論　文

分析的データベース問合せ処理を対象とするディスクストレージの消費エネルギーコスト推定手法

早水　悠登[†a)]　　合田　和生[†]　　喜連川　優[†,††]

Energy Cost Estimation of Disk Storage for Analytical Database Query Processing

Yuto HAYAMIZU[†a)], Kazuo GODA[†], and Masaru KITSUREGAWA[†,††]

あらまし
データセンタのエネルギー消費は増加の一途をたどっている．近年では所謂ビッグデータブームや IoT ブームがドライバとなり，大規模データの積極的な活用を指向して，ストレージを中心とする膨大な IT 資源を投入したデータ分析基盤の構築が珍しくなく，その中心的役割を担うデータベースシステムのエネルギー効率向上は重要な課題である．こうしたシステムでは取り分けストレージのエネルギー消費が大きい傾向にあるため，著者らは関係データベースシステムを対象とし，ストレージシステムのエネルギー消費に着目したコストベース問合せ最適化による省エネルギー化に取り組んでいる．本論文では，コストベース最適化の核となるストレージの消費エネルギーコスト推定手法として，ディスクアレイに於けるストレージ構成のバリエーションを考慮したコスト推定手法を提案する．そして，JBOD ストレージ，サーバ及び高精度電力計を用いた評価実験により，提案手法により良好な消費エネルギーコスト推定が可能であることを示す．

キーワード　データベースシステム，問合せ最適化，ディスクストレージ，消費エネルギー，コスト推定

1. はじめに

データセンタのエネルギー消費は増加を続けており，2013 年には米国において大規模発電所 34 基分に相当する 910 億 kWh ものエネルギーが消費され，2020 年までに 1,390 億 kWh に増大すると予測されている [1]．また米国に限らず，日本 [2]，欧州 [3] や新興国 [4, 5] における報告から示されるように，世界的にデータセンタのエネルギー消費は増加の一途をたどっているだけでなく，エネルギーの全発電量に対する消費割合も高まる傾向にある．近年では，いわゆるビッグデータや IoT ブームがドライバとなり，大量に蓄積されたデータの積極的な利活用が企業や国家の競争戦略上不可欠との見方も広まりつつあり，ストレージを中心とする膨大な IT 資源を投入し，大規模データ分析基盤を構築することも珍しくない．しかしながら，こうしたエネルギー消費の増大に依拠する IT システム拡大が持続可能であるとは考えにくく，データセンタにおける IT システムの省エネルギー化は重要な課題である．

本論文では，IT システムの中でもとりわけ大規模データの蓄積・管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の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では，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，実行コストが最小と見込まれる実行方法を選択するという，いわゆるコストベースの問合せ最適化の枠組みが広く用いられている．従前の関係データベースシステムにおける問合せ最適化は主に実行時間の最小化を指向しており，入出力量や演算処理量の予測値や，利用可能な資源制約の情報などから実行時間を推定し，コストとして用いる手法が主であった．これに対し，著者らは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に取り組んでいる [6, 7]．本論文では [7] に基づき分析的

† 東京大学生産技術研究所, 東京都
Institute of Industrial Science, the University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo, 157-8505 Japan
†† 国立情報学研究所, 東京都
National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
a) E-mail: haya@tkl.iis.u-tokyo.ac.jp

データベース問合せ処理を対象として，多数のディスクドライブを搭載したディスクアレイストレージを想定し，そのストレージ構成を変化させた場合の性能・電力特性を実測によって明らかにする．そして，性能・電力特性に基づいてストレージの消費エネルギーコスト推定手法を拡張し，様々なストレージ構成のバリエーションに対して良好な推定を与えるコスト推定手法を提案する．そして，JBOD ストレージ，サーバ及び高精度電力計を用いた評価実験により，提案手法の有効性を示す．評価実験に際しては，オープンソースのデータベース管理システム PostgreSQL に加えて，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 [8] の PostgreSQL をベースとする試作実装 [9] を用いて実験を行い，問合せ実行方式毎のエネルギー効率を比較する．

本論文の構成を以下に示す．第 2 節では，関連研究についてまとめる．第 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，第 4 節において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について説明し，第 5 節では，評価実験結果を示し，第 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 [10] 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 [11, 12] 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 [13] ではトランザクション処理の消費電力特性，文献 [14] 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 [15] では問合せ実行計画を構成するアルゴリズム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 [16, 17] 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分析している．文献 [18] は，プロセッサ及びメモリ資源使用状況と消費エネルギーの関係を分析するためのベンチマークの提案を行っている．文献 [19] では，ARM プロセッサと Xeon プロセッサの電力効率を比較分析している．省電力化のアプローチとしては，プロセッサの周波数スケーリングに着目した取り組みとして，文献 [20–22] のアプリケーション性能等に応じた周波数スケーリング制御手法，文献 [23] の電力効率を最大化する周波数同定手法，文献 [24] の周波数スケーリングと複数問合せの共通処理集約の組み合わせによる省エネルギー化手法などがみられる．分析ワークロードを主眼とした省エネルギー化のアプローチとしては，文献 [25] のマテリアライズドビューの生成・維持に要する電力と性能のトレードオフ調整手法や，文献 [26–28] のワークロードのモデル化に基づくディスクアレイのドライブ電源制御による省電力化手法，文献 [29, 30] の並列データベースシステムにおける電力効率分析とクラスタ構成最適化手法，文献 [31] の並列処理基盤におけるノード電源制御手法などがある．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 [32]．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 [33] 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 [34] 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．これらの取り組みは，いずれも本論文が対象とする問合せ最適化におけるコスト推定手法とは技術的に直行するものである．

問合せ最適化に関する省エネルギー化の議論として，先駆的な取り組みは 20 年以上前に文献 [35] が存在するが，主に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のデータベースシステムが対象として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 [36] では，データベースシステムの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 [37, 38] では，問合せ処理性能とプロセッサの消費電力に着目したモデル化や，その評価が行われている．性能と消費電力のトレードオフの調整に関して，文献 [39] 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．また文献 [40] 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調整可能とするアプローチも見られる．一方で，本論文で提案する，多数のディスク

ドライブを搭載するストレージアレイを想定した消費エネルギーコスト推定手法は，著者らの発表文献 [6, 7] を除いては，著者らの知る限りにおいて存在せず，新規な取り組みである．本論文では，文献 [7] の内容に基づき，ハッシュ結合のコスト推定手法を改善し，コスト推定に用いる入出力性能・電力特性のスループット変化に対する変化量を実測によって明らかにするとともに，評価に用いる問合せを追加し，提案手法の評価を行っている．

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 関係データベースシステムにおける問合せ最適化

関係データベースシステムでは，ユーザが SQL 等の宣言的言語によって記述した問合せ要求を受け付け，**関係演算子**から構成される式に翻訳して問合せを表現する．そして，問合せを実行するための具体的な手続きを示す問合せ実行計画を生成し，これに基づいて問合せ処理を駆動する．問合せ実行計画は，多くの場合**関係演算アルゴリズム**を節点とする木構造によって表現される．関係演算アルゴリズムは，例えば関係演算子のリレーション選択 σ に対応する全表走査や索引走査，あるいはリレーション結合 $\bowtie$ に対応するネステッドループ結合やハッシュ結合など，1 つの関係演算子に対して複数の選択肢が存在する．つまり，単一の問合せに対して，関係演算アルゴリズムの組み合わせの数だけ問合せ実行計画の選択肢が存在するため，関係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推定される問合せ実行計画を選択する．

初期の関係データベースシステム実装においては，実装の容易性から事前に規定されたヒューリスティックなルールに基づいて問合せ実行計画を選択するルールベース最適化がしばしば用いられたが，昨今の関係データベースシステム実装においては，所定のコストモデルを用いて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ように定式化できる．

実行時間最小化問題:

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

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

これに対し，消費電力や消費電力量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，許容される消費電力量 E の制限下において，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定式化できる．

エネルギー制約のある実行時間最小化問題:

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

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定式化できる．

消費電力量最小化問題:

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

3.2 問合せ実行計画のストレージ消費エネルギーコスト推定

問合せ実行計画の木構造において，節点を結ぶ辺は関係演算アルゴリズム同士の入力・出力関係を表し，子が出力したタプルを親が入力として受け取り，処理する方式が一般的である．関係演算アルゴリズムは，入力タプル毎に逐次演算実行が可能である**パイプライン動作可能アルゴリズム**と，入力タプルが一定量蓄積されないと演算実行を開始できない**ブロッキングアルゴリズム**の 2 種類に大別できる．問合せ実行計画がすべてパイプライン動作可能アルゴリズムから構成される場合には，全ての節点がパイプライン動作によって同時並行で駆動されながら処理が進行する．これに対し，ブロッキングアルゴリズムが含まれる場合，まず当該アルゴリズムの入力側部分木の実行が駆動され，その実行結果が入力タプルとして一定量蓄積された後に，当該アルゴリズム及びその出力側部分木の実行が駆動される．つまり，ブロッキングアルゴリズムの入力となる辺を境界として，問合せ実行計画を複数の部分木に分割した場合，各々の部分木はパイプライン動

作可能な関係演算アルゴリズムの集合となる．このように規定される部分木を**実行計画ブロック**と称することとする．一般に問合せ実行計画 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アルゴリズム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．

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

即ち，p_{bi} を構成する関係演算アルゴリズム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．

3.3 分析的データベース問合せ処理の基本的な関係演算アルゴリズム

本論文が対象とする分析的データベース問合せ処理は，TPC-H ベンチマーク [33] に代表されるように，リレーションに格納された多数のタプルの選択とその結合が主要な演算としてしばしば用いられる．選択の基本的な関係演算アルゴリズムとしては，リレーション全体を操作する全表走査 (FTS) や索引走査 (IS) が，結合の基本的な関係演算アルゴリズムとしては，ハッシュ結合 (HJ) やネステッドループ結合 (NLJ) が広く用いられている [41]．本論文では，これらの 4 種類の関係演算アルゴリズムに焦点を当て，実行時間，ストレージの消費電力量，ストレージの最大消費電力のコスト推定手法を示す．(注1)

なお，これ以降のコスト推定手法に関する議論に用いる表記については，表 1 を参照されたい．

a） **FTS**(R): 全表走査によるリレーション R の選択

リレーション R は，ストレージ上に実体を持つ関係表から構成される場合と，問合せ実行計画の部分木の出力から構成される場合との 2 通りが存在するが，全表走査の場合には実体を持つ関係表のみを考えればよく，各コスト指標は次のように与えられる．

$$\begin{aligned}
p_{bi} &= \mathbf{FTS}(R) \\
\tau(p_{bi}) &= ||R|| \cdot \lambda_R^{seq} \\
e(p_{bi}) &= \tau(p_{bi}) \cdot w(p_{bi}) \\
w(p_{bi}) &= \sum_{d \in \mathcal{D}} \Omega_d^{seq} \left(\frac{||R_d||}{||R||} \cdot \frac{B_R}{\lambda_R^{seq}} \right) + w_c
\end{aligned}$$

b） **IS**(R): 二次索引走査によるリレーション R の選択

二次索引走査の場合には，全表走査と同様にリレーション R は実体を持つ関係表から構成される場合のみを考慮すればよく，各コスト指標は次のように与えられる．

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

c） **HJ**(R, S): リレーション R, S のハッシュ結合

ハッシュ結合の処理においては，まずリレーション R より入力タプルを受け取りながら，当該タプルを随時ハッシュ表に挿入するビルド処理を行い，その後にリレーション S より入力タプルを受け取りながら，結合対象タプルをハッシュ表から探索して随時結合結果を出力するプローブ処理を行う．ビルド処理の実行時間は，リレーション R を構成する問合せ実行計画の処理が律速要因となる場合と，ハッシュ表への挿入演算が律速要因となる場合が考えられ，同様にプローブ処理の実行時間は，リレーション S を構成する問合せ実行計画の処理が律速要因となる場合と，ハッシュ表の探索演算が律速要因となる場合が考えられる．したがって，各コスト指標は次のように与えられる．

$$p_{bi} = \mathbf{HJ}(R, S)$$

(注1)：分析的データベース問合せ処理において用いられる関係演算アルゴリズムは必ずしもこれら 4 種類に限定されないが，既存の関係演算アルゴリズムに対し網羅的にコスト推定手法を与えることは本論文の対象とする議論の範囲を超えるため，ここでは扱わないこととする．

$$\begin{aligned}\tau(p_{bi}) =& \max\left(\tau(R), |R|/\mu_{\bowtie}^{build}\right)\\ &+ \max\left(\tau(S), |S|/\mu_{\bowtie}^{probe}\right)\\ e(p_{bi}) =& \tau(R)w(R) + \tau(S)w(S)\\ w(p_{bi}) =& \max(w(R), w(S))\end{aligned}$$

d） **NLJ**(**IS**(R),**IS**(S)): 索引走査を伴うリレーション R, S のネステッドループ結合

ネステッドループ結合は，左辺からの入力タプル毎に，右辺の問合せ実行計画の処理を駆動する．全体の演算量は左辺の出力タプル数と右辺の出力タプル数の積によって決まるため，それぞれの出力タプル数が一定程度小さい場合に有効なアルゴリズムである．実用上は，主として左辺と右辺が索引走査によって絞り込まれる問合せにおいて有効であることから，ここでは R, S がそれぞれ実体を持つ関係表の索引走査であるものとする．この場合，各コスト指標は次のように与えられる．

$$\begin{aligned}p_{bi} =& \mathbf{NLJ}(\mathbf{IS}(R), \mathbf{IS}(S))\\ \tau(p_{bi}) =& \zeta_\sigma \cdot (1 + c_a) \cdot |R| \cdot (\lambda_R^{rnd} + j_{R,S} \cdot \lambda_S^{rnd})\\ e(p_{bi}) =& \tau(p_{bi}) \cdot w(p_{bi})\\ w(p_{bi}) =& \sum_{d \in \mathcal{D}} \Omega_d^{rnd}(\theta_d) + w_c\end{aligned}$$

$$\text{where } \theta_d = \frac{|R_d|}{|R|} \cdot \frac{1}{\lambda_R^{rnd} + j_{R,S} \cdot \lambda_S^{rnd}} + \frac{|S_d|}{|S|} \cdot \frac{j_{R,S}}{\lambda_R^{rnd} + j_{R,S} \cdot \lambda_S^{rnd}}$$

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，ディスクドライブ単体がストレージとして用いられることは殆どなく，多数のディスクドライブを搭載したディスクアレイ環境が一般的となっている．また消費電力を低減することを目的とし，ディスクアレイ上の全てのディスクドライブを常時稼働させることなく，一部のディスクドライブを一時的にスピンダウンして休止させる機能を備える商用ディスクアレイ製品も珍しくない．即ち，ある特定のディスクアレイ構成を前提とした環境であっても，データベースシステムが利用可能な入出力帯域や，その際の消費電力の特性は運用中のストレージ構成によって大きく異なる．そこで本節では，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．

表 1　コスト推定手法における変数の一覧
Table 1　A list of variables used in cost estimation

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
$\|R\|$	リレーション R のタプル数
$\|\|R\|\|$	実体を持つ関係表 R を構成するページ数
$\|\|R_d\|\|$	実体を持つ関係表 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
$\|R_d\|$	実体を持つ関係表 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
$j_{R,S}$	結合 $R \bowtie S$ の結合増幅率
$\mu_{\bowtie}^{build}$	ハッシュ結合 $\bowtie$ のビルド処理において，ハッシュ表に毎秒挿入可能なタプル数の最大スループット
$\mu_{\bowtie}^{probe}$	ハッシュ結合 $\bowtie$ のプローブ処理において，ハッシュ表を毎秒検索可能な回数の最大スループット
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_R	実体を持つ関係表 R のページ長
λ_R^{seq}	シーケンシャルアクセスにより実体を持つ関係表 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実体を持つ関係表 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
$\Omega_d^{seq}(\theta)$	ストレージデバイス d からシーケンシャルアクセスによって転送レート θ でデータを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
$\Omega_d^{rnd}(\theta)$	ストレージデバイス d からランダムアクセスによってスループット θ でデータを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

4.1 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．

$$\lambda_{\mathcal{D}}^{seq} = \begin{cases} \frac{1}{\sum_{d \in \mathcal{D}} \left(1/\lambda_d^{seq}\right)} & (|\mathcal{D}| \leqq |\mathcal{D}_M|) \\ B_R / M^{seq} & (|\mathcal{D}| > |\mathcal{D}_M|) \end{cases}$$

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．

$$\tau_{\mathcal{D}}(p_{bi}) = \frac{||R||}{\sum_{d\in\mathcal{D}}(1/\lambda_d^{seq})}$$

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

$$w_{\mathcal{D}}(p_{bi}) = \sum_{d\in\mathcal{D}} \Omega_d^{seq} \left(B_R/\lambda_d^{seq}\right) + w_c$$

$e_{\mathcal{D}}(p_{bi})$ 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，$|\mathcal{D}| \leqq |\mathcal{D}_{\mathcal{M}}|$ においては消費電力量 $e_{\mathcal{D}}(p_{bi})$ 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．

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

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

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

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 $w_{\mathcal{D}}(p_{bi})$ ならびに消費電力量 $e_{\mathcal{D}}(p_{bi})$ 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わず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た場合には，実行時間の短縮にはつながらず，消費電力および消費電力量が増加する．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 [8] 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，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 4 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時期ディスクドライブを 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

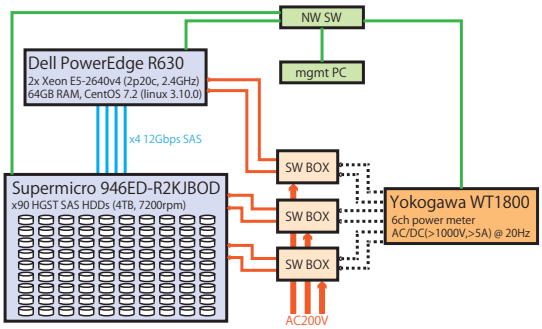

図 1　実験環境の構成概要
Fig. 1　Overview of the experimental environment

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経由で発行し，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，利用ボリュームを構成するドライブ数を変化させ，当該ボリュームに含まれない磁気ディスクドライブをスピンダウンして停止状態とした上で，入出力処理性能と消費電力の計測を行った．利用ボリュームは 1024KB 単位でストライピングを行い作成した．8KB 単位のランダム読込(注2)，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，入出力要求発行レートを変化させながら，入出力スループットと消費電力を計測した．　それぞれの測定点は 1 回ずつ測定を行い，1 回の測定においては 1 分間入出力マイクロベンチマークを一定の負荷で実行し，入出力性能を 1 秒毎にと消費電力を 0.05 秒毎に記録し，1 分間の平均値を測定値として採用した．

図 2 にシーケンシャル読込の測定結果を示す．各点は実測値を，線は各ストレージ構成について実測値から導出した性能・電力特性を表す．磁気ディスクドライブ数が多いストレージ構成ほど消費電力が高く，増分はほぼ磁気ディスクドライブ数に比例した．また各ストレージ構成について，データ転送レートに応じて消費電力が増加する傾向が見られるが，その増加量は相対的に小さく，90 ドライブ構成については，0 MB/s（入出力無し）の場合と 1986 MB/s（最大性能）の場合の消費電力の差は 46.9 W であった．

図 3 にランダム読込の測定結果を示す．各点は実測値を，線は各ストレージ構成について実測値から導出した性能・電力特性を表す．シーケンシャル読込の場合と同様に，磁気ディスクドライブ数が多いストレー

(注2)：データベースのページ長は 8〜16KB とされる場合が多く，本論文で用いた PostgreSQL ではページ長が 8KB であるため，本論文ではランダム読込については 8KB をアクセス単位として設定した．

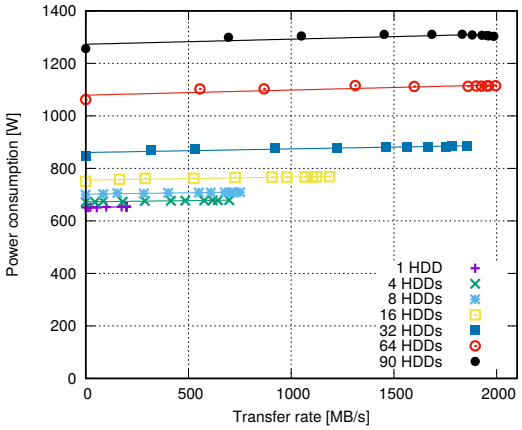

図 2　シーケンシャル読込 (128KB 単位) における性能・電力特性
Fig. 2　Power-performance profile of sequential access (128KB)

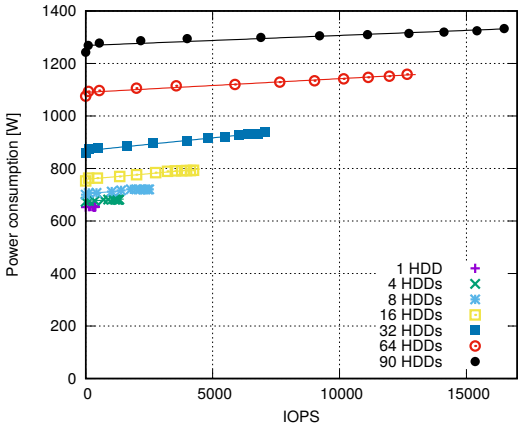

図 3　ランダム読込 (8KB 単位) における性能・電力特性
Fig. 3　Power-performance profile of random access (8KB)

ジ構成ほど消費電力が高く，その増分はほぼ磁気ディスクドライブ数に比例した．入出力負荷の変動による消費電力の変動量は，シーケンシャル読込の場合と比べて大きく，90 ドライブ構成については，0 IOPS（入出力なし）の場合と 16400 IOPS （最大性能） の場合の消費電力の差は 90.0 W であった．

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，消費電力量の推定値と実測値の比較を行った． 評価においては，Q.1 に関しては FTS を用いる場合と IS を用いる場合それぞれについて 1 つの固定した問合せ実行計画を指定し，当該問合せ実行計画に関する消費エネルギーコストの推定値と実測値を比較した．また，Q.2，Q.3 に関しては HJ を用いる場合と，NLJ を用いる場合それぞれについて 1 つの固定した問合せ実行計画を指定し，同様に比較を行った． 計測に際しては，磁気ディスクドライブ数が 1, 4, 8, 16, 32, 64, 90 であるボリュームをそれぞれ構成し，それぞれに TPC-H データセットをロードした同一のデータベースイメージを格納した上で，利用していない磁気ディスクドライブはスピンダウンして停止状態として実験を行った．また，IS と NLJ に関しては，通常の PostgreSQL による実行に加えて，PostgreSQL におけるアウトオブオーダ型データベースエンジンの試作実装 [9] を用いて計測を行った． 消費エネルギーコスト推定手法における変数について，B_R は PostgreSQL のページ長 16KB とし，リレーションのデータ量に関わる変数 $|R|, |R_d|, ||R||, ||R_d||$ については PostgreSQL のリレーション毎の統計情報から算出した．問合せの選択率 ζ_σ，結合増幅率 $j_{R,S}$，索引等のオーバヘッド c_a については PostgreSQL の問合せ最適化器の生成する推定値を用いた．(注3) 消費電力に関わる変数 $\Omega_d^{seq}(\theta), \Omega_d^{rand}(\theta), w_c$ は図 2,3 に示す測定結果からストレージ構成毎の線形モデルを構築し利用した．入出力の平均応答遅延 $\lambda_R^{seq}, \lambda_R^{rnd}, \lambda_d^{rnd}, \lambda_d^{rnd(m)}$ は，事前に入出力のベンチマークを用いて計測した値を利用した(注4)．ハッシュ結合演算のスループット

表 2 評価用問合せの概要
Table 2 Queries for experimental evaluations

問合せ	説明
Q.1	σ(LINEITEM), $\zeta_\sigma \simeq 0.01\%$
Q.2	σ(PART) $\bowtie$ LINEITEM, $\zeta_\sigma \simeq 0.003\%$
Q.3	σ(CUSTOMER) $\bowtie$ ORDERS $\bowtie$ LINEITEM, $\zeta_\sigma \simeq 0.003\%$

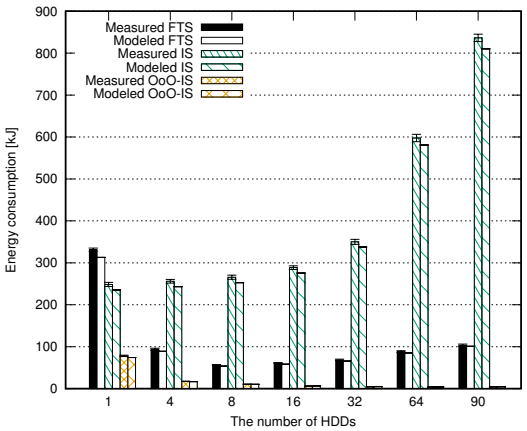

図 4 Q.1: 問合せの消費エネルギーの実測値と推定値の比較
Fig. 4 Q.1: Comparisons of measured and modeled energy consumption

$\mu_{\bowtie}^{build}, \mu_{\bowtie}^{probe}$ は，リレーションのデータを予め主記憶バッファにキャッシュした状態で，各問合せと同等のハッシュ結合を実行して事前計測を行い，処理データ量と実行時間からこれら変数の値を算出して用いた．(注5)． 実験に際しては，それぞれの測定点に関して 5 回ずつ測定を行い，問合せ実行に要した消費電力量の 5 回の平均値を測定値として採用した．

Q.1 の実行に要した消費電力量の測定結果を 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 までは消費電力量が低下し，それ以上のドライブ数では消

(注3)：TPC-H に関して PostgreSQL は高精度な問合せ実行コスト推定を与えることが報告されており [42, 43]，本論文に用いた問合せに関してもその誤差は 1%以下であることを予備実験により確認した．
(注4)：α,m については，論理的には $\lambda_d^{rnd}, \lambda_d^{rnd(m)}$ の導出に用いられる変数である．ただし，$\lambda_d^{rnd}, \lambda_d^{rnd(m)}$ の値を入出力マイクロベンチマークによって直接事前計測し，評価実験においては α,m の算出は不要であったため省略した．
(注5)：本論文では提案する消費エネルギーコスト推定手法自体の有効性を評価するため，各変数は正確な値が得られていることを想定し $\mu_{\bowtie}^{build}, \mu_{\bowtie}^{probe}$ は事前計測した値を用いた．しかし，問合せ最適化器のカーディナリティ推定と実行時間推定から実行前にこれらの予測値を算出する方法や，あるいは [44] に示される問合せ実行計画のノード毎の実行時処理カウンタを用いる方法などにより，事前計測を行わずに当該変数の値は算出，補正可能であると考えられる．

表 3　各問合せ，各実行方式の消費電力量の平均推定誤差
Table 3　Average estimation error of energy consumption of each query and method

	FTS	IS	OoO-IS
Q.1	4.4%	4.1%	3.6%

	HJ	NLJ	OoO-NLJ
Q.2	3.9%	3.6%	2.5%
Q.3	4.1%	3.8%	2.4%

費電力量が増加する結果となった．ドライブ数の増加に応じて単位時間あたりの消費電力は増加するが，ドライブ数 8 までは，ドライブ数に応じて入出力の最大転送速度が増加し，問合せ実行時間が減少することにより消費電力量は減少した．一方で，ドライブ数 8 以上では最大転送速度が飽和して，問合せ実行時間が概ね一定となり，消費電力量はドライブ数に応じて増加した．IS においてはドライブ数に応じて消費電力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，消費電力量は 23.5%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 64 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 実行に要する消費電力量は低下したが，ドライブ数 90 においては消費電力量が 0.3kJ（17%）増加した．これは，実行時間はドライブ数の増加により 0.2 秒程度短縮する一方で，平均消費電力が 193W 増加したためである．

以上の結果は，第 4. 1. 1 節, 第 4. 1. 2 節, 第 4. 2 節におけるコスト推定から導出される消費電力量の傾向と整合する．消費電力量のコスト推定と実測値の誤差は表 3 に示すように, FTS は平均 4.4%, IS は平均 4.1%, OoO-IS は平均 3.6%であった．また，PostgreSQL と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 90 のときであり，PostgreSQL による FTS に比して 25.3 倍のエネルギー効率であった．

Q.2, Q.3 の実行に要した消費電力量の測定結果を図 5，6 に示す．Q.2 は 2 つのリレーションの結合，Q.3 は 3 つのリレーションの結合であり，それぞれ 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．いずれの問合せ

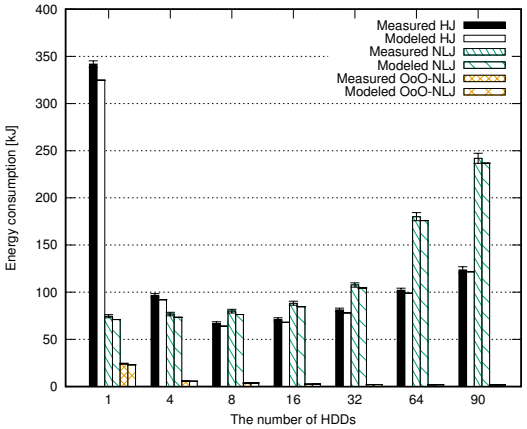

図 5　Q.2: 問合せの消費エネルギーの実測値と推定値の比較
Fig. 5　Q.2: Comparisons of measured and modeled energy consumption

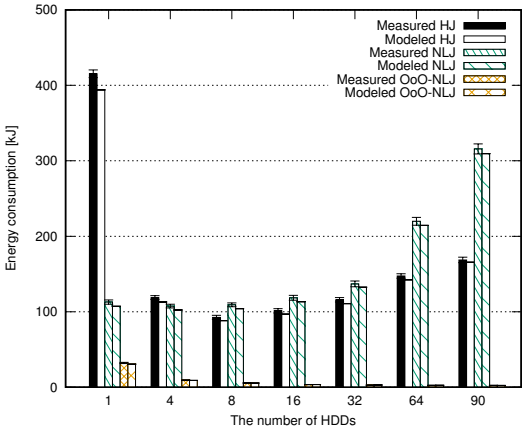

図 6　Q.3: 問合せの消費エネルギーの実測値と推定値の比較
Fig. 6　Q.3: Comparisons of measured and modeled energy consumption

についても，PostgreSQL による HJ 実行では，Q.1 の場合と同様にドライブ数 8 の場合が消費電力量が最小となり，PostgreSQL による NLJ 実行では，ドライブ数の増加に応じて消費電力量は単調に増加した．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 実行についても，Q.1 の場合の同様にドライブ数 64 までは消費電力量が減少し，ドライブ数 90 では 20%程度の増加に転じた．消費電力量のコスト推定と実測値の誤差は，表 3 に示すように 2.4%–4.1%程度であった． IS および NLJ の消費電力量の推定値については，第 3. 3 節に示すように選択率 ζ_σ および結合増幅

率 $j_{R,S}$ に比例する項を持つため，$\zeta_\sigma, j_{R,S}$ の推定誤差に対して線形に影響を受ける．また複数の結合演算を有する問合せでは，各結合演算での推定誤差が乗じられて増幅される．本論文で用いた Q.2, Q.3 については，事前の予備実験により $\zeta_\sigma, j_{R,S}$ の推定誤差はいずれも 1%以下であることを確認したため，表 3 に示す消費電力量のコスト推定に $\zeta_\sigma, j_{R,S}$ の見積もり誤差が与える影響は 1%以下程度であると考えられる． また，PostgreSQL と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 Q.2 のドライブ数 90 のときであり，PostgreSQL による HJ に比して 71.2 倍のエネルギー効率であった．

6. おわりに

本論文では，多数のディスクドライブを搭載したディスクアレイストレージを想定し，ストレージ構成を変化させた場合の性能・電力特性を実測によって明らかにするとともに，ストレージ構成のバリエーションを考慮したストレージ消費エネルギーコスト推定手法を提案した．評価実験においては，JBOD ストレージの利用ドライブ数を 1 から 90 まで変化させ，複数の問合せ実行方式により消費電力量を計測した結果，各ストレージ構成，各問合せ，各問合せ実行方式の平均推定誤差は高々 4.4%程度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方式と比較する実験により，最大で 71.2 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）の委託業務「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」の結果得られたものである．

文　献

[1] J. Whitney and P. Delforge, “Data center efficiency assessment,” Issue paper on NRDC (The Natural Resource Defense Council), 2014.

[2] 朽網道徳，“グリーン it 推進協議会調査分析委員会 総合報告（2008 年度〜2012 年度）,” Technical report，グリーン IT 推進協議会，2013.

[3] M. Avgerinou, P. Bertoldi, and L. Castellazzi, “Trends in data centre energy consumption under the european code of conduct for data centre energy efficiency,” Energies, vol.10, no.10, pp.1–18, 2017.

[4] I.F. Sulaiman, “The emerging indonesian data center market and energy efficiency opportunities,” Technical report, Asian Development Bank, 2017.

[5] “Data centre energy efficiency benchmarking: Final report,” National Environment Agency (NEA).

[6] 合田和生，早水悠登，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,” The 1st. cross-disciplinary Workshop on Computing Systems, Infrastructures, and Programming (xSIG 2017)，2017.

[7] 早水悠登，合田和生，喜連川優，“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,” 電子情報通信学会第 9 回データ工学と情報マネジメントに関するフォーラム／第 15 回日本データベース学会年次大会 (DEIM2017)，2017.

[8] 喜連川優，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，vol.8，no.1，pp.131–136，jun 2009.

[9] 早水悠登，合田和生，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），vol.7，no.2，pp.104–116，jun 2014.

[10] R.E. Brown, R. Brown, E. Masanet, B. Nordman, B. Tschudi, A. Shehabi, J. Stanley, J. Koomey, D. Sartor, P. Chan, et al., “Report to congress on server and data center energy efficiency: Public law 109-431,” Technical report, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US), 2007.

[11] R. Agrawal, A. Ailamaki, P.A. Bernstein, E.A. Brewer, M.J. Carey, S. Chaudhuri, A. Doan, D. Florescu, M.J. Franklin, H. Garcia-Molina, J. Gehrke, L. Gruenwald, L.M. Haas, A.Y. Halevy, J.M. Hellerstein, Y.E. Ioannidis, H.F. Korth, D. Kossmann, S. Madden, R. Magoulas, B.C. Ooi, T. O’Reilly, R. Ramakrishnan, S. Sarawagi, M. Stonebraker, A.S. Szalay, and G. Weikum, “The claremont report on database research,” SIGMOD Rec., vol.37, no.3, pp.9–19, Sept. 2008.

[12] Y.-c. Tu, X. Wang, and Z. Xu, “Power-aware dbms: Potential and challenges,” pp.598–599, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[13] M. Poess and R.O. Nambiar, “Energy cost, the key challenge of today’s data centers: A power consumption analysis of tpc-c results,” Proc. VLDB Endow., vol.1, no.2, pp.1229–1240, Aug. 2008.

[14] M. Poess and R. Othayoth Nambiar, “A power consumption analysis of decision support systems,” Proceedings of the First Joint WOSP/SIPEW International Conference on Performance Engineering, pp.147–152, WOSP/SIPEW ’10, ACM, New York, NY, USA, 2010.

[15] D. Tsirogiannis, S. Harizopoulos, and M.A. Shah,

"Analyzing the energy efficiency of a database server," Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, pp.231–242, SIGMOD '10, ACM, New York, NY, USA, 2010.

[16] N. Nishikawa, M. Nakano, and M. Kitsuregawa, "Energy aware raid configuration for large storage systems," 2011 International Green Computing Conference and Workshops, pp.1–5, July 2011.

[17] S. Harizopoulos, M.A. Shah, J. Meza, and P. Ranganathan, "Energy efficiency: The new holy grail of data management systems research," CIDR Perspectives 2009, 2009.

[18] Y. Zhou, M. Alghamdi, S. Taneja, W.-S. Ku, and X. Qin, "Towards energy-efficient multicore database systems," 2016 Seventh International Green and Sustainable Computing Conference (IGSC), pp.1–8, Nov. 2016.

[19] U. Sirin, R. Appuswamy, and A. Ailamaki, "Oltp on a server-grade arm: Power, throughput and latency comparison," Proceedings of the 12th International Workshop on Data Management on New Hardware, pp.10:1–10:7, DaMoN '16, ACM, New York, NY, USA, 2016.

[20] Y. Hayamizu, K. Goda, M. Nakano, and M. Kitsuregawa, "Application-aware power saving for online transaction processing using dynamic voltage and frequency scaling in a multicore environment," pp.50–61, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[21] M. Korkmaz, A. Karyakin, M. Karsten, and K. Salem, "Towards dynamic green-sizing for database servers," International Workshop on Accelerating Data Management Systems Using Modern Processor and Storage Architectures - ADMS 2015, Kohala Coast, Hawaii, USA, August 31, 2015., eds. by R. Bordawekar, T. Lahiri, B. Gedik, and C.A. Lang, pp.25–36, 2015.

[22] Y.-C. Tu, X. Wang, B. Zeng, and Z. Xu, "A system for energy-efficient data management," SIGMOD Rec., vol.43, no.1, pp.21–26, May 2014. http://doi.acm.org/10.1145/2627692.2627696

[23] S. Götz, T. Ilsche, J. Cardoso, J. Spillner, T. Kissinger, U. Aβmann, W. Lehner, W.E. Nagel, and A. Schill, "Energy-efficient databases using sweet spot frequencies," Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, pp.871–876, UCC '14, IEEE Computer Society, Washington, DC, USA, 2014.

[24] W. Lang and J.M. Patel, "Towards eco-friendly database management systems," CIDR 2009, Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2009, Online Proceedings, 2009.

[25] A. Roukh, L. Bellatreche, A. Boukorca, and S. Bouarar, "Eco-dmw: Eco-design methodology for data warehouses," Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP, pp.1–10, DOLAP '15, ACM, New York, NY, USA, 2015.

[26] P. Behzadnia, W. Yuan, B. Zeng, Y.-C. Tu, and X. Wang, "Dynamic power-aware disk storage management in database servers," pp.315–325, Springer International Publishing, Cham, 2016.

[27] N. Nishikawa, M. Nakano, and M. Kitsuregawa, "Application sensitive energy management framework for storage systems," IEEE Transactions on Knowledge and Data Engineering, vol.27, no.9, pp.2335–2348, Sept. 2015.

[28] N. Nishikawa, M. Nakano, and M. Kitsuregawa, "Energy efficient storage management cooperated with large data intensive applications," 2012 IEEE 28th International Conference on Data Engineering, pp.126–137, April 2012.

[29] W. Lang, S. Harizopoulos, J.M. Patel, M.A. Shah, and D. Tsirogiannis, "Towards energy-efficient database cluster design," Proc. VLDB Endow., vol.5, no.11, pp.1684–1695, July 2012.

[30] B. Feng, J. Lu, Y. Zhou, and N. Yang, "Energy efficiency for mapreduce workloads: An in-depth study," Proceedings of the Twenty-Third Australasian Database Conference - Volume 124, pp.61–70, ADC '12, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 2012.

[31] W. Lang and J.M. Patel, "Energy management for mapreduce clusters," PVLDB, vol.3, pp.129–139, 2010.

[32] K. Goda and M. Kitsuregawa, "Power-aware remote replication for enterprise-level disaster recovery systems," USENIX 2008 Annual Technical Conference, pp.255–260, ATC'08, USENIX Association, Berkeley, CA, USA, 2008.

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

[34] S. Rivoire, M.A. Shah, P. Ranganathan, and C. Kozyrakis, "Joulesort: A balanced energy-efficiency benchmark," Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, pp.365–376, SIGMOD '07, ACM, New York, NY, USA, 2007.

[35] R. Alonso and S. Ganguly, "Energy efficient query optimization," Technical report, Matsushita Info Tech Lab, 1992.

[36] G. Graefe, "Database servers tailored to improve energy efficiency," Proceedings of the 2008 EDBT Workshop on Software Engineering for Tailor-made Data Management, pp.24–28, SETMDM '08, ACM, New York, NY, USA, 2008.

[37] Z. Xu, Y.C. Tu, and X. Wang, "Exploring power-performance tradeoffs in database systems," 2010

IEEE 26th International Conference on Data Engineering (ICDE 2010), pp.485–496, March 2010.
[38] Z. Xu, Y.C. Tu, and X. Wang, “Dynamic energy estimation of query plans in database systems,” 2013 IEEE 33rd International Conference on Distributed Computing Systems, pp.83–92, July 2013.
[39] W. Lang, R. Kandhan, and J.M. Patel, “Rethinking query processing for energy efficiency: Slowing down to win the race.,” IEEE Data Eng. Bull., vol.34, no.1, pp.12–23, 2011.
[40] Z. Xu, Y.-C. Tu, and X. Wang, “Pet: Reducing database energy cost via query optimization,” Proc. VLDB Endow., vol.5, no.12, pp.1954–1957, Aug. 2012.
[41] R. Ramakrishnan and J. Gehrke, Database management systems, McGraw Hill, 2000.
[42] V. Leis, A. Gubichev, A. Mirchev, P. Boncz, A. Kemper, and T. Neumann, “How good are query optimizers, really?,” Proc. VLDB Endow., vol.9, no.3, pp.204–215, Nov. 2015.
[43] W. Wu, Y. Chi, S. Zhu, J. Tatemura, H. HacigĂijmĂijs, and J.F. Naughton, “Predicting query execution time: Are optimizer cost models really unusable?,” 2013 IEEE 29th International Conference on Data Engineering (ICDE), pp.1081–1092, April 2013.
[44] S. Chaudhuri, V. Narasayya, and R. Ramamurthy, “Estimating progress of execution for sql queries,” Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp.803–814, SIGMOD ’04, ACM, New York, NY, USA, 2004.

（平成 xx 年 xx 月 xx 日受付）

早水　悠登

2009 年東京大学工学部電子情報工学科卒業．2014 年同大学院情報理工学系研究科電子情報学専攻博士課程単位取得満期退学．同年，博士（情報理工学）．日本学術振興会特別研究員 DC2 を経て，現在，東京大学生産技術研究所特任助教．データベースシステムに関する研究に従事．情報処理学会，日本データベース学会，各会員

合田　和生　（正員）

2000 年東京大学工学部電気工学科卒業．2005 年同大学院情報理工学系研究科電子情報学専攻博士課程単位取得満期退学．同年，博士（情報理工学）．現在，東京大学生産技術研究所特任准教授．データベースシステム，ストレージシステムの研究に従事．本会，情報処理学会，日本データベース学会，ACM，IEEE，USENIX 各会員．

喜連川　優　（正員：フェロー）

1983 年東京大学工学系研究科情報工学専攻博士課程修了，工学博士．東京大学生産技術研究所教授．2002 年より日本データベース学会理事，2013 年 4 月より国立情報学研究所所長，2013 年 6 月より情報処理学会会長（2015 年 5 月まで）．データベース工学の研究に従事．本会業績賞，情報処理学会功績賞，ACM SIGMOD E.F. Codd Innovations Award，C&C 賞受賞．本会，情報処理学会，ACM，IEEE 各フェロー．

Abstract Data center energy consumption continues to increase. Database system plays a central role in data centers for managing and utilizing a huge amount of data, and improvement of its energy efficiency is an important issue. The authors are working on storage energy-aware cost-based query optimization for energy saving of database system. In this paper, we propose a cost estimation method of storage energy consumption considering the variation of storage configurations of a disk array, and show that the method provides good estimates with the experimental evaluation using a JBOD storage, a server, and a high precision wattmeter.

Key words database system, query optimization, disk storage, energy consumption, cost estimation