ストレージ消費電力特性に基づく関係データベース演算子の省電力指向コストモデル

早水　悠登† 　合田　和生† 　喜連川　優†,††

† 東京大学生産技術研究所　〒 153–8505 東京都目黒区駒場 4–6–1
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †{haya,kgoda,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　昨今の IT 基盤に於いては少なからぬ電力がストレージにより消費されており，データセンタの省電力化においてストレージの消費電力制御の重要性は高い．本論文では，消費エネルギーを考慮した問合せ最適化の枠組みにおいて，ストレージ構成に応じた消費エネルギーコスト推定手法を示し，幅広いストレージ構成に対する有効性を実験的に評価した．また，当該コスト推定を著者らが研究に取り組むアウトオブオーダ型データベースエンジンへと適用し，従来型の問合せ実行に対するエネルギー効率を実験的に評価した．

キーワード　問合せ最適化, ストレージ, 消費電力, 消費エネルギー, コストモデル

1. はじめに

ビッグデータ時代の到来が喧伝されて久しく，爆発的に増加を続けるデータの積極的な利活用は，今日の IT システムにおいて必須の要素となりつつある．データの格納・管理を担うストレージ資源を中心とし，大規模データ活用を支えるためのデータセンタにおける IT 資源は拡大を続けている．一方で，その消費エネルギー増大が近年特に問題視されており，2013 年には米国のみで大規模発電所 34 基分に相当する 910 億 kWh ものエネルギーを消費しており，2020 年までに 1,390 億 kWh に増大すると予測されている [1]．昨今大きな注目を集めている IoT や人工知能技術に限らず，今日の IT 技術はデータ処理能力の拡大をドライバとして目覚ましい発展を続けているが，それに伴う急速な消費エネルギー増大を継続することは困難になる一方であることから，データセンタにおける IT システムの省エネルギー化は極めて重要な課題であるといえる．

本論文では，IT システムの中でもとりわけデータ処理の中心的役割を果たすストレージシステム，およびその活用を担う関係データベースシステムに着目した省エネルギー化に着目したい．関係データベースシステムにおいて，ある問合せ処理の実行方法は通常複数存在することから，問合せ最適化によって，利用可能な資源制約の下で実行コストが最小と見込まれる実行方法を選択する，所謂コストベース最適化の枠組みが広く用いられている．従前のデータベースシステムにおいて，殆どの場合問合せ最適化の指標は問合せ実行時間であり，即ち利用可能な資源を投入して最も実行時間の短くなる実行方法が選択されるよう，問合せのコスト推定手法が設計されてきた．これに対し，著者らは文献 [2] に於いて，ストレージの消費エネルギーコスト推定手法を核とする，消費エネルギーを考慮したコストベース最適化手法の枠組みを示し，評価実験により関係データベースの基本演算の消費エネルギーコストを良好な精度で推定可能であることを確認するとともに，当該コスト推定に基いたストレージシステムの能動的なエネルギー制御の有効性を示した．本論文では，ストレージ構成の変化が消費エネルギーコスト推定に与える影響を考慮するよう当該コスト推定手法を拡張し，幅広いストレージ構成に対して消費エネルギーコスト推定の適用性があることを示す．また，当該コスト推定を著者らが研究に取り組むアウトオブオーダ型データベースエンジンへと適用し，従来型の問合せ実行に対する省エネルギー性を実験的に評価する．

本論文の構成を以下に示す．2 節では，関連研究についてまとめる．3 節では，ストレージの消費エネルギーを考慮した問合せ最適化ならびにストレージの消費エネルギーコスト推定手法の概要を示し，4 節においてストレージ構成に応じた消費エネルギーコスト推定手法について説明し，5 節では，高精度電力計を備えた実験環境における評価実験について説明し，6 節で本論文をまとめる．

2. 関連研究

データベースシステムの省エネルギー化に関しては，データセンタにおける電力消費量の急速な増加傾向を米国環境保護庁が文献 [3] において報告した時期に前後して，データベースシステムの省エネルギー化の重要性が文献 [4] などにおいて指摘され，以降その議論が徐々に本格化を始めた．データベースシステムにおける消費電力の特性を明らかにするため，文献 [5] ではトランザクション処理の消費電力特性，文献 [6] では分析系処理に着目した消費電力特性の分析が行われており，また文献 [7] では問合せ実行計画を構成する演算子やシステムコンポーネント毎の電力消費傾向の分析が行われている．文献 [8] では，ストレージ構成に応じた性能と消費エネルギーの関係を実験的に確認し，その結果を基にデータベースシステムのソフトウェアアーキテクチャを見直すことで，大きな省エネルギー化の余地が見いだせることを指摘している．省電力化のアプローチとしては，文献 [9] ではプロセッサの周波数スケーリングを用い

て，アプリケーションにおける性能要求を満たしながら消費電力を削減する手法について議論しており，文献 [10] では周波数スケーリングと複数の問合せにおける共通処理集約と組み合わせた省エネルギー化手法を提案している．またミッションクリティカルなシステムでは，アクティブスタンバイしている 2 次系データベースシステムが少なからぬ電力を消費することに着目したリモートレプリケーションにおけるエネルギー効率技法なども見られる [11]．データベースシステムのエネルギー効率の評価については，TPC ベンチマーク [12] が性能あたりの消費電力の指標を採用しているほか，文献 [13] においてソートアルゴリズムのエネルギー効率を評価するためのベンチマークが提案されるなど一定の取り組みが見られる．

問合せ最適化に関する省エネルギー化の議論としては，先駆的な取組みとしては 20 年以上前から文献 [14] などが見られるが，多くは携帯端末など電源制約の厳しい環境を対象としたものであり，データセンタにおいて展開される規模でのデータベースシステムが論じられるようになったのは，主に 2000 年代以降である．文献 [15] では，データベースシステム省エネルギー化が見込まれる領域として，問合せ最適化に加え，演算・入出力資源スケジューリングなどが指摘されている．文献 [14] では，消費電力と問合せ処理性能のトレードオフについてモデル化し，実験によりその評価を行っている．また，性能と消費電力のトレードオフの調整に関しては，文献 [16] では，定められたサービスレベルの範囲で省エネルギー化の余地をするアプローチについて議論し，ERP なる指標を提唱し，当該指標により問合せ最適化を省エネルギー指向とする枠組みを提案している一方，文献 [17] に見られるように，データベース管理者がトレードオフを指定可能とするアプローチもみられる．

3. ストレージの消費エネルギーを考慮した問合せ最適化

3.1 データベースシステムにおける問合せ最適化

本設における問合せ最適化の定式化は文献 [2] に基づく．本論文の議論の前提とする範囲を改めて以下に説明するが，仔細については文献 [2] を参照されたい．

関係データベースシステムでは，ユーザからの問合せは SQL 等の宣言的言語によって記述される．データベースシステムは実行の具体的な手続きに対応する問合せ実行計画を生成し，実行計画に従って問合せ処理を駆動する．問合せ処理におけるリレーション走査方法や結合順序などにより，一般に 1 つの問合せに対して問合せ実行計画は一意に定まらないことから，データベースシステムは問合せ最適化によって，最も望ましいと考えられる問合せ実行計画を選択する．初期のデータベースシステムにおいては，実装の容易性からヒューリスティックに規定されたルールに基いて選択するルールベース最適化がしばしば用いられたが，昨今のデータベースシステムにおいては問合せ実行計画のコストを推定し，コストが最小となる実行計画を選択するコストベース最適化が広く用いられている．従前のデータベースシステムにおいては，コストの指標は専ら問合せ実行時間であり，問合せ最適化は次の問題に言い換えられる．

実行時間最小化問題:

$$\begin{aligned} &\text{obj.} \quad \tau(p) \to \min \\ &\text{s.t.} \quad p \in \mathcal{P}(q) \end{aligned}$$

ここで，$\mathcal{P}(q)$ は問合せ q を処理可能な実行計画の集合であり，$\tau(p)$ は実行計画 p の実行に要することが推定される時間である．

これに対し，消費エネルギーを最適化の指標に新たに加えることで，様々な問合せ最適化のバリエーションが考えられる．例えば，利用可能な消費電力の上限 W や，消費電力量の E が与えられた場合に，問合せ実行に要する実行時間 $\tau(p)$ を最小化する最適化は次のように規定できる．

エネルギー制約のある実行時間最小化問題:

$$\begin{aligned} &\text{obj.} \quad \tau(p) \to \min \\ &\text{s.t.} \quad p \in \mathcal{P}(q), e(p) \leqq E, w(p) \leqq W \end{aligned}$$

ここで，$e(p)$ は実行計画 p の実行に要すると推定される消費電力量，$w(p)$ は実行計画 p の実行に要することが推定される最大消費電力を表す．

また，問合せ実行時間の上限 T が与えられた場合に，消費電力量 $e(p)$ を最小化する問合せ最適化は次のように規定できる．

消費電力量最小化問題:

$$\begin{aligned} &\text{obj.} \quad e(p) \to \min \\ &\text{s.t.} \quad p \in \mathcal{P}(q), \tau(p) \leqq T, w(p) \leqq W \end{aligned}$$

3.2 基本的な実行計画ブロックに基づく消費エネルギーコスト推定

問合せ実行計画は，関係データベース演算子を節とする木構造によって表現されるものとする．問合せ実行計画の木構造において，辺は関係データベース演算子同士の入力・出力関係を表す．関係データベース演算子は，入力データに対して逐次実行が可能であるパイプライン動作可能な演算子と，入力データが一定量蓄積されないと実行が開始できないブロッキング演算子の 2 つに大別できる．ここで，問合せ実行計画の木構造を，ブロッキング演算子に対する入力に対応する辺において分割して生じた部分木を実行計画ブロックと称することとする．即ち，問合せ実行計画 p は，1 つ以上の実行計画ブロック p_{bi} から構成され，p_{bi} はいずれもパイプライン動作可能な演算子の集合として規定される．

$$p_{bi} = p_{b0}, p_{b1}, \cdots, p_{b(n-1)}$$

本論文においては，問合せ実行計画 p の実行は，実行計画ブロック $p_{b0}, p_{b1}, \cdots, p_{b(n-1)}$ を直列に実行するものとして以降の議論を行う．この場合，問合せ実行計画 p のコスト推定は次に示す性質を満たすことから，各実行計画ブロック p_{bi} のコスト推定へと問題を分割可能である．

$$\tau(p) = \sum_{i=0}^{n-1} \tau(p_{bi}), e(p) = \sum_{i=0}^{n-1} e(p_{bi}), w(p) = \max_{i=0}^{n-1} w(p_{bi})$$

即ち，p_{bi} を構成する関係データベース演算子の種類毎にコスト推定手法を与えることで，各実行計画ブロック p_{bi} のコスト推定を行うことができる．ここでは，基本的な実行計画ブロックとして次の 4 種類のケースについて，実行時間，消費電力量，最大消費電力のコスト推定手法について示す．

（1） **FTS**: 全表走査による単一関係表 R の選択 $\sigma(R)$

$$\tau(p_{bi}) = ||R|| \cdot \lambda_R^{seq}$$
$$e(p_{bi}) = \tau(p_{bi}) \cdot w(p_{bi})$$
$$w(p_{bi}) = \sum_{d \in \mathcal{D}} \Omega_d^{seq} \left(\frac{||R_d||}{||R||} \cdot \frac{B}{\lambda_R^{seq}} \right) + w_c$$

（2） **IS**: 二次索引走査による単一関係表 R の選択 $\sigma(R)$

$$\tau(p_{bi}) = \zeta_\sigma \cdot (1 + c_a) \cdot |R| \cdot \lambda_R^{rnd}$$
$$e(p_{bi}) = \tau(p_{bi}) \cdot w(p_{bi})$$
$$w(p_{bi}) = \sum_{d \in \mathcal{D}} \Omega_d^{rnd} \left(\frac{|R_d|}{|R|} \cdot \frac{1}{\lambda_R^{rnd}} \right) + w_c$$

（3） **HJ**: 全表走査による関係表 R, S のハッシュ結合 $R \bowtie S$ 省略する (注1)

（4） **NLJ**: 二次索引走査による関係表 R, S のネステッドループ結合 $R \bowtie S$

$$\tau(p_{bi}) = \zeta_\sigma \cdot (1 + c_a) \cdot |R| \cdot (\lambda_R^{rnd} + j_{\bowtie} \cdot \lambda_S^{rnd})$$
$$e(p_{bi}) = \tau(p_{bi}) \cdot w(p_{bi})$$
$$w(p_{bi}) = \sum_{d \in \mathcal{D}} \Omega_d^{rnd}(\theta_i) + w_c$$
$$\theta_i = \frac{|R_i|}{|R|} \cdot \frac{1}{\lambda_R^{rnd} + j_{\bowtie} \cdot \lambda_S^{rnd}} + \frac{|S_d|}{|S|} \cdot \frac{j_{\bowtie}}{\lambda_R^{rnd} + j_{\bowtie} \cdot \lambda_S^{rnd}}$$

表 1 に各コスト推定手法における変数をまとめる．

表 1　コスト推定手法における変数の一覧

変数	説明
$\mathcal{D}$	データベースシステムを構成するストレージデバイスの集合
d	ストレージデバイス
$\|\|R\|\|$	関係表 R を構成するページ数
$\|\|R_d\|\|$	関係表 R を構成するページのうち，ストレージデバイス d に属するものの数
$\|R\|$	関係表 R のタプル数
$\|R_d\|$	関係表 R のタプルのうち，ストレージデバイス d に属するものの数
ζ_σ	選択 σ の選択率
$j_{\bowtie}$	結合 $\bowtie$ の結合増幅率
c_a	二次索引等の補助データ構造へのアクセスコストに掛かる計数
B	ページ長
λ_R^{seq}	シーケンシャルアクセスにより関係表 R を二次記憶から読み出す際の 1 ページあたりの平均応答時間
λ_R^{rnd}	ランダムアクセスにより関係表 R を二次記憶から読み出す際の 1 ページあたりの平均応答時間
$\Omega_d^{seq}(\cdots)$	ストレージデバイス d をシーケンシャルアクセスによって二次記憶から読み出す際の消費電力．ただし d が停止している場合はその消費電力
$\Omega_d^{rnd}(\cdots)$	ストレージデバイス d をランダムアクセスによって二次記憶から読み出す際の消費電力．ただし d が停止している場合はその消費電力
w_c	ストレージデバイスによらないファンや周辺回路等による固定的な消費電力

4. ストレージ構成を考慮した消費エネルギーコスト推定手法の拡張

昨今のデータセンタにおいては，ストレージシステムの高密度化が顕著であり，多数のドライブを高密度に搭載したストレージ環境は一般的となりつつある．また，クラウドの急速な普及や所謂 Software Defined Storage 技術の隆盛などにより，データベースシステムが利用可能なストレージ構成は非常に幅広いものとなってきている．これらのストレージ資源を，消費エネルギーの観点から活用する技法を議論する端緒として，本節ではデータベースシステムを構成するストレージデバイスの集合 $\mathcal{D}$ を変数として捉えた場合の，消費エネルギーコスト推定手法の拡張について議論する．

4.1 FTS における消費エネルギーコスト推定手法の拡張

ストレージデバイスに対してシーケンシャルアクセスを行う場合，OS やストレージコントローラ，ストレージデバイス等の各入出力スケジューリング層において先読み効果が期待される．よって，ストレージデバイス集合 $\mathcal{D}$ が有する入出力帯域の総和が，入出力バス帯域やプロセッサによる処理速度によって規定されるシステム上限 M^{seq} であるときに，$\mathcal{D} = \mathcal{D}_M$ であるとすると，下記の関係が成り立つ．

$$\lambda_{\mathcal{D}}^{seq} = \begin{cases} \frac{1}{\sum_{d \in \mathcal{D}} (1/\lambda_d^{seq})} & (|\mathcal{D}| \leq |\mathcal{D}_M|) \\ \frac{B}{M^{seq}} & (|\mathcal{D}| > |\mathcal{D}_M|) \end{cases} \quad (1)$$

ただし，λ_d^{seq} はストレージデバイス d からシーケンシャルアクセスで 1 ページを読み出す平均応答時間を表す．

4.1.1 $|\mathcal{D}| \leqq |\mathcal{D}_{\mathcal{M}}|$ の場合

FTS における各指標は次のように導くことができる．

$$\tau_{\mathcal{D}}(p_{bi}) = \frac{||R||}{\sum_{d \in \mathcal{D}} (1/\lambda_d^{seq})}$$
$$e_{\mathcal{D}}(p_{bi}) = ||R|| \cdot \frac{\sum_{d \in \mathcal{D}} \Omega_d^{seq} \left(\frac{B}{\lambda_d^{seq}} \right)}{\sum_{d \in \mathcal{D}} \frac{1}{\lambda_d^{seq}}} + \frac{||R|| \cdot w_c}{\sum_{d \in \mathcal{D}} \frac{1}{\lambda_d^{seq}}}$$
$$w_{\mathcal{D}}(p_{bi}) = \sum_{d \in \mathcal{D}} \Omega_d^{seq} \left(\frac{B}{\lambda_d^{seq}} \right) + w_c$$

$e_{\mathcal{D}}(p_{bi})$ の第一項は，分母と分子ともにストレージデバイスの追加によって増加するため，$\mathcal{D}$ の変化に伴う増減は小さいと見込まれ，特にストレージデバイス d が全て同一の特性を有する場合には定数項となる．$e_{\mathcal{D}}(p_{bi})$ の第二項は $\mathcal{D}$ の追加に応じて減少する．よって，$|\mathcal{D}| \leqq |\mathcal{D}_{\mathcal{M}}|$ においては消費エネルギー量 $e_{\mathcal{D}}(p_{bi})$ はストレージデバイスの追加により減少することがわかる．

4.1.2 $|\mathcal{D}| > |\mathcal{D}_{\mathcal{M}}|$ の場合

FTS における各指標は次のように導くことができる．

(注1)：関係表 R のビルド処理を行った後，関係表 S に対するプローブ処理が行われるため，各関係表の FTS へと帰着可能である．

$$\tau_{\mathcal{D}}(p_{bi}) = \frac{||R|| \cdot B}{M^{seq}}$$

$$e_{\mathcal{D}}(p_{bi}) = ||R|| \cdot \left(\sum_{d \in \mathcal{D}} \Omega_d^{seq} \left(\frac{||R_d||}{||R||} \cdot B \cdot M^{seq} \right) + \frac{B}{M^{seq}} \right)$$

$$w_{\mathcal{D}}(p_{bi}) = \sum_{d \in \mathcal{D}} \Omega_d^{seq} \left(\frac{||R_d||}{||R||} \cdot B \cdot M^{seq} \right) + w_c$$

ストレージデバイス d の入出力帯域の総和がシステムの性能上限 M^{seq} を超過しており，実行時間はストレージ構成に影響されず一定値となることから，ストレージデバイスを追加するほど消費電力 $w_{\mathcal{D}}(p_{bi})$ ならびに消費エネルギー量 $e_{\mathcal{D}}(p_{bi})$ は増加する．ただし，ストレージデバイス追加により各ストレージデバイスの利用率が低下することから，1 デバイス追加あたりの消費電力増加量は鈍化する．

4.2 IS, NLJ における消費エネルギーコスト推定手法の拡張

単一のストレージデバイス d に対して 1 ページずつランダムアクセスで読み出す際には，基本的には入出力処理はほぼ直列化されて実行されることとなるが，確率 α で隣接するブロックに対するアクセスが生じ，1 回のシークで 2 つ以上の入出力が処理される可能性を考慮すると，平均シーク時間 ι_d に対し平均応答時間 $\lambda_d^{rnd} = (1-\alpha)\iota_d$ となる．

ストレージが複数のストレージデバイス集合 $\mathcal{D}$ によって構成される場合，連続する入出力が同一デバイスの隣接するブロックに対して行われる確率の低下を加味すると，1 ページ読み込みに要する平均応答時間は $\lambda_{\mathcal{D}}^{rnd} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} (1 - \alpha/|\mathcal{D}|)\iota_d^{rnd}$ となる．IS や NLJ においては，実行時間について $\tau_{\mathcal{D}}(p_{bi}) \propto \lambda_{\mathcal{D}}^{rnd}$ であるため，ストレージデバイス追加によって僅かに実行時間が増加することが予想される．また，追加されたストレージデバイス分だけクエリ実行に要するエネルギーは増加する．

4.3 非順序型問合せ実行を用いる場合の消費エネルギーコスト推定手法

従来型のデータベースシステムでは，IS や NLJ の実行に関して，ランダムアクセスで 1 ページずつデータを読み出すことから，結果として 1 台以上のストレージデバイスの入出力帯域を十分に活用することができず，ストレージデバイスが追加されるほど消費エネルギーコストが増加するものであった．これに対し，ストレージに対する高多重なランダムアクセスによって高速な問合せ実行を実現する，アウトオブオーダ型データベースエンジン [18] による IS や NLJ の消費エネルギーコストについて考察したい．アウトオブオーダ型データベースエンジンによる問合せ実行は，単一の問合せ実行を動的に分解し，高多重に非同期入出力を発行することに特徴を有する．消費エネルギーコスト推定の観点からは，ストレージに対するランダムアクセスが多重度 m で行われ，入出力帯域が高効率に活用されることで 1 ページ読み出しに要する平均応答時間 $\lambda_{\mathcal{D}}^{rnd(m)}$ が大幅に低下することが期待される．

5. 評価実験

5.1 実験環境

実験環境の概要を図 1 に示す．当該環境は，実験用サーバ，

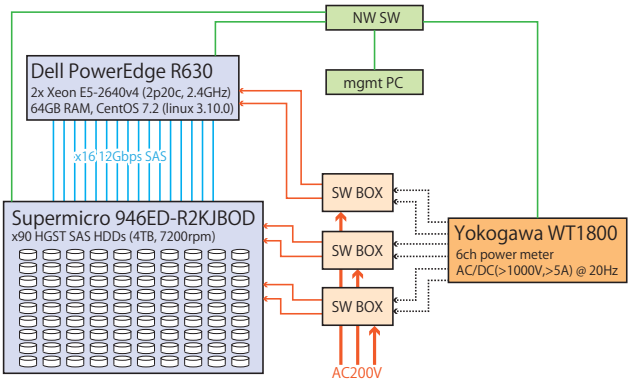

図 1　実験環境の構成概要

JBOD ストレージ，高精度電力計，管理用 PC サーバにより構成される．実験用サーバと JBOD ストレージは miniSAS HD ケーブルを用いて 12Gbps SAS 8 レーンにて接続されており，JBOD ストレージに搭載される各ハードディスクを SCSI ターゲットとして個別に認識し，SCSI プロトコルによる入出力発行や制御が可能である．実験用サーバ及び JBOD ストレージの消費電力を測定するため，それぞれの機器は著者らが製作した電源スイッチボックスを経由して電力の供給を受ける．電源スイッチボックス内には電源回路の電流，電圧が測定可能なプローブ用端子が設けられており，高精度電力計へ接続することで消費電力の測定を行う．管理用 PC から実験用サーバへ処理の実行開始命令や，高精度電力計における測定開始・終了命令をネットワークを介して発行することで，実験環境を構成する機器全体の制御を行う．

5.2 ストレージ構成と入出力性能・消費電力

JBOD ストレージにおける 90 台の磁気ディスクドライブのうち，磁気ディスクアレイを構成するドライブ数を変化させ，当該アレイに含まれないドライブを Sleep 状態とした上で，入出力処理性能と消費電力の計測を行った．ストレージアレイは 64KB 単位でストライピングを行い，8KB 単位のランダム読込，128KB 単位のシーケンシャル読込のそれぞれについて入出力マイクロベンチマークを用いた測定を実施し，ページグループ長として想定する 8KB あたりの入出力に要するエネルギーを算出した．

図 2 に示すシーケンシャル読込の測定結果においては，入出力スレッド数を 1, および 4 としたそれぞれの場合について，8KB 読込に要する入出力あたりのエネルギーを示す．4.1.1,4.1.2 節において示したように，磁気ディスクアレイの入出力帯域がシステムが処理可能な性能上限に至るまでは，ドライブ数増加に対して入出力あたりに要するエネルギーが低下し，その後は増加に転じる．入出力スレッド数が 1 の場合には，5 ドライブの場合に入出力あたりのエネルギーが最小であったが，入出力スレッド数が 4 の場合には，システムの処理性能上限が約 4 倍となり，入出力あたりのエネルギーが最小となるドライブ数は 21 であった．

測定結果を図 3,2 に示す．ランダム読込の測定においては，入出力キュー長を 1, 4, 16, 480 とした場合それぞれの結果を示

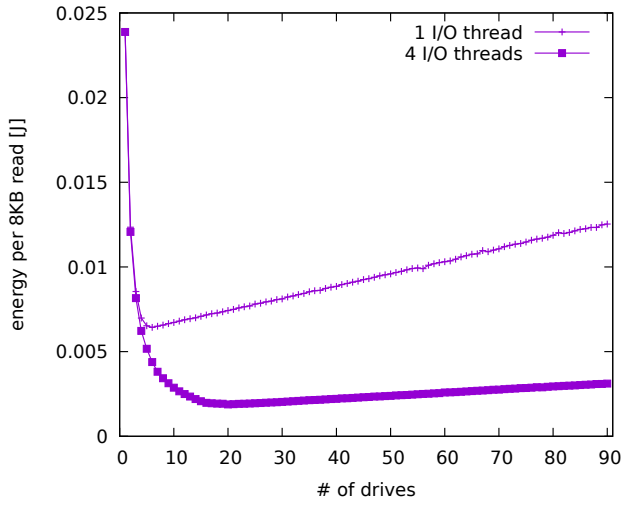

図 2　シーケンシャル読込みにおける利用ドライブ数と 8KB 入出力あたり消費電力量

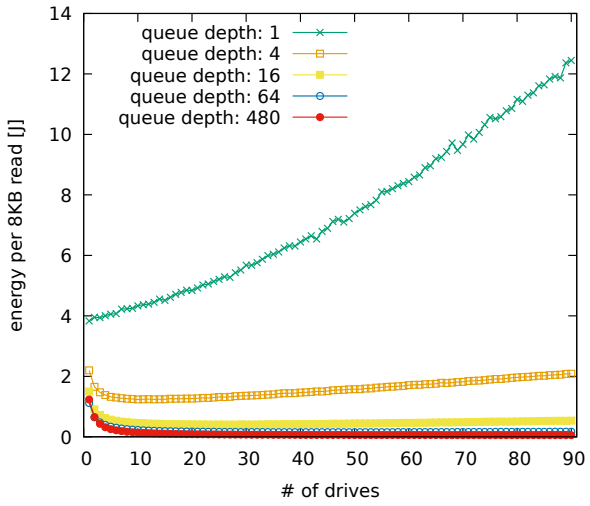

図 3　ランダム読込み (8KB) における利用ドライブ数と 8KB 入出力あたり消費電力量

す．ランダム読込においては，多重に発行される入出力数の上限は入出力キュー長によって制限されるため，入出力キュー長が大きいほど磁気ディスクドライブの入出力帯域を活用可能であり，入出力あたりのエネルギーは小さくなる．入出力キュー長が 1 の場合には，4.2 節における議論の場合に相当し，入出力あたりのエネルギーはドライブ数増加に対して単調に増加することがわかる．一方，入出力キュー長が 1 より大きい場合には，磁気ディスクアレイの入出力帯域が活用される分だけ入出力に要するエネルギーは低下する．

5.3　ストレージ構成を考慮した消費エネルギーコスト推定手法の評価

ストレージ構成を考慮した消費エネルギーコスト推定手法の有効性を評価するため，TPC-H データセット (SF=100)，ならびに PostgreSQL9.4 を用いて表 2 に示す評価用問合せを実行し，消費エネルギー量の推定値と実測値の比較を行った．計測に際しては，磁気ディスクストレージ数が 1, 4, 8, 16, 32, 90

表 2　評価用問合せの概要

問合せ	説明
Q.1	σ(LINEITEM), $\zeta_\sigma \simeq 0.01\%$
Q.2	σ(CUSTOMER) $\bowtie$ ORDERS $\bowtie$ LINEITEM, $\zeta_\sigma \simeq 0.003\%$

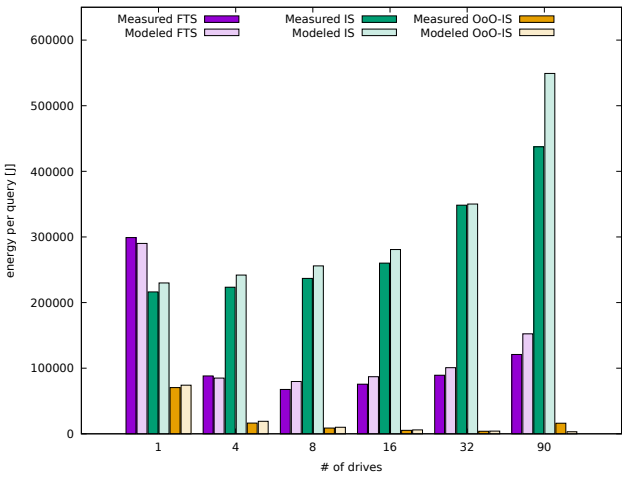

図 4　Q.1: 問合せの消費エネルギーの実測値と推定値の比較

であるボリュームをそれぞれ構成し，それぞれに同一のデータベースを格納した上で，利用していない磁気ディスクドライブは Sleep 状態として実験を行った．また，IS と NLJ に関しては，PostgreSQL における [19] アウトオブオーダ型データベースエンジンの試作実装を用いて計測を行った．

Q.1 の実行に要した消費エネルギー量の測定結果を 4 に示す．Q.1 は単一表走査であり，PostgreSQL による FTS，IS，ならびにアウトオブオーダ型データベースエンジンによる IS(OoO-IS) の 3 つの実行方式について測定を行った．FTS においては，ドライブ数 8 までは消費エネルギー量が低下し，それ以上のドライブ数では消費エネルギー量が増加する結果となり，また，IS においてはドライブ数に応じて消費エネルギー量は単調に増加した．アウトオブオーダ型データベースエンジンによる IS 実行では，ドライブ数 1 の場合においても，PostgreSQL による IS 実行と比べて，消費エネルギー量は 32.5%であった．これは，磁気ディスクドライブ 1 台であっても，入出力要求を多重に発行することでスケジューリング効果による性能向上が生じるためである．ドライブ数 32 までは，アウトオブオーダ型データベースエンジンによる IS 実行に要する消費エネルギー量は低下したが，ドライブ数 90 においては消費エネルギー量が増加した．これは，FTS の性能上限の場合と同様に，ストレージ以外によって性能上限が律速されるためであると考えられる．以上の結果は，4.1.1,4.1.2,4.2 節におけるコスト推定から導出される消費エネルギー量の傾向と一致する．消費エネルギー量のコスト推定と実測値の誤差は平均して 14.8%であり，誤差が顕著であるドライブ数 90 の場合を除くと平均 9.0%であった．また，アウトオブオーダ型データベースエンジンのエネルギー効率に着目すると，最大となるのがドライブ数 32 のときであり，PostgreSQL による FTS に比して 23.0 倍のエネルギー効率であった．

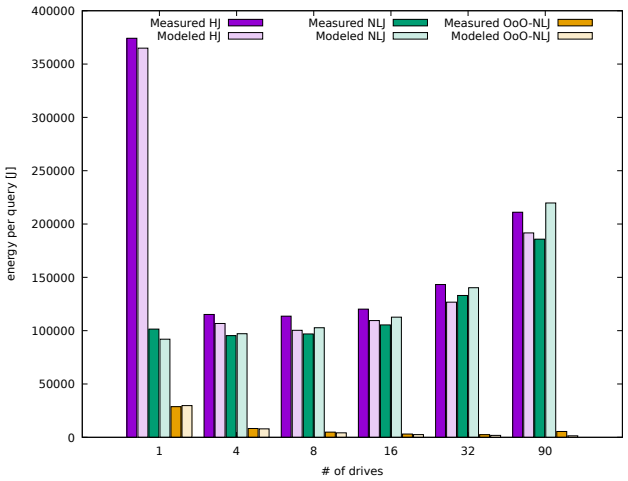

図 5　Q.2: 問合せの消費エネルギーの実測値と推定値の比較

Q.2 の実行に要した消費エネルギー量の測定結果を図 5 に示す．Q.2 は，3 つの関係表の結合であり，PostgreSQL による HJ,NLJ，ならびにアウトオブオーダ型データベースエンジンによる NLJ(OoO-NLJ) の 3 つの実行方式について測定を行った．PostgreSQL による HJ,NLJ 実行については，Q.1 の FTS,IS の場合とそれぞれ同様の傾向が見られ，またアウトオブオーダ型データベースエンジンによる NLJ 実行については，Q.1 のアウトオブオーダ型データベースエンジンによる IS 実行と同様の傾向が見られた．消費エネルギー量のコスト推定と実測値の誤差は平均して 13.4%であった．また，アウトオブオーダ型データベースエンジンのエネルギー効率に着目すると，最大となるのがドライブ数 32 のときであり，PostgreSQL による HJ に比して 53.9 倍のエネルギー効率であった．

6. おわりに

本論文では，ストレージ構成に応じた消費エネルギーコスト推定手法を示し，複数の異なるストレージ構成における問合せ実行性能および消費電力の測定を行う実験により，その有効性の評価を行った．JBOD ストレージ環境を用いた評価実験により，各問合せの平均推定誤差が高々14.8%程度であることが確認された．また，当該コスト推定をアウトオブオーダ型データベースエンジンへと適用し，従来型の問合せ実行に対する問合せ実行との比較実験により，最大で 53.9 倍のエネルギー効率向上が得られることを確認した．

謝　辞

本研究の一部は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）委託業務「エネルギー・環境新技術先導プログラム／革新的な省エネルギー型データベース問合せコンパイラの研究開発」及び「IoT 推進のための横断技術開発プロジェクト／先進 IoT サービスを実現する革新的超省エネルギー型ビッグデータ基盤の研究開発」に拠る．

文　献

[1] J Whitney and P Delforge. Data center efficiency assessment. *Issue Paper, Aug*, 2014.

[2] 合田 和生, 早水 悠登, and 喜連川 優. ストレージシステムの消費エネルギーを考慮したコストベース型のデータベース問合せ最適化手法の提案. In *xSIG 2017*, 2017 (to appear).

[3] Richard Brown et al. Report to congress on server and data center energy efficiency: Public law 109-431. *Lawrence Berkeley National Laboratory*, 2008.

[4] Rakesh Agrawal, Anastasia Ailamaki, Philip A. Bernstein, Eric A. Brewer, Michael J. Carey, Surajit Chaudhuri, AnHai Doan, Daniela Florescu, Michael J. Franklin, Hector Garcia-Molina, Johannes Gehrke, Le Gruenwald, Laura M. Haas, Alon Y. Halevy, Joseph M. Hellerstein, Yannis E. Ioannidis, Hank F. Korth, Donald Kossmann, Samuel Madden, Roger Magoulas, Beng Chin Ooi, Tim O'Reilly, Raghu Ramakrishnan, Sunita Sarawagi, Michael Stonebraker, Alexander S. Szalay, and Gerhard Weikum. The claremont report on database research. *SIGMOD Rec.*, 37(3):9–19, September 2008.

[5] Meikel Poess and Raghunath Othayoth Nambiar. Energy cost, the key challenge of today's data centers: A power consumption analysis of tpc-c results. *Proc. VLDB Endow.*, 1(2):1229–1240, August 2008.

[6] Meikel Poess and Raghunath Othayoth Nambiar. A power consumption analysis of decision support systems. In *Proceedings of the First Joint WOSP/SIPEW International Conference on Performance Engineering*, WOSP/SIPEW '10, pages 147–152, New York, NY, USA, 2010. ACM.

[7] Dimitris Tsirogiannis, Stavros Harizopoulos, and Mehul A. Shah. Analyzing the energy efficiency of a database server. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 231–242, New York, NY, USA, 2010. ACM.

[8] Stavros Harizopoulos, Mehul A. Shah, Justin Meza, and Parthasarathy Ranganathan. Energy efficiency: The new holy grail of data management systems research. In *In CIDR*, 2009.

[9] Yuto Hayamizu, Kazuo Goda, Miyuki Nakano, and Masaru Kitsuregawa. *Application-Aware Power Saving for Online Transaction Processing Using Dynamic Voltage and Frequency Scaling in a Multicore Environment*, pages 50–61. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[10] Willis Lang and Jignesh M. Patel. Towards eco-friendly database management systems. In *CIDR*. www.cidrdb.org, 2009.

[11] Kazuo Goda and Masaru Kitsuregawa. Power-aware remote replication for enterprise-level disaster recovery systems. In *USENIX 2008 Annual Technical Conference*, ATC'08, pages 255–260, Berkeley, CA, USA, 2008. USENIX Association.

[12] Transaction Processing Performance Council. Tpc-h benchmark specification, 2008.

[13] Suzanne Rivoire, Mehul A. Shah, Parthasarathy Ranganathan, and Christos Kozyrakis. Joulesort: A balanced energy-efficiency benchmark. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 365–376, New York, NY, USA, 2007. ACM.

[14] Z. Xu, Y. C. Tu, and X. Wang. Exploring power-performance tradeoffs in database systems. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pages 485–496, March 2010.

[15] Goetz Graefe. Database servers tailored to improve energy efficiency. In *Proceedings of the 2008 EDBT Workshop on Software Engineering for Tailor-made Data Management*, SETMDM '08, pages 24–28, New York, NY, USA, 2008.

ACM.

[16] Willis Lang, Ramakrishnan Kandhan, and Jignesh M Patel. Rethinking query processing for energy efficiency: Slowing down to win the race. *IEEE Data Eng. Bull.*, 34(1):12–23, 2011.

[17] Zichen Xu, Yi-Cheng Tu, and Xiaorui Wang. Pet: Reducing database energy cost via query optimization. *Proc. VLDB Endow.*, 5(12):1954–1957, August 2012.

[18] 喜連川 優 and 合田 和生. アウトオブオーダー型データベースエンジン ooode の構想と初期実験. **日本データベース学会論文誌**, 8(1):131–136, jun 2009.

[19] 早水 悠登, 合田 和生, and 喜連川 優. アウトオブオーダ型クエリ実行に基づくプラグイン可能なデータベースエンジン加速機構. **情報処理学会論文誌データベース（*TOD*）**, 7(2):104–116, jun 2014.