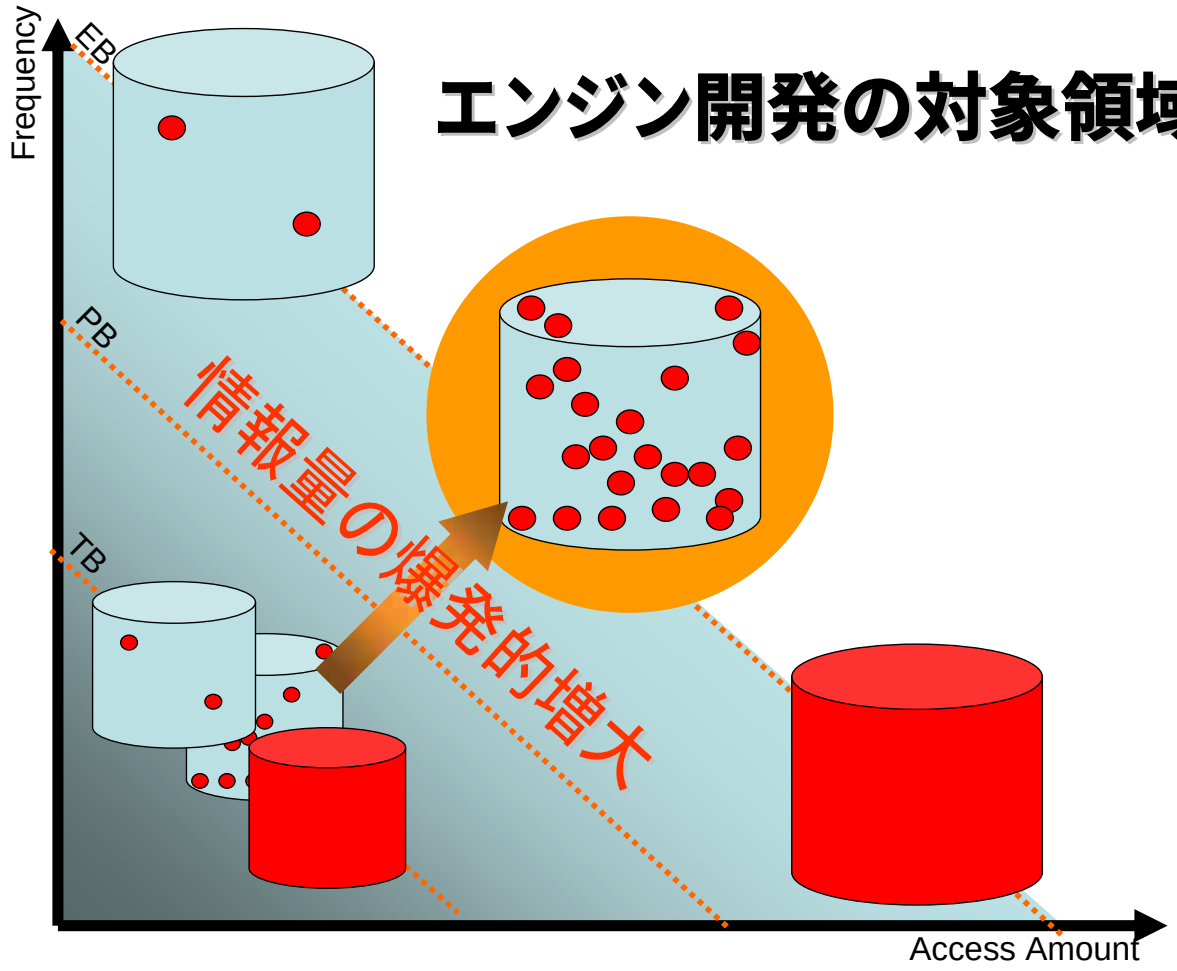
- 米国最大手小売業WalMart社 (年間売上高3450億ドル)
  - 2005年より部分的な電子タグ導入試験
    - 全商品の0.021%の荷台に電子タグを付与
    - 販売・流通の効率化、新規事業の創出に活用を予定
  - 2006年6月現在、0.6ペタバイト (PB) の商品明細データベースを保有
    - 本格的な電子タグ適用により少なくともエクサバイト (EB) 級に拡大
    - 全個別商品への電子タグ適用により7.7EB／日のデータ生成を予測

RFID対応の業者
100社 (2005.01) → 600社以上 (2008.09)
データベースは小さくとも数PB級以上に成長



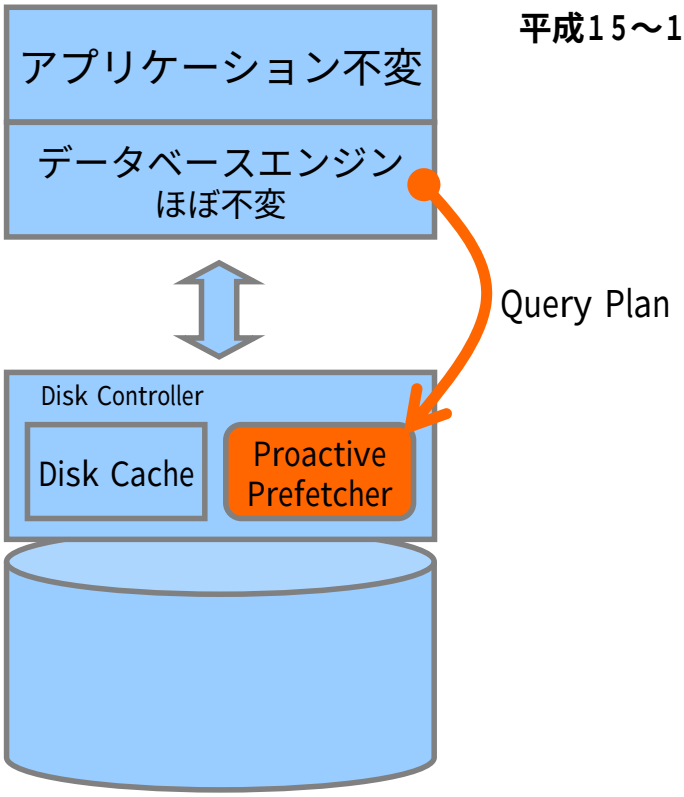
アットマークアイティー電子版、及びInformationWeek電子版の報道記事を元に作成

エンジン開発の対象領域



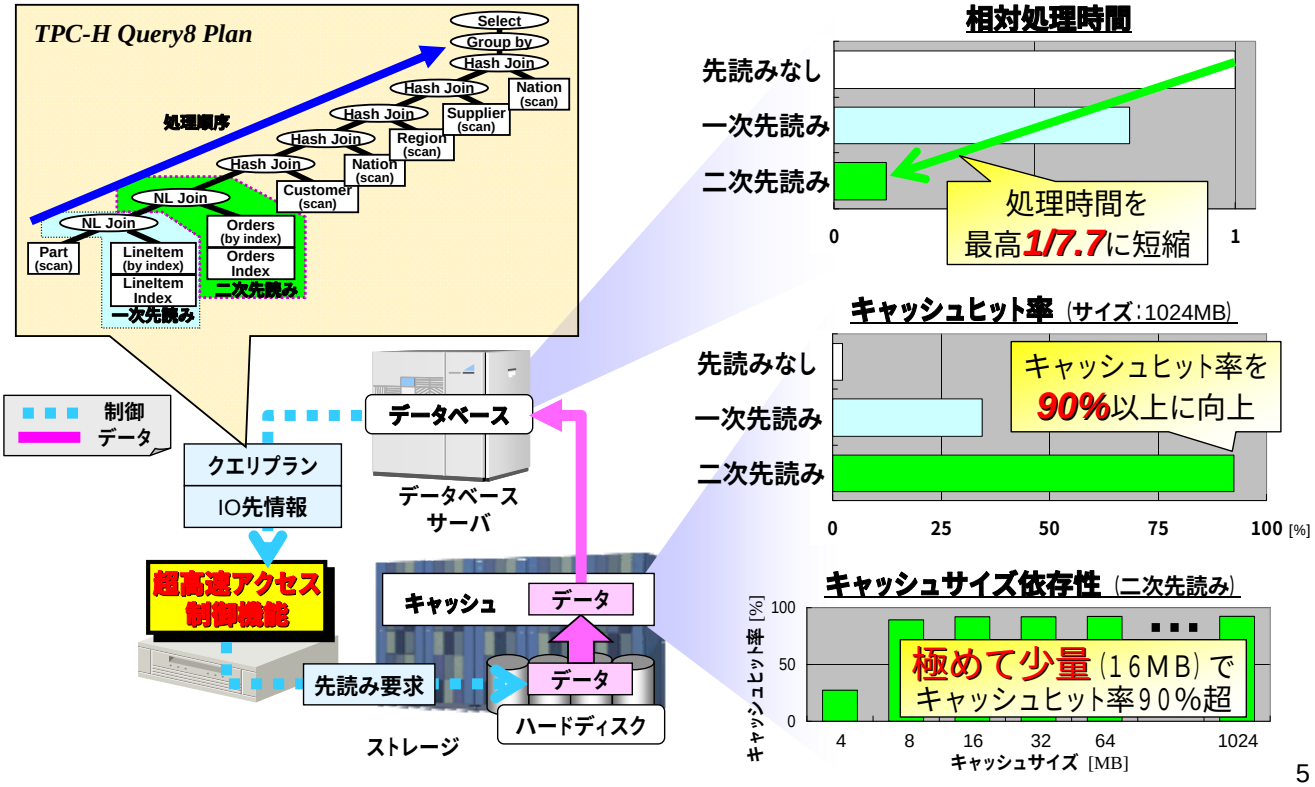
文部科学省　リーディングプロジェクト「ストレージによるプロアクティブキャッシング」

平成15～19年度実施



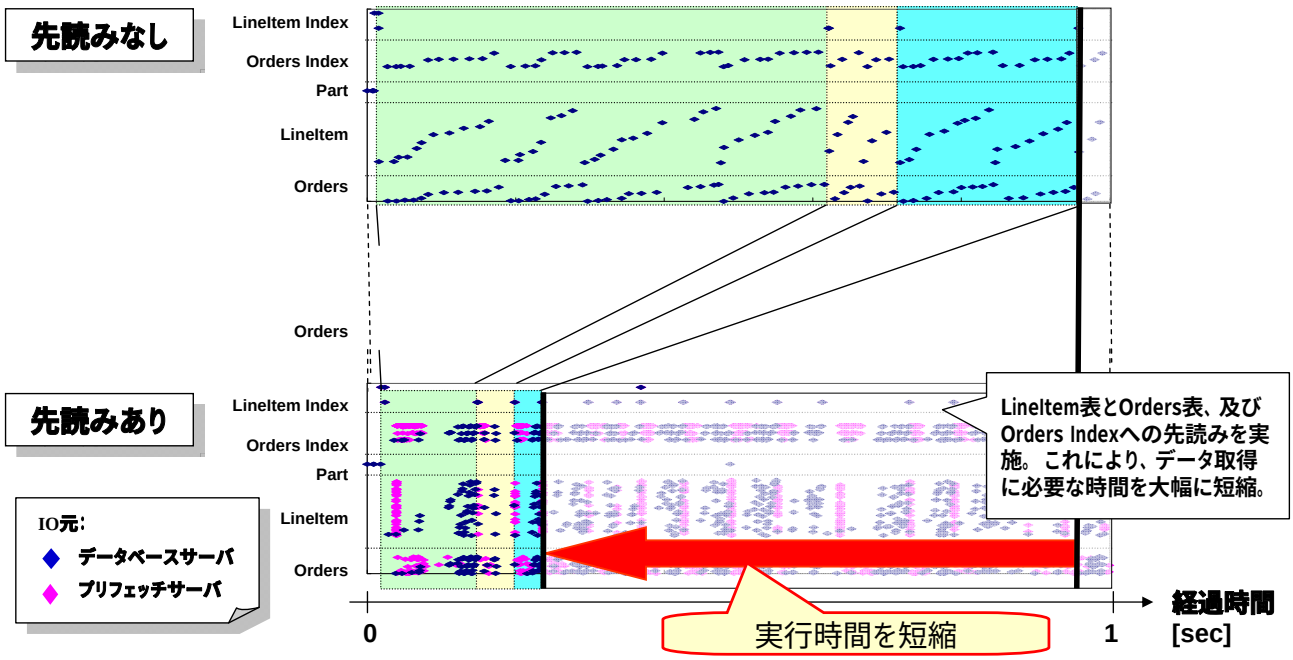
クエリプラン利用
プロアクティブキャッシング

最大7.7倍の性能向上を達成

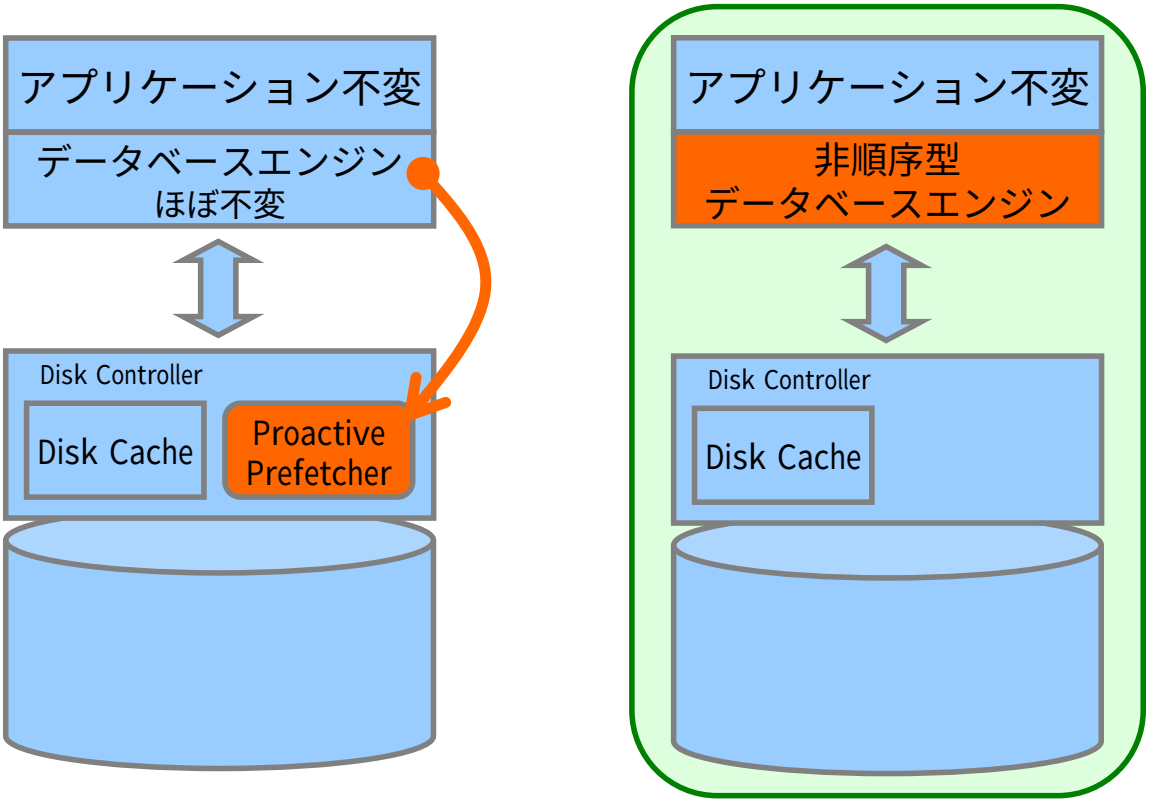


クエリプラン利用
プロアクティブキャッシング

- 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



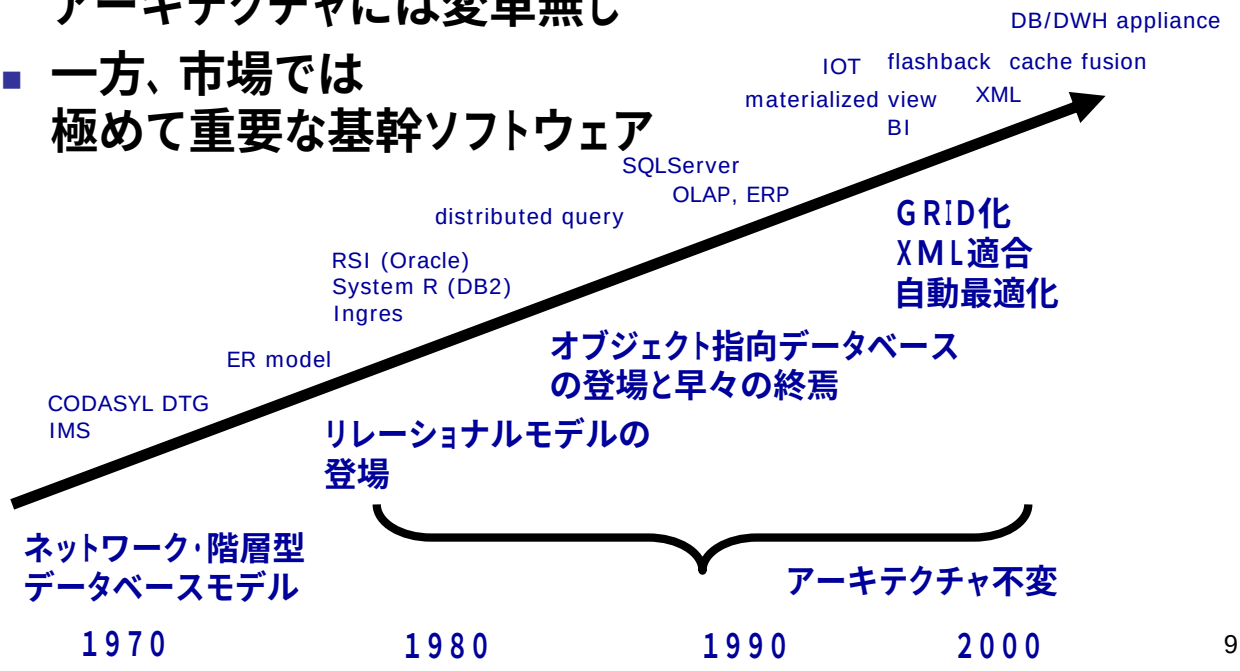
OoODE (Out-of-Order Database Engine):
非順序型データベースエンジン

- 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

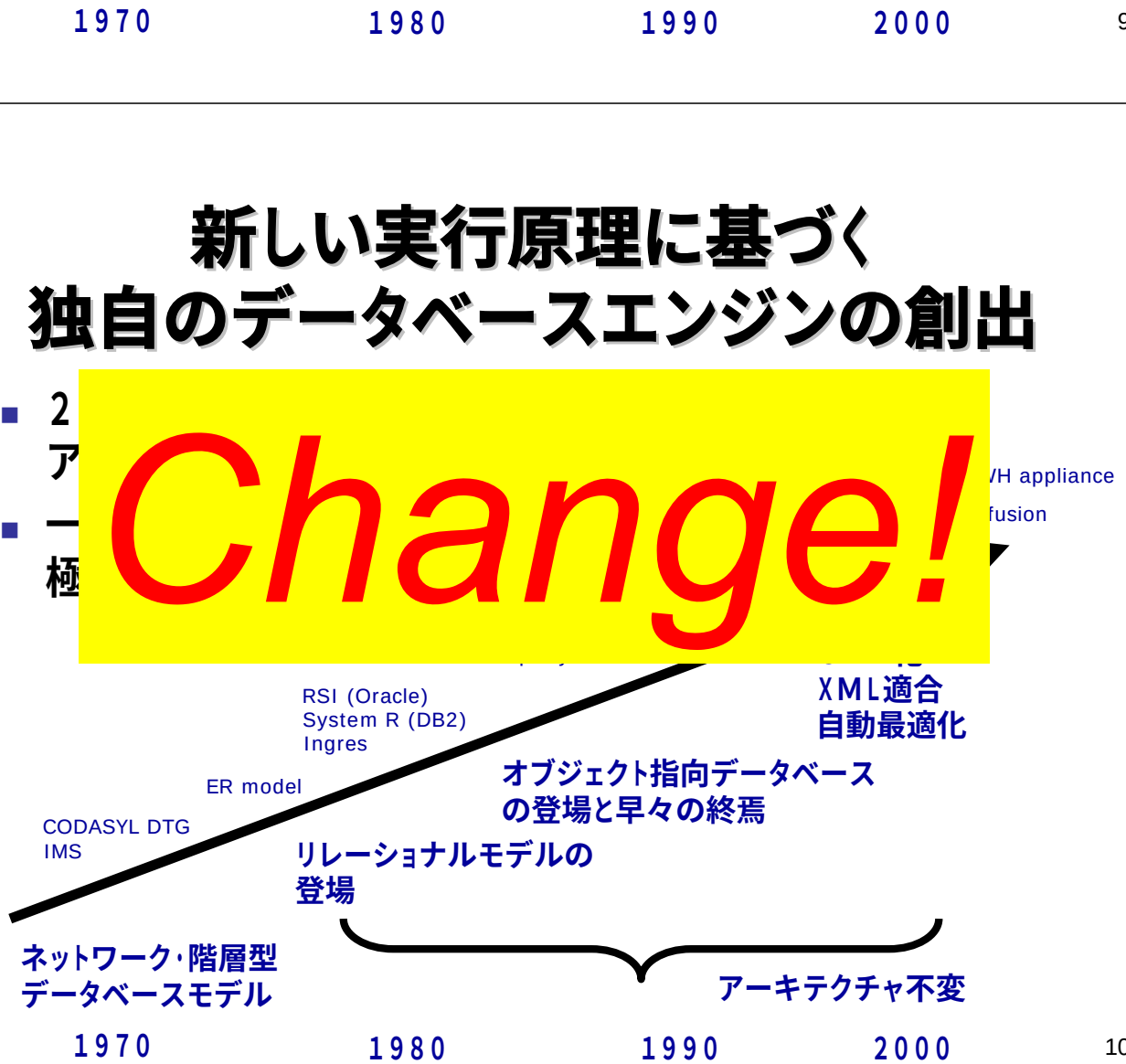
- 20年以上、リレーショナルデータベースエンジンの アーキテクチャには変革無し
- 一方、市場では 極めて重要な基幹ソフトウェア



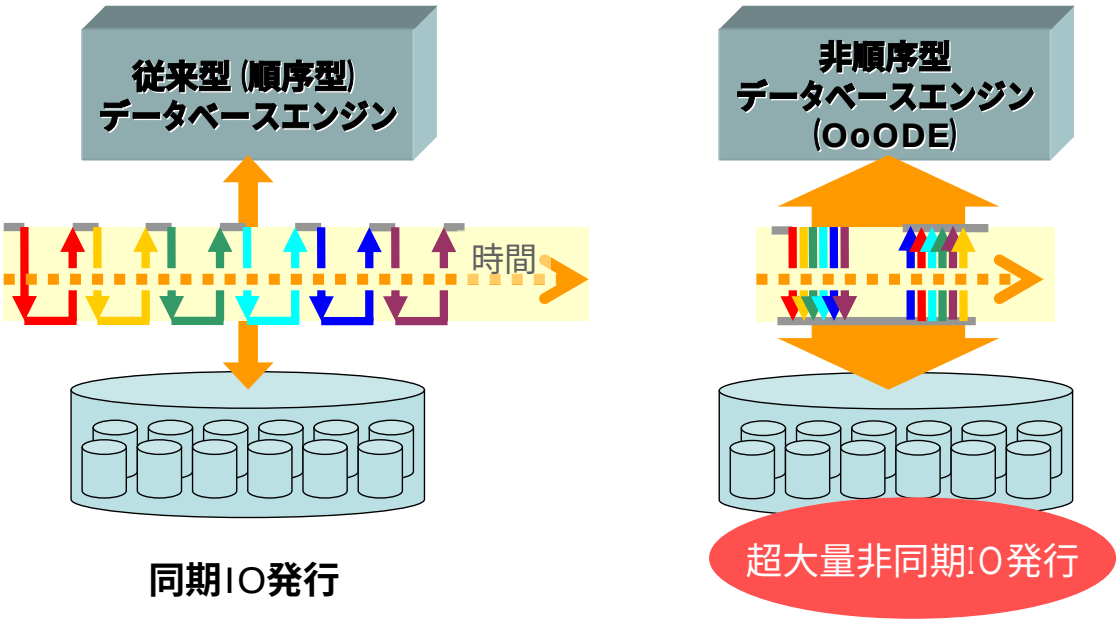
新しい実行原理に基づく 独自のデータベースエンジンの創出

- 2
- 一

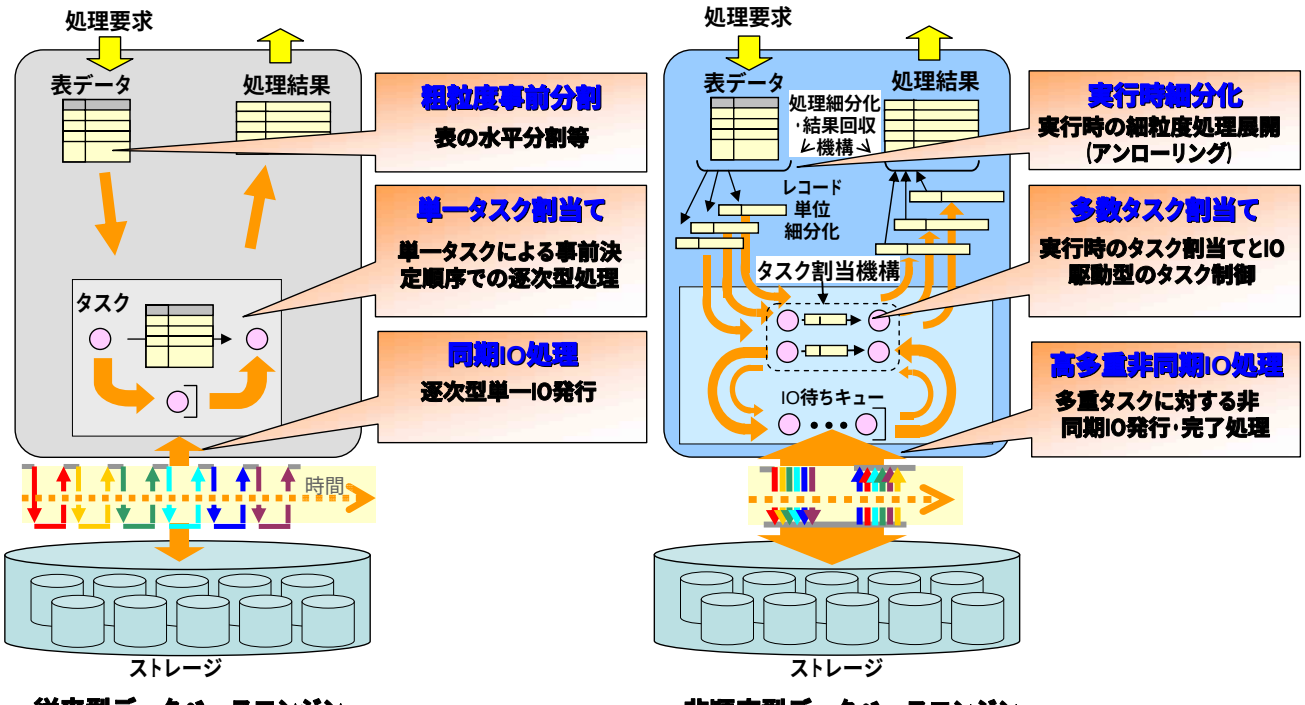
Change!



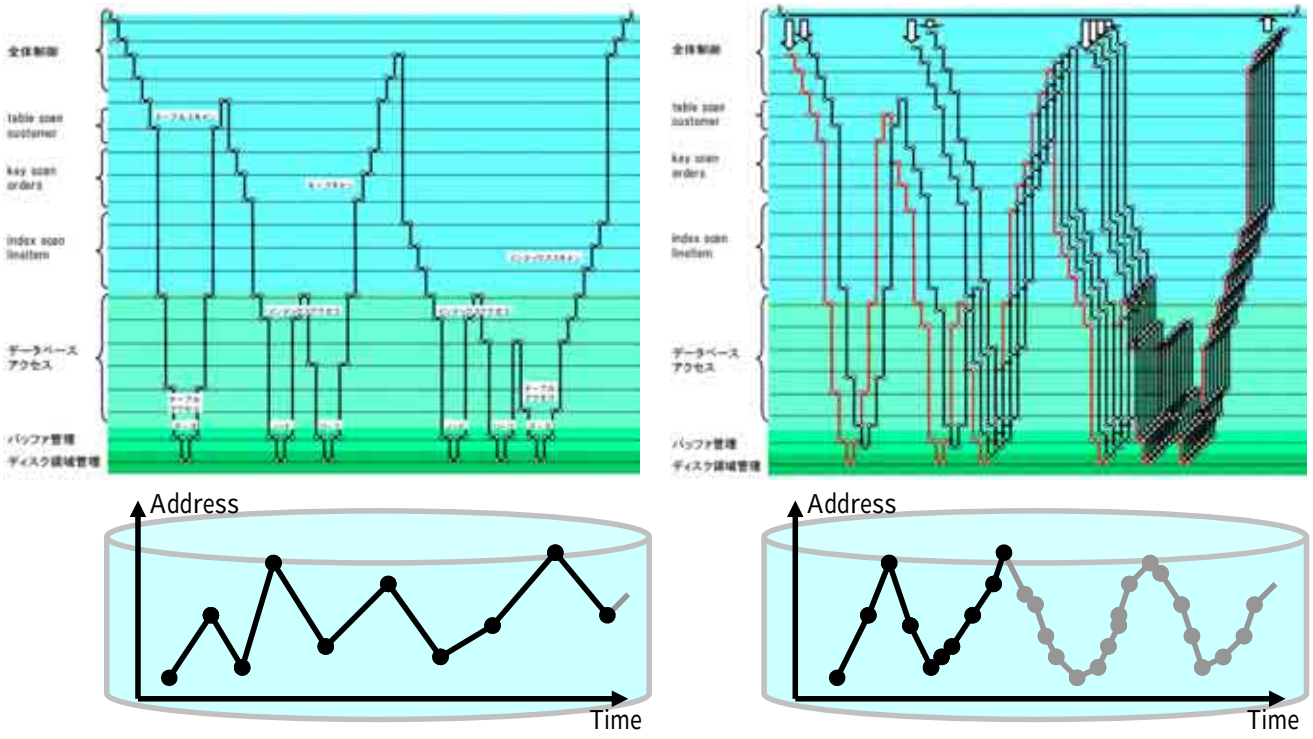
従来型データベースエンジンの問題と非順序型データベースエンジン



OoODEの基本構成



処理フロー及びストレージアクセス比較



従来型データベースエンジン　　非順序型データベースエンジン

研究開発計画

	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 基本設計	設計・一部実装	限定版OoODEの開発 (約10倍の性能向上を達成予定) 実装・評価 基本検討	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 高資源調整機構の開発 設計	高度化 実装
(3) OoODEのモニタリング技術に関する研究	設計	挙動モニタリング機構の開発 一部実装	実装・評価 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	基本調査	実証アプリケーションの検討		非常に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 設計・部分構築 詳細調査	全体構築 実証評価基盤の構築と実証実験 設計 構築・実証実験

研究構成

サブテーマ1：
OoODE技術に関する研究

サブテーマ2：
OoODEの資源調整技術に関する研究

サブテーマ3：
OoODEのモニタリング技術に関する研究

サブテーマ4：
OoODEの実証評価に関する研究

研究構成

サブテーマ1：
OoODE技術に関する研究

サブテーマ2：
OoODEの資源調整技術に関する研究

サブテーマ3：
OoODEのモニタリング技術に関する研究

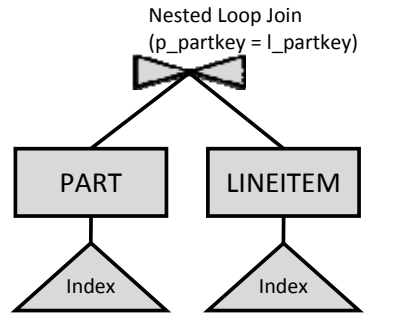
サブテーマ4：
OoODEの実証評価に関する研究

2方向からのOoODE化戦略

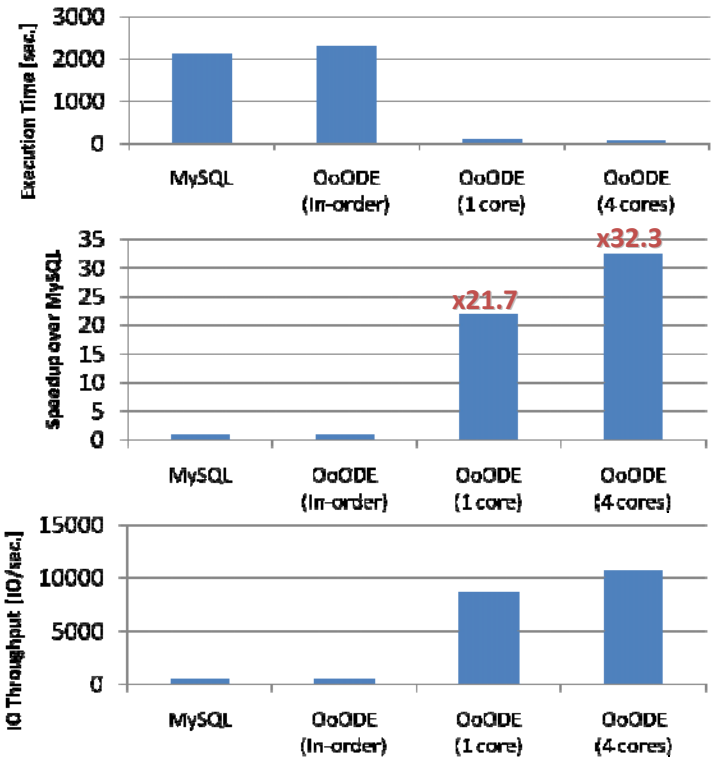
- 東大版 OoODE
  - オープンソースデータベースMySQLのOoODE化
  - マルチコア環境における多重スレッド機構の追求
- 日立版 OoODE
  - 商用データベースHiRDBのOoODE化
  - 現行コードへの組込みの難易度
  - トランザクション、再編成など他のデータベース機能への影響の精査
  - マルチスレッド化未対応、1コアでの実現

オープンソースデータベースMySQLにおけるマルチコア活用

TPC-Hベンチマーク
データセット：SF=100 (100GB)
問合せ：Q.14,17相当



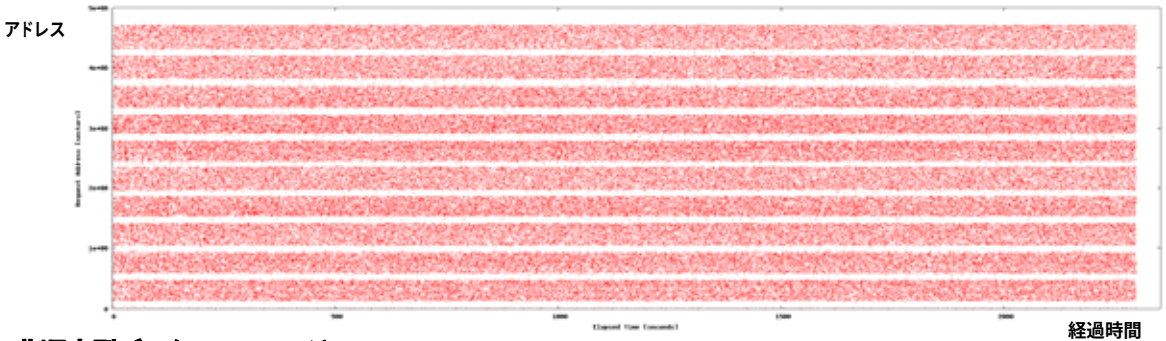
4x Quad-core Xeon Processors
32GB Memory (Only 1.6GB used)
4x FC HBAs (4Gbps)
20x 10Krpm 146GB FC HDDs (RAID-0)
RedHat Enterprise Linux 5.3
MySQL 5.1 InnoDB Storage Engine



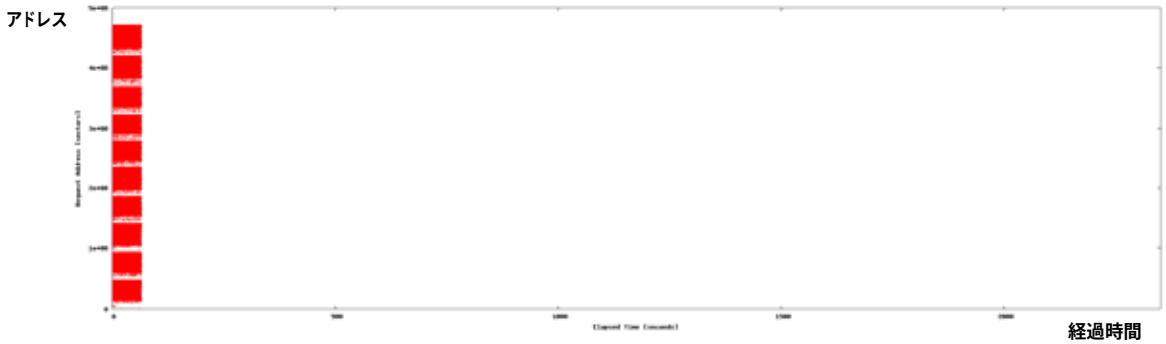
超大量IO発行による大幅な性能向上

順序型データベースエンジン

TPC-Hベンチマーク, SF=100, Q.14相当



非順序型データベースエンジン (マルチコア: 4)

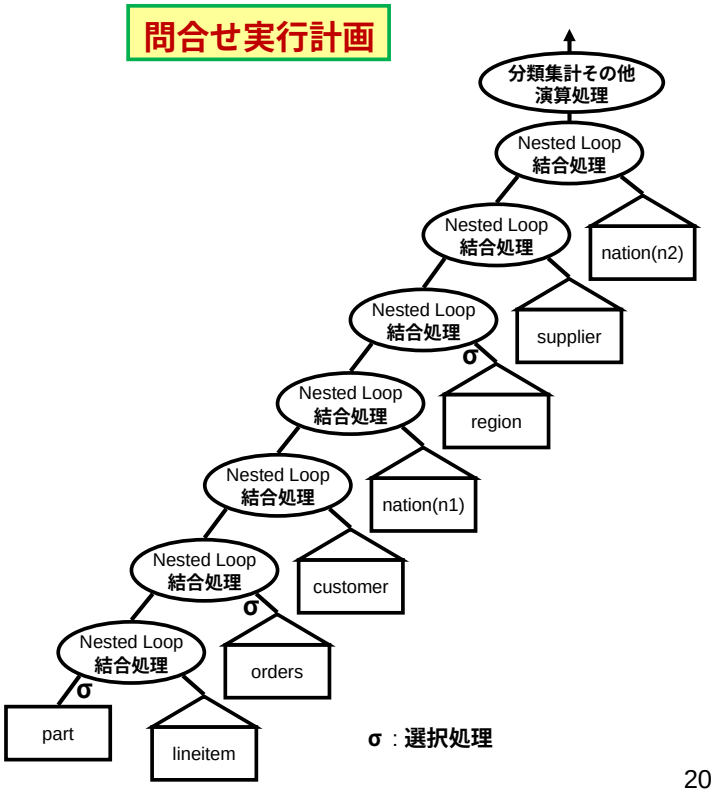


評価対象 TPC-H Q8相当問合せ

問合せ文 (SQL)

```
SELECT  o_year,
          sum(case when nation = 'BRAZIL'
                then volume else 0 end)
              / sum(volume) as mkt_share
FROM (
   SELECT
      extract(year from o_orderdate) as o_year,
      l_extendedprice * (1-l_discount) as volume,
      n2.n_name as nation
   FROM
      part, supplier, lineitem, orders, customer,
      nation n1, nation n2, region
   WHERE
      p_partkey = l_partkey
      and s_suppkey = l_suppkey
      and l_orderkey = o_orderkey
      and o_custkey = c_custkey
      and c_nationkey = n1.n_nationkey
      and n1.n_regionkey = r_regionkey
      and r_name = 'AMERICA'
      and s_nationkey = n2.n_nationkey
      and o_orderdate between date '1995-01-01'
                          and date '1996-12-31'
      and p_type = 'ECONOMY ANODIZED STEEL'
      and p_size < 3
   ) as all_nations
GROUP BY o_year
ORDER BY o_year;
```

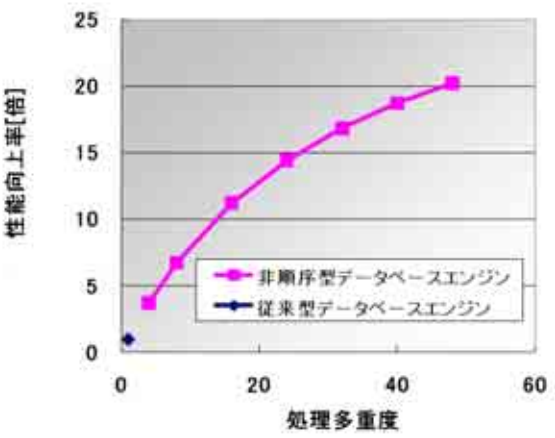
問合せ実行計画



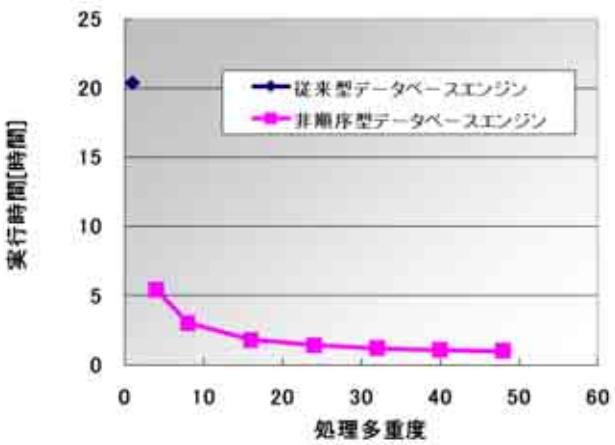
商用データベースHiRDBベースのOoODEプロトタイプにおける性能向上

**業界標準データベースベンチマークTPC-Hにおいて、
同一環境でソフトウェアのみの変更により約20倍の高速化を達成**

処理多重度に対する性能向上率の変化



処理多重度に対する性能向上率の変化



評価環境
CPU: Intel Xeon 2.66 GHz
メモリ: 32GB
HDD台数: 20台 (10krpm)

データ: TPC-H SF=1000 (約1TB)
処理: TPC-H Q8相当 (選択率:0.028%)

商用データベースHiRDBベースのOoODEプロトタイプの適用範囲

**頻繁に利用される基本的なデータベース処理に焦点を絞り設計・実装を実施中
(平成21年度内にTPC-Hの問合せ22個のうち半数の11個へ対応する機能限定版OoODEの開発を完了予定)**

TPC-H問合せ適用範囲 (総数22個)

機能限定版OoODE
(平成21年度内開発完了予定)
問合せ:11個
Q1, 3, 5, 6, 7, 8, 9,
Q10, 12, 14, 19

【基本処理】
単一表検索処理
ネステッドループ結合処理
分類集計処理
整列処理

本格版OoODE
問合せ:11個
Q2, 4, 11, 13, 15, 16,
Q17, 18, 20, 21, 22

【その他の処理】
副問合せ処理 (相関条件有・無)
集合演算処理
導出表処理
ハッシュ 結合処理

研究構成

サブテーマ1：
OoODE技術に関する研究

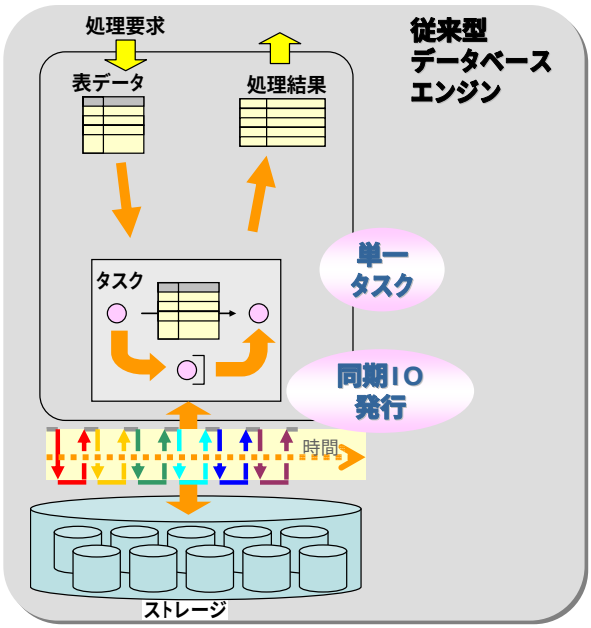
サブテーマ2：
OoODEの資源調整技術に関する研究

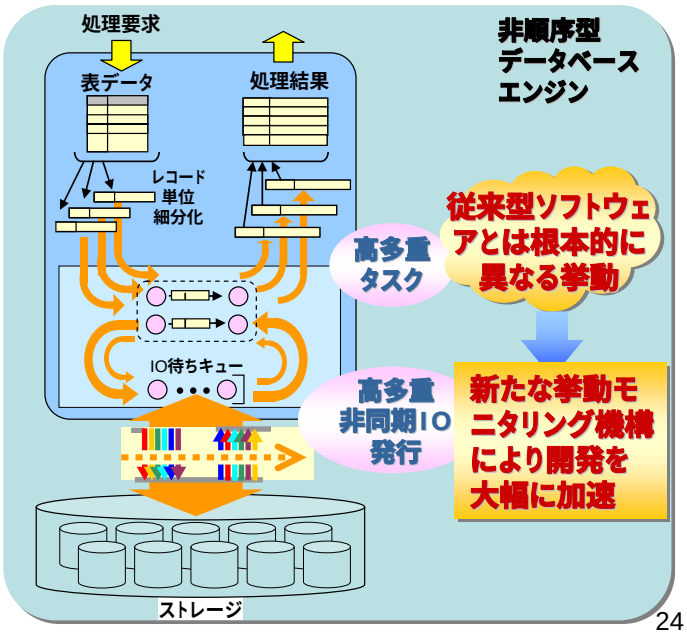
サブテーマ3：
OoODEのモニタリング技術に関する研究

サブテーマ4：
OoODEの実証評価に関する研究

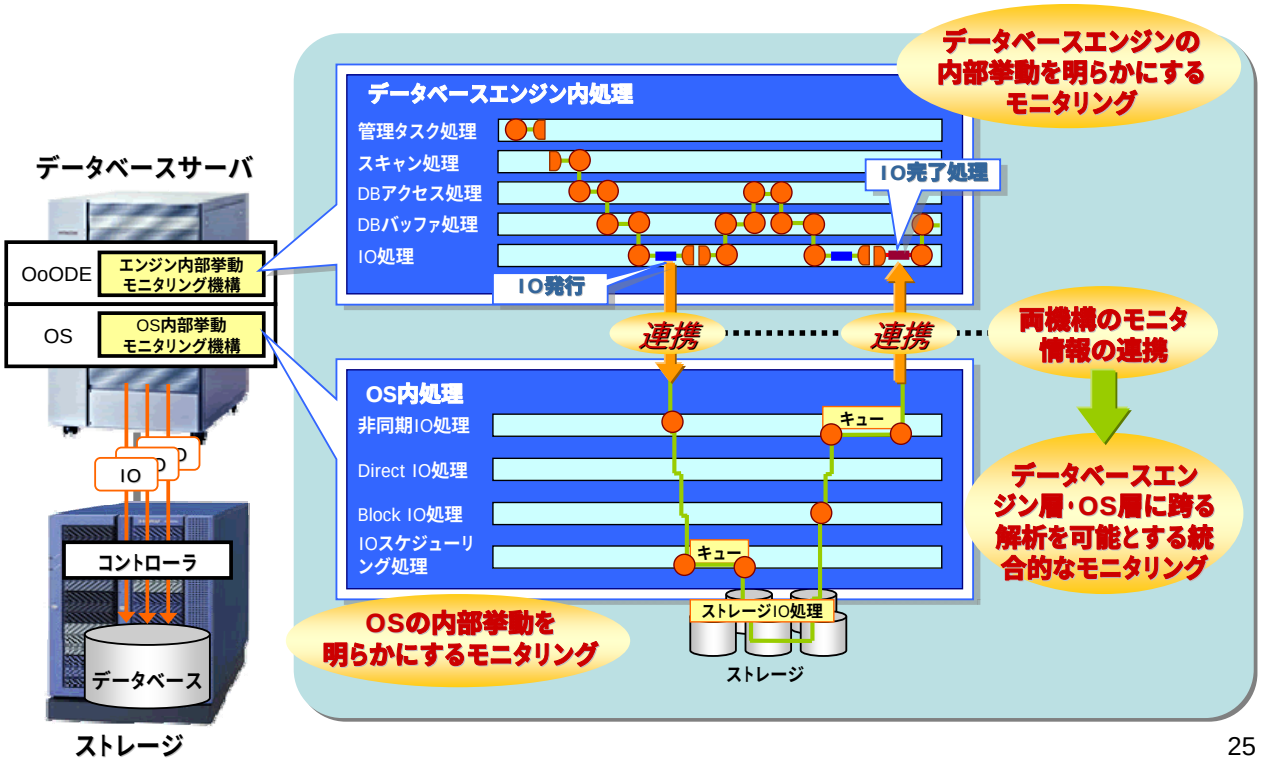
新たな挙動モニタリング技術の必要性

・非順序型データベースエンジンにおける処理動作は極めて非決定的
・高多重非同期IOに対するOS挙動は未解明
ソフトウェア内部挙動のモニタリング機構によりソフトウェア動作を解析
（データベース処理の正当性確認、システム性能デバッギング）



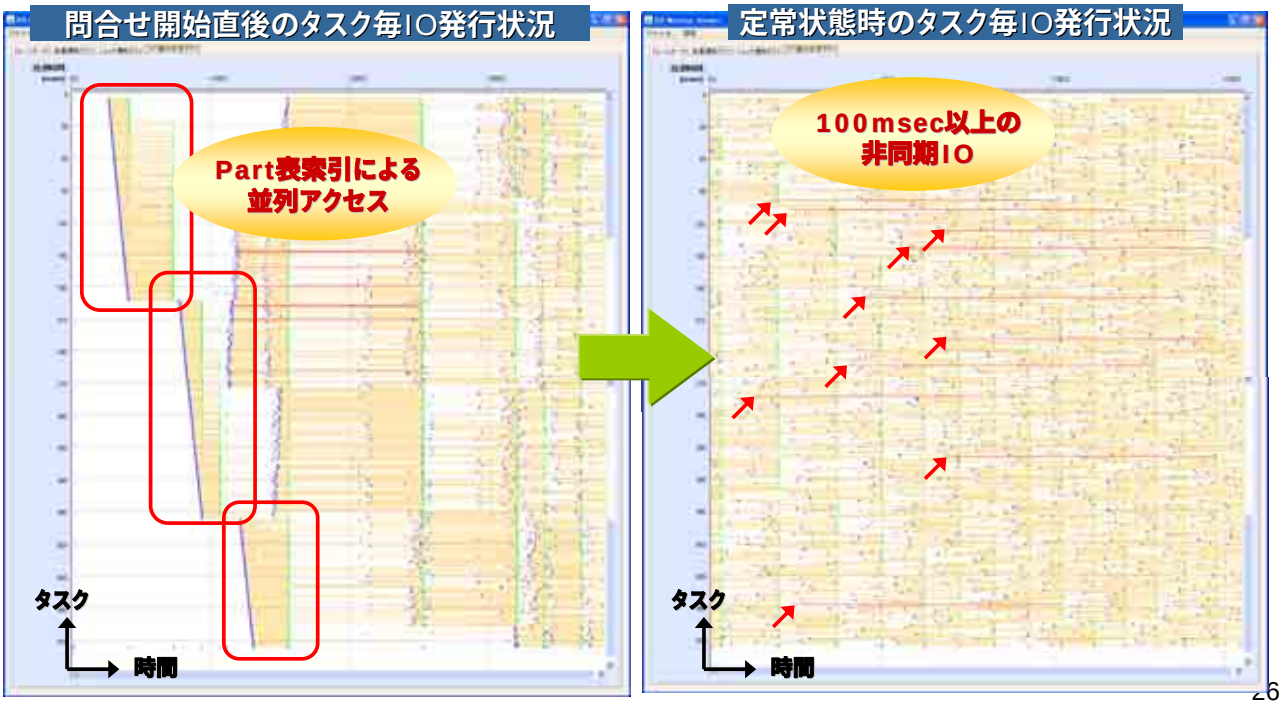


挙動モニタリング機構の概要



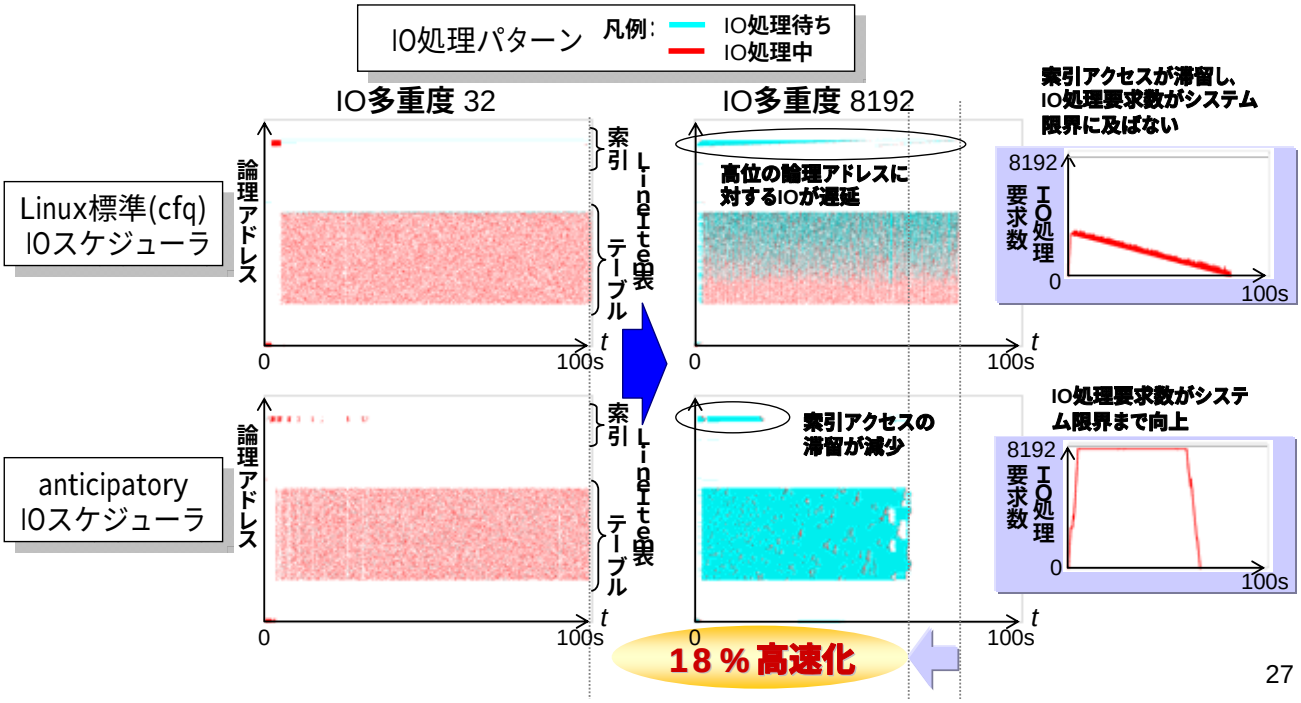
OoODE内部挙動の解析

- 問合せ開始直後：Part表索引によるPart表への並列アクセスが逐次的に開始
- 定常状態時：高多重非同期IO（所々IOに遅延を確認）



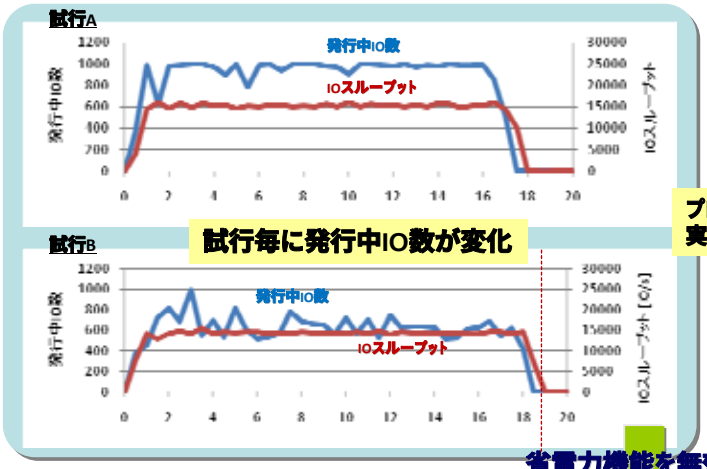
OS内部挙動の解析 (IOスケジューラ)

これまで未解明であったIOスケジューラの非同期IOに対する特異挙動を解明
ボトルネックの除去により18%問合せ性能を向上

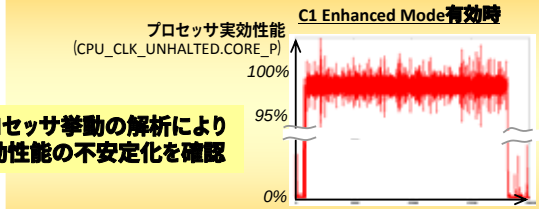


OS内部挙動の解析 (プロセッサ省電力化)

再現性の低い問合せ性能問題を解析、プロセッサ省電力化機能による影響を解明



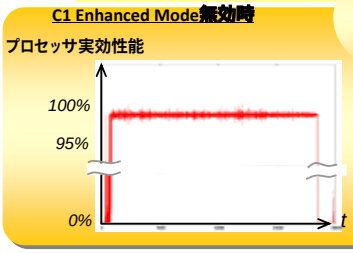
プロセッサ挙動を詳細解析
(コア動作のパフォーマンスカウンタ計測)



プロセッサ挙動の解析により実効性能の不安定化を確認

省電力機能(C1 Enhanced Mode)の副作用により、アイドル状態からの復帰後10us間に性能が低下
(Intel Xeon Processor 53xx資料)

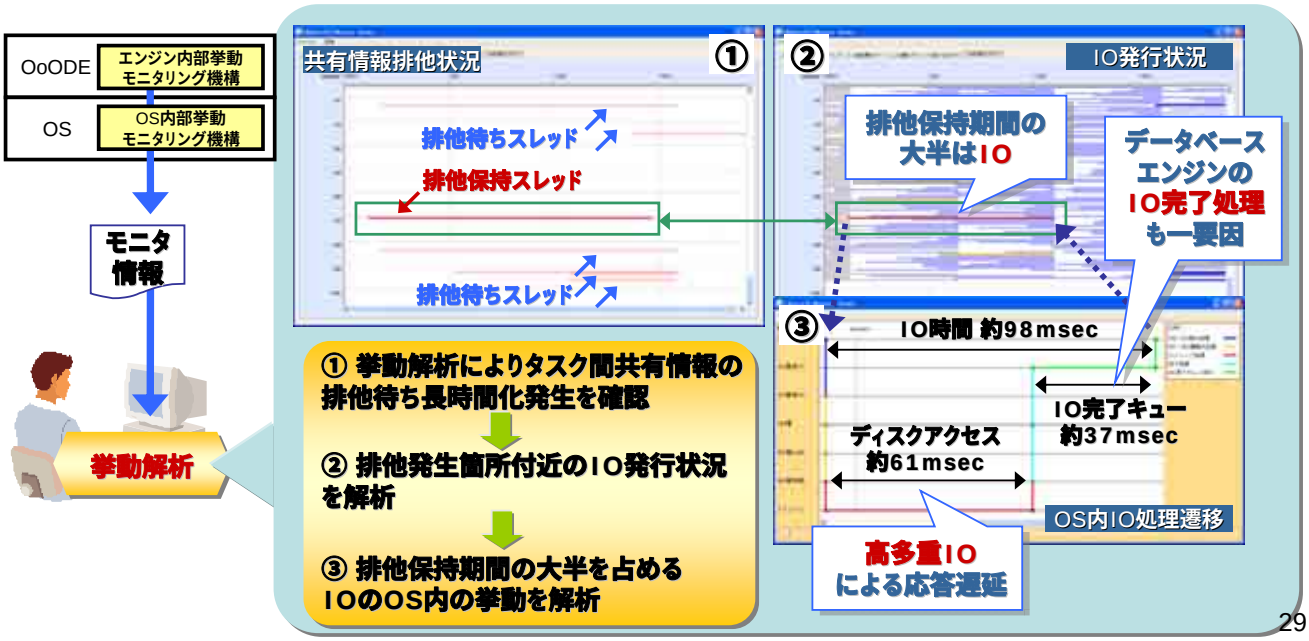
プロセッサ実効性能を安定化
発行中IO数を安定化



挙動不安定化の原因を解明

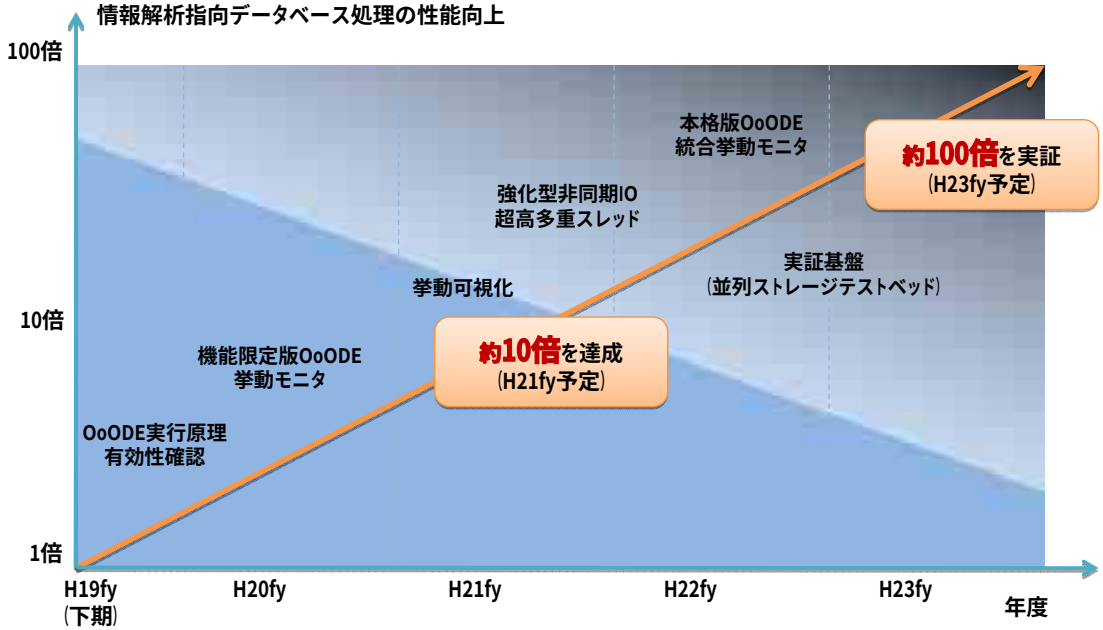
OoODE・OSの連携解析（基礎実験）

OoODEとOS双方の内部挙動モニタリング機構の取得情報を連携して解析し、タスク間の長い排他待ち現象の根本原因を解析
今後、ストレージシステムを含めた統合モニタリング機構・可視化機構として検討を深めたい

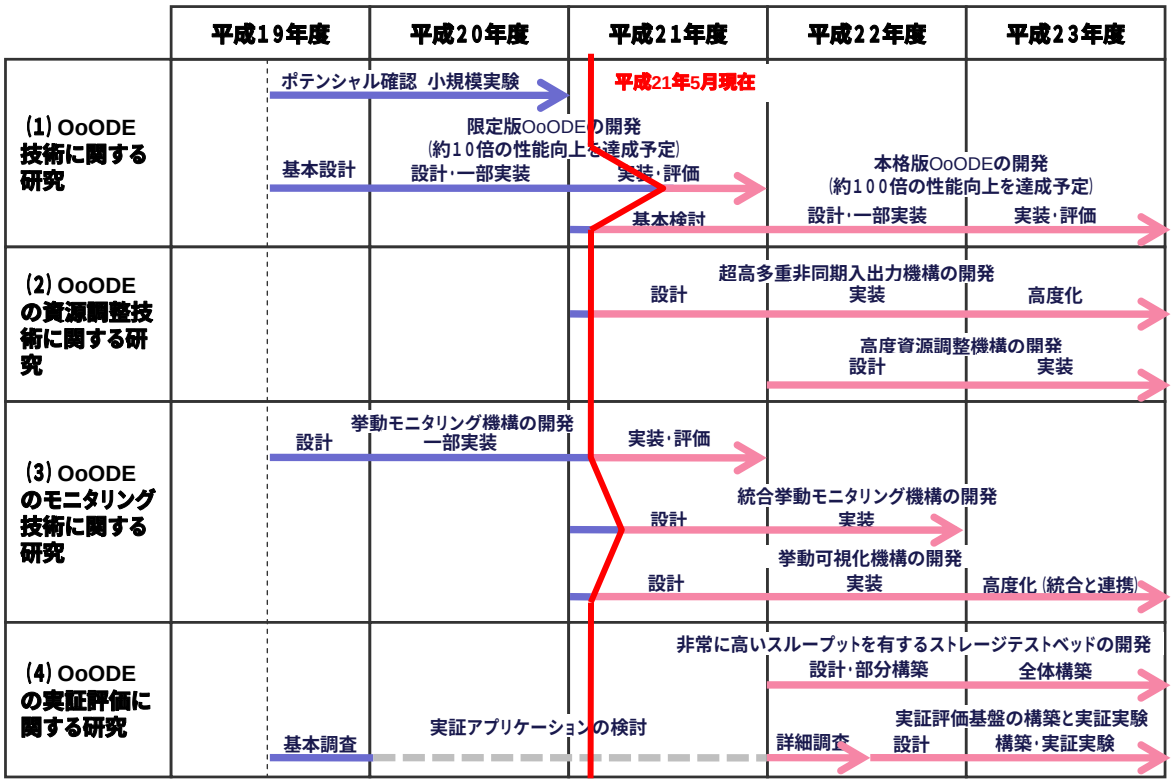


ロードマップ

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）



研究開発計画



研究構成

サブテーマ1：
OoODE技術に関する研究

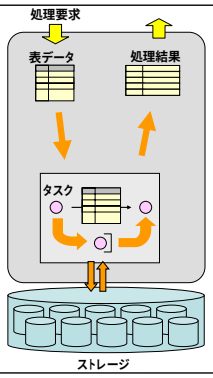
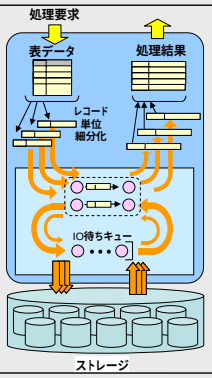
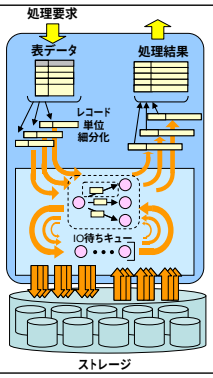
サブテーマ2：
OoODEの資源調整技術に関する研究

サブテーマ3：
OoODEのモニタリング技術に関する研究

サブテーマ4：
OoODEの実証評価に関する研究

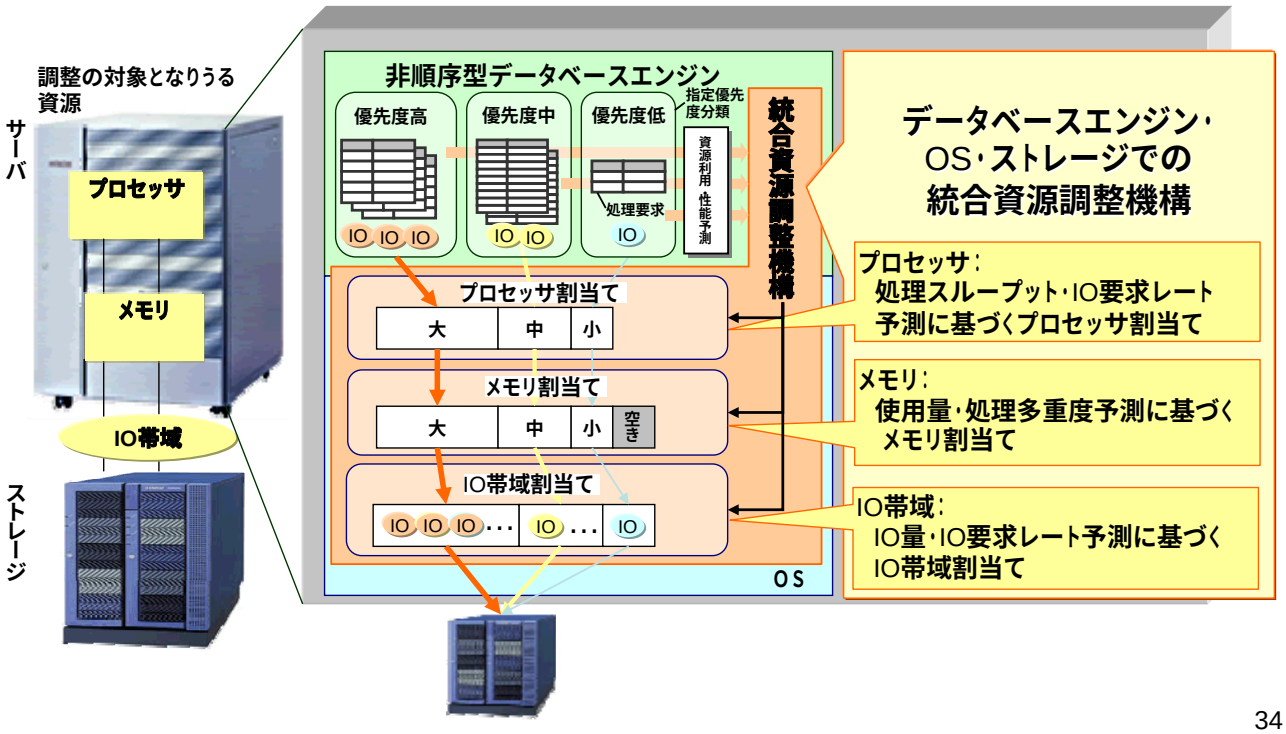
非順序型データベースエンジン技術の予定

- 機能限定版OoODE：解析系処理の約10倍高速化を予定（平成21年度）
- 本格版OoODE：解析系処理の約100倍高速化を実証予定（平成23年度）

	従来型 データベースエンジン	機能限定版非順序型 データベースエンジン	本格版非順序型 データベースエンジン
開発期間	・・・	平成19 ～ 21年度	平成21～ 23年度
概要			
非順序適用範囲	・・・	単表検索、NL結合、分類集計	副問合せ、その他演算処理
想定実行環境	・・・	中規模実験環境	大規模実証環境
処理多重度	1	最大数百規模	最大数千規模
性能向上率	1	10倍以上	100倍以上

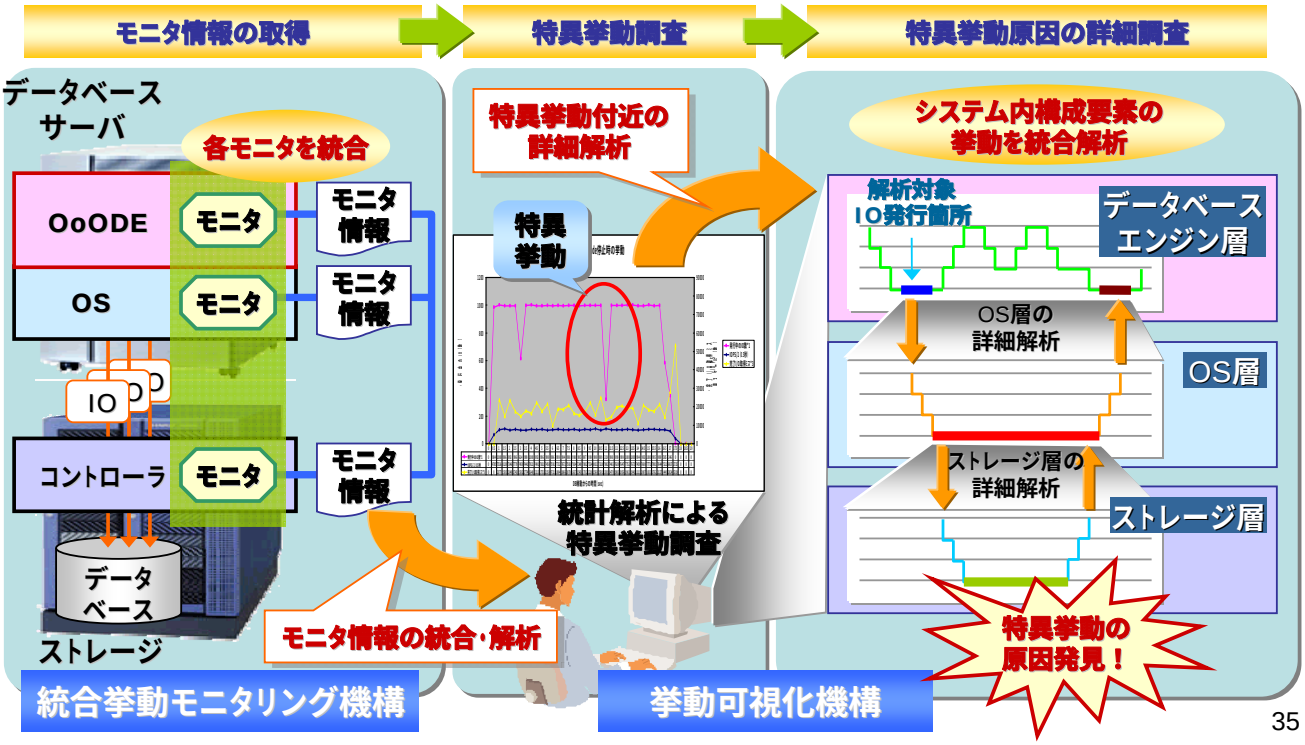
資源調整技術の予定

- 資源利用量・処理内容を基にした性能予測を利用して資源利用量・処理順序を調整、処理性能を制御



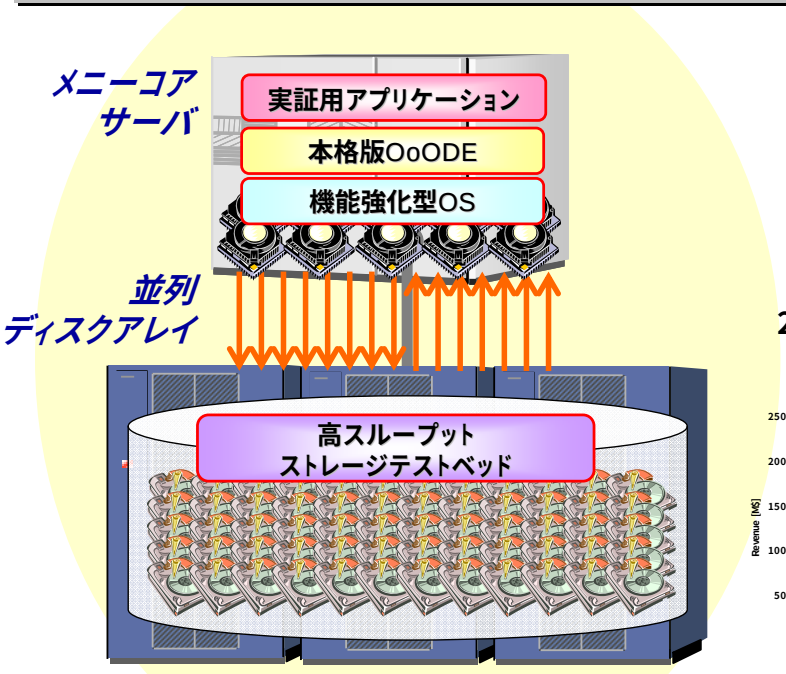
統合挙動モニタリング機構・挙動可視化機構の予定

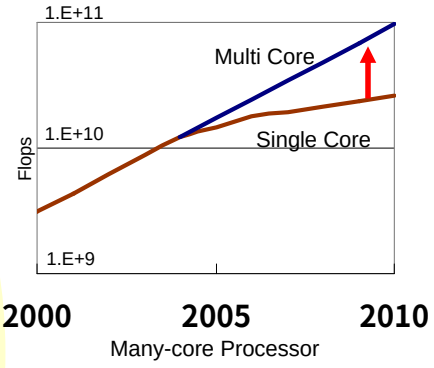
- データベース、OS、ストレージの各層のモニタリング情報を統合・可視化、システム内構成要素の特異挙動の原因調査時間を短縮

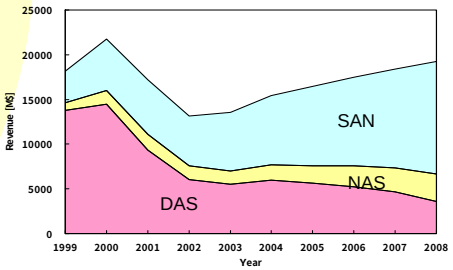


実証評価の予定

- 500TB級高スループットストレージによる実証環境を構築
  (業界標準ベンチマークTPC-H換算でSF=200K規模を想定)
- トレーサビリティ等の超巨大データ活用アプリケーションを用いた実証実験を実施







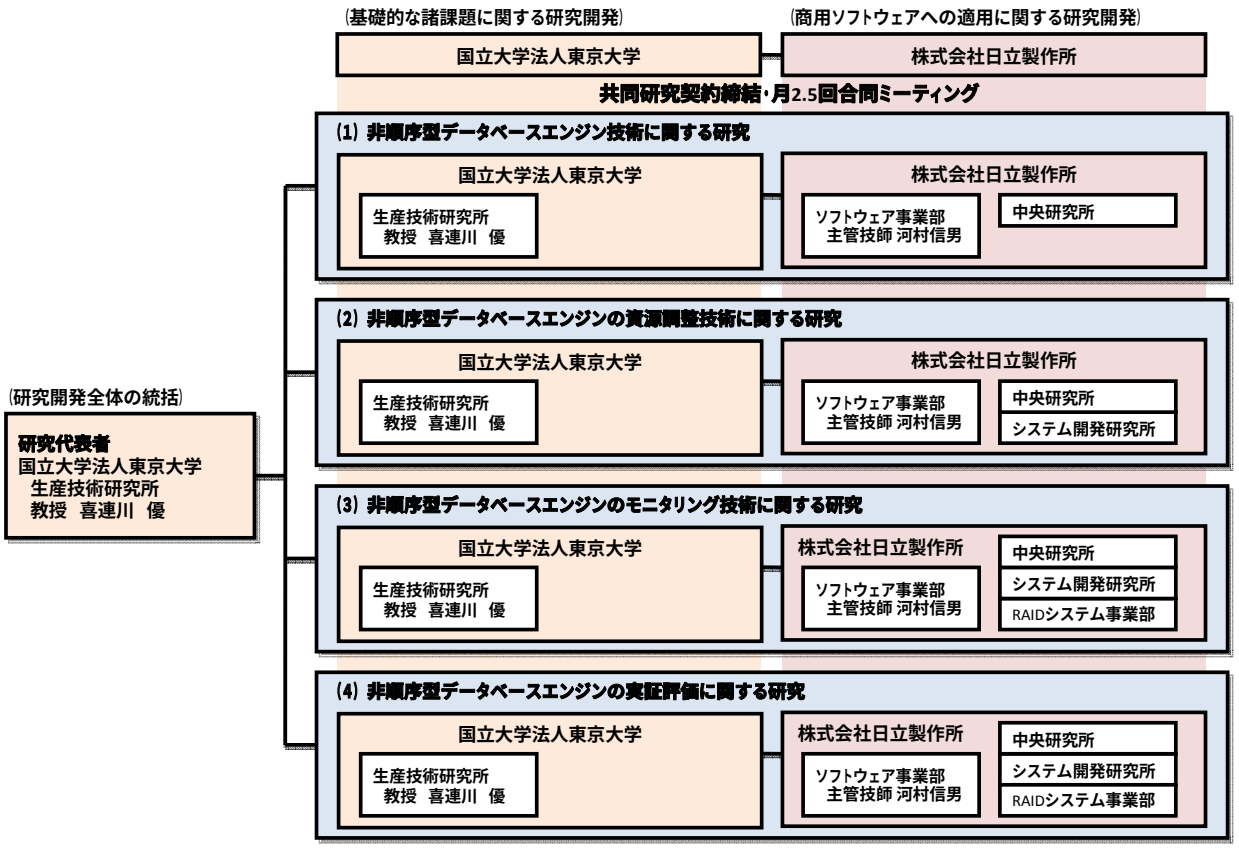
Storage Consolidation

独創性・優位性

- 従来とは全く異なるDBMSソフトウエアアーキテクチャの確立
  - 網羅的調査の実施、過去類似技術無し
  - 非順序型実行による非決定的動作
  - 実行時資源調整
- 2桁の大幅な性能向上の達成
- マルチコアプロセッサ指向戦略的ソフトウエア
- ＯＳ、ストレージを含めた統合的システム開発
  - 膨大な非同期IOスレッド管理
  - キャッシュ指向ではなく、スループット指向ストレージシステム

- 国内外DBMS（Oracle, IBM, Microsoft等）
  - 従来型アプローチ、むしろ上位アプリへの開発意欲
- DWHアプライアンス（Netezza, Oracle等）
  - 80年代のデータベースマシンソリューション、導入容易性の追求

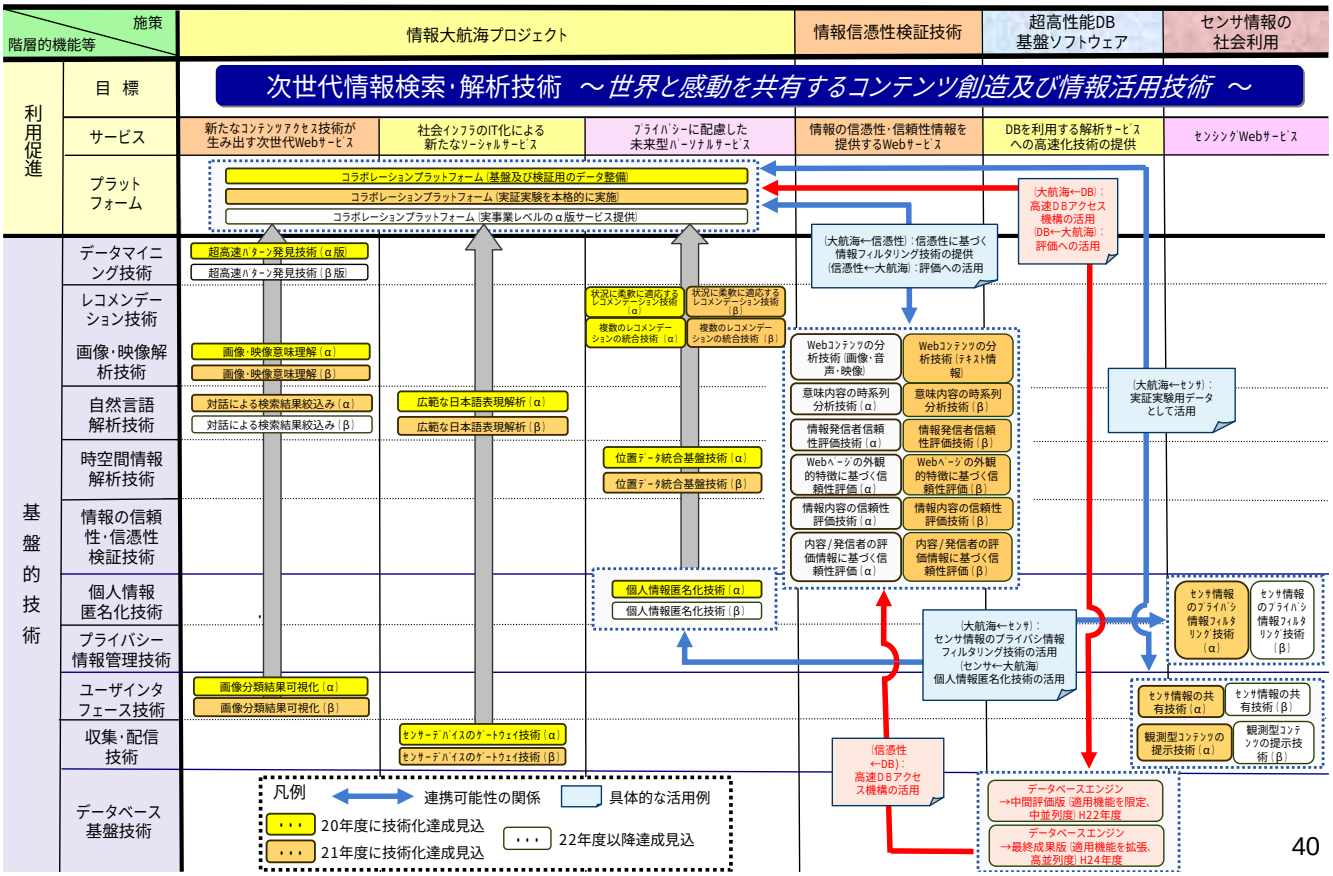
研究開発体制



研究開発体制

- 東京大学と、日立製作所における広範な部署が密接に連携
- 月２．５回程度の合同ミーティングを実施
  （日立ソフト、HGST等からも適宜参加）
- 大学研究室での共同実験、企業事業部における技術研修
- 研究開発プロジェクト会議（坂内NII所長ほか有識者）
- CSTP省庁間連携施策群「情報の巨大集積化と利活用」
  （毎年：連携シンポジウム）
  （年４回：連携タスクフォース）

他プロジェクトとの連携



成果の利活用

- 従来技術では困難であった応用への適応により「超巨大データの戦略的活用」に挑戦
  - 多様な社会的・経済的な波及効果
  - データベース基盤ソフトウェアは巨大かつ高成長市場
    （平成１９年度世界市場規模：1兆８８００億円、年率１２．６％成長、IDC出典）
  - マルチコア時代に向けて非順序型データベース実行処理系を先駆的に開発
  - 高次の情報処理を実現し、我が国の国際競争力の向上に寄与、
- 最終成果の一部製品化を予定
- 実証実験を通じた成果アピールを予定
  - 超巨大データ活用型アプリケーションを用いた実証実験により約１００倍の性能向上を実証予定
  - 総合小売業購買情報分析システム、流通トレーサビリティなどを調査中
- 標準技術への配慮
  - OoODEの適用には基本的にアプリケーションの改編は一切不要
- 知的財産権への取り組み
  - １件の特許を出願済み、１０件を出願準備中

人材育成

- 大学の若手研究者
  - 特任教員等4名が参加
  - 合同ミーティングに参加（技術検討・評価への深く参画、戦略推進へも関与）
  - 企業事業部における技術研修に参加
  - ➔基盤的な基礎研究から産業的な応用開発までを俯瞰することのできる高度なリーダを育成
    - 大学・企業間の知の還流を促進
    - 大学における研究成果の産業界への迅速な展開
- 企業の研究者・技術者
  - 広範な部署（ソフトウェア事業部、システム開発研究所、中央研究所、RAIDシステム事業部）から 15人が参加
  - 大学研究室で共同実験を実施、共同論文を執筆中
  - ➔ 企業内研究者・技術者の学位取得、トップレベル国際会議への参加
    - 我が国企業の国際プレゼンスの向上