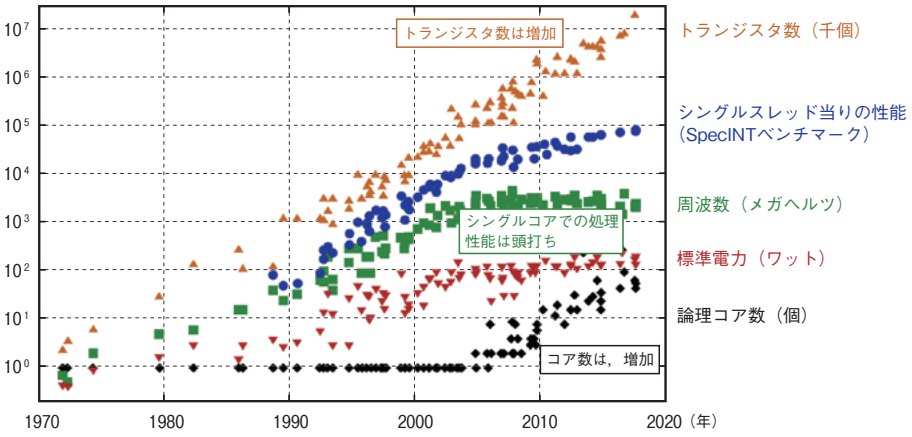
ディスアグリゲーテッドコンピューティングの実現に向けて

NTTソフトウェアイノベーションセンタでは，社会・技術の進化に対応するため，そのシステム基盤を支えるディスアグリゲーテッドコンピューティングの実現に向けた技術開発を行っています．本稿では，不揮発性メモリや高速インターコネクトのような特定用途に特化したハードウェアを活用する技術と，多数のコアを持つCPUを並列処理で性能を引き出す技術を取り上げ，ハードウェアの進化だけでなくソフトウェアの革新も重要であることを紹介します．

石崎 晃朗（いしざき てるあき）　中園 翔（なかぞの しょう）
内山 寛之（うちやま ひろゆき）　小宮 輝之（こみや てるゆき）
NTTソフトウェアイノベーションセンタ

ポストムーア時代のコンピューティング技術

半導体の集積率が18カ月で2倍になるという「ムーアの法則」が限界に達し，CPUシングルスレッドの性能は頭打ちになっています（**図1**）．ポストムーア時代を迎えてGPUやFPGAなどのハードウェアが進化して次世代のコンピュータのあり方が模索されています．一方で，CPUを中心に発展してきた



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp
This chart is provided under the permissive 'Creative Commons Attribution 4.0 International Public License'
Adjusting points are adding comments and translating the subject and the legends.
2010年までのデータとその図示は，M.Horowitz, F. Labonte, O. Shacham, K.Olukotun, L.Hammond, and C. Battenによって実施された．
2010-2017年のデータとその図示は，K. Ruppによって実施された．この図は，クリエイティブコモンズライセンス4.0に基づき，提供されている．
オリジナルの図に対して，翻訳とコメントを付加した．
出典：https://github.com/karlrupp/microprocessor-trend-dataを基に作成．

図1　マイクロプロセッサのトレンドデータ（42年間）

従来のソフトウェアでは，特定用途に特化したハードウェア向けに最適化されていないため性能を十分に発揮できないこともあります．そのため，私たちはさまざまなサービスやアプリケーションを支えるコンピューティングシステムを実現するためにはハードウェアの進化だけでは不十分であり，これらの先進的なハードウェアの能力を引き出すソフトウェアの革新との両輪によって実現することが必要と考えています．具体例として，NTT物性科学基礎研究所と取り組む光を用いたイジング型計算機という新しいハードウェア[(1)]，および，NTT先端集積デバイス研究所と取り組む光インターコネクト技術[(2)]と，それらの技術を効率的に活用するソフトウェアとの両輪により，これまで計算困難だった組合せ最適化問題の解決や情報処理システムのさらなる性能向上をめざす研究開発を実施しています．

NTTが掲げるIOWN（Innovative Optical and Wireless Network）構想では，超大容量・超低遅延・超低消費電力を特徴としたネットワークと情報処理基盤の実現をめざしていますが[(3)]，ネットワークの高速化だけではなく，データ処理にかかる処理遅延の低減化や処理の高効率も解決すべき大きな課題となっています．これらの課題を解決するためには，高速化や省電力化に限界のある従来のコンピューティングアーキテクチャを抜本的に見直し，用途に応じて多様なハードウェアをソフトウェアで柔軟に組み合わせて活用することで高速・高効率なデータ処理を行うディスアグリゲーテッドコンピューティングの実現が必要となっています．この新しいコンピューティングアーキテクチャでは，リアルワールドの多種多様なデータを安全に結び付けて，価値を創出するサービスを高速・高効率に実現できるだけでなく，電力効率を最大化させることで持続可能社会へ新たな価値を創出することも可能となります．

NTTソフトウェアイノベーションセンタ（SIC）では，ソフトウェアの変革を通じたディスアグリゲーテッドコンピューティングの実現に向けて，①CPUだけに依存せず特定用途に特化したハードウェアをソフトウェアで柔軟かつ効率的に活用してデータ処理を行う技術，②多数のコアを持つCPUをソフトウェアにより並列処理で性能を引き出す技術の開発を行っています．本稿では，①の取り組みの1つであるメモリセントリックコンピューティング技術から重要な要素技術として「不揮発性メモリを活用したストレージI/O性能向上」と「高速インターコネクトを活用したネットワークI/O性能向上」を，②の取り組みの1つである高速トランザクション処理技術から「トランザクショナルストレージライブラリLineairDBのOSS（オープンソースソフトウェア）化」について紹介します．

不揮発性メモリに関する取り組み

近年，次世代高速ストレージであるSCM（Storage Class Memory）といわれる，DRAM（Dynamic Random Access Memory）に近い遅延時間でアクセス可能

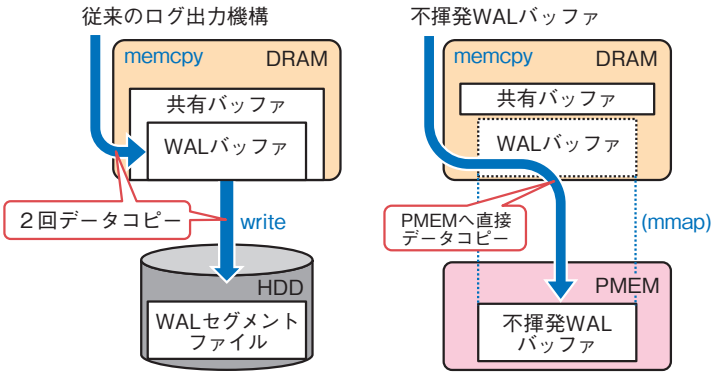


図2　PostgreSQLのログ出力機能への不揮発WALバッファの適用例

でかつ，NAND Flash SSDと同程度の大容量化が可能な，不揮発性のストレージとして利用可能なメモリデバイスが注目されています．SCMの中でもDIMMスロットへ直接挿すタイプの製品として，Intel Optane PMEM（Persistent Memory）が2019年に販売開始されています．

デバイスそのものが非常に高速なため，既存のソフトウェアを単純にPMEM上で動作させることでも「それなり」の性能が得られますが，これだけではデバイス性能を十分に引き出すことができません．その理由は，ソフトウェアが新たな特性を持つデバイスへ適応できていないためです．

従来型のHDDやNAND Flash SSDなどのストレージは，低速*かつランダムアクセスに弱いデバイス特性のため，書き込み対象のデータをDRAM上へバッファリングし，その後まとめ書きを行うようなスループット重視のソフトウェア実装が一般的です．この，DRAM上へのバッファリング・書き込み処理には主にCPUリソースが利用されています．

このソフトウェア実装モデルは，DRAMのアクセス速度とストレージデバイスとの性能差が大きいときには有利ですが，SCMは数100 nsオーダーという低い遅延時間でアクセス可能であるため，従来のソフトウェア実装モデルのままPMEM環境で実施すると，逆に遅延時間が大きくなってしまいます．

この問題に対して，リレーショナルデータベースのOSS製品であるPostgreSQLのログ出力（WAL: Write Ahead Logging）機能を対象として，DRAMと低速ストレージの2層構成となっている従来のソフトウェア構造を見直し，PMEMの特性に適応したソフトウェア実装モデルの検討を進めています．

図2は，PostgreSQLのログ出力機能について示しています．従来のPostgreSQLではDRAM上の独自バッファ機構（共有バッファ内のWALバッファ領域）へログデータ

＊ アクセス速度は，HDDは約10 ms，NAND Flash SSDは約数100 μs，DRAMは数10 ns．

がバッファリングされ，その後ログデータはストレージ領域へ書き出されます．この2層構造のログ機構を見直し，DRAMへのバッファリングをせずPMEM上へ直接データ書き込みを行うのが，提案している不揮発WALバッファです．

この方式のメリットは，2層構造で発生していた「ログデータのストレージ書き出しに伴うWALバッファ領域のロック待ち時間削減」と「データコピー回数削減によるCPU・メモリリソース削減」であり，PostgreSQLのinsert処理について約20%の性能改善を達成しています[4]．この検討は，将来的にはストレージとメモリの統合によるCPU処理削減に向けたソフトウェア実装検討と位置付けており，現在も継続的に検討を進めています[5]．

高速インターコネクトに関する取り組み

不揮発性メモリに関する取り組みにより，ソフトウェア処理のストレージI/O性能のボトルネック解消が見込めますが，次のボトルネック要因としてはネットワークI/Oが考えられます．

特に，データ交換を複数ノード間で行う分散データ処理ソフトウェアでは，ネットワークI/Oの遅延時間が致命的となりますが，HPC（High Performance Computing）の領域では，高速インターコネクトであるInfinibandとインメモリ処理を前提としてメモリの低遅延転送技術であるRDMA（Remote Direct Memory Access）を用いて，この問題へ対応しています．RDMAは，転送元サーバ上のメモリから転送先サーバ上のメモリへのデータコピーについて，CPUを介さずネットワークデバイス側で行うための技術です（**図3**）．CPUが介在しないことで処理受け渡しの遅延時間削減や，TCP/IPのプロトコル処理などもないため，低遅延での転送処理が可能なのが特徴です．私たちは，主にHPC領域で用いられてきたこの低遅延転送技術について，エンタープライズ向けソフトウェアへの適応をめざしており，RDMAの基礎評価や分散学習フレームワークであるMXNet等を対象としてソフトウェア実装モデルの検討を進めています[6]．RDMAによる低遅延転送処理はFPGAやGPUなどのハードウェア上のメモリにおいても，実行可能となりつつあり，この検討は，将来的にはネットワーク処理に関するCPU処理の削減に重要な技術と位置付け，継続的に検討を進めています．

LineairDB：高速トランザクショナルストレージライブラリのOSS化

ディスアグリゲーテッドコンピューティングにおいて多数のコアを持つCPUを並列処理で性能を引き出す技術として，144コアを持つCPUにおいて処理性能がスケールするメニーコア向けトランザクション処理技術に取り組んできました[7]．本技術をベースとした，トランザクショナルストレージライブラリとしてLineairDBを開発し，2020年4月にOSSとして公開しました[8]．近年，CPU

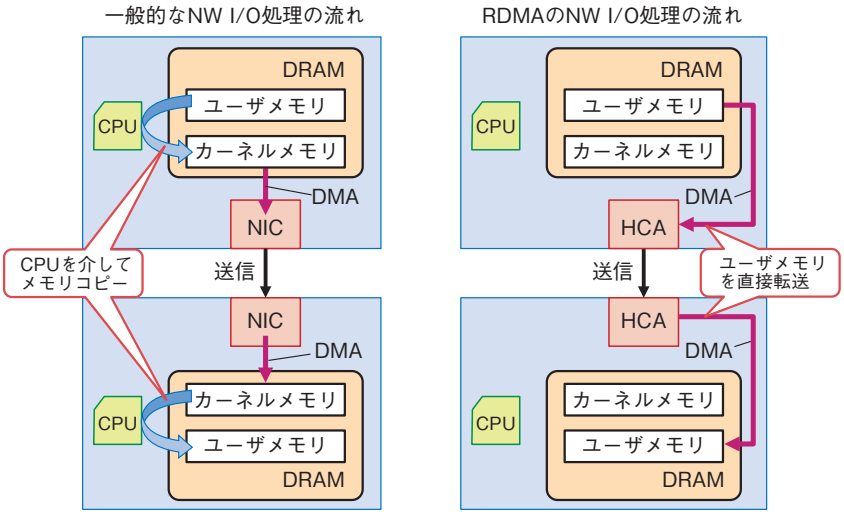


図3　一般的なネットワーク処理とRDMAを活用したネットワーク処理

上のトランジスタ数は増加していますが，それらはCPUのコア数を増加させることで実現されています（図1）．1980年代に基本的な技術群が完成した既存のデータベース管理システムの多くは，多数のコアを持つCPUを想定していないアーキテクチャに基づいています．そのため，コア数に対して処理をスケールすることができず，コア数を増やすと処理スループットが減少することが知られています[9]．昨今の研究により，読み込みに対してこの問題を解決する手法が提案されてきましたが，書き込みに対しては優れた解決手法がなく既存技術を利用せざるを得ませんでした．LineairDBでは，コア数が増加した場合でも書き込みの処理性能をスケールすることができる高速トランザクション処理技術を搭載しており，メニーコアCPUの能力を最大限引き出すことが可能です．具体的には，144個のコアを持つCPUを搭載したマシン上において，キーバリューストアに対するベンチマークとして有名なYCSB（Yahoo! Cloud System Benchmark）-A［Read50%，Write50%］に対して，既存技術の約3倍となる1秒間に1000万トランザクションの処理を実現しました[8]．

LineairDBはキーバリューストアとしてお使いいただけるインタフェースを用意しており，さまざまなシーンでご利用いただけます．また，ライセンスは，Apache License, version 2.0を採用しており，さまざまなライセンスとの親和性も高く利用しやすいため，是非一度お試しください．LineairDBコミュニティには，slackを通じたコミュニケーションも用意しています．疑問や要望などありましたら，お気軽にご相談ください．また，開発者として参画いただける方も大歓迎です．今後は，より利用しやすいインタフェースの整備やレンジクエリへの

対応を図り，さまざまなユースケースで利用いただける機能拡張を行っていきます．

今後の展開

私たちはソフトウェアの革新を強みとして，高速・不揮発な共有メモリを介して多様なハードウェアが直接データ交換するメモリセントリックコンピューティング技術や，メニーコアCPUの能力を最大限引き出すことが可能な高速トランザクション処理技術の確立により，社会・技術の進化に対応できるシステム基盤を支えるディスアグリゲーテッドコンピューティングを実現します．

■参考文献

⑴ 新井・八木・内山・冨田・宮原・巴・堀川：“イジング型計算機による組合せ最適化のためのハイブリッド計算基盤，”NTT技術ジャーナル，Vol. 31，No. 11，pp. 27-31，2019.
⑵ Focus on the News：“光通信で培った技術を応用し，情報処理システムの性能向上を実現する「光インターコネクト技術」を開発，”NTT技術ジャーナル，Vol. 31，No. 11，pp. 65-66，2019.
⑶ 特集：“IOWN構想特集―オールフォトニクス・ネットワーク実現に向けた光電融合技術―，”NTT技術ジャーナル，Vol. 32，No. 8，pp. 6-28，2020.
⑷ https://www.slideshare.net/ntt-sic/wal-234538063
⑸ https://www.postgresql.org/message-id/002f01d5d28d$23c01430$6b403c90$@hco.ntt.co.jp_1
⑹ https://www.slideshare.net/ntt-sic/rdma-programming-design-and-case-studies-for-better-performance-distributed-applications
⑺ 中園・内山：“メニーコア向け高速トランザクション処理技術，”NTT技術ジャーナル，Vol. 31，No. 11，pp. 32-35，2019.
⑻ https://github.com/LineairDB/LineairDB
⑼ X. Yu, G. Bezerra, A. Pavlo, S. Devadas, and M. Stonebraker: “Staring into the Abyss: An Evaluation of Concurrency Control with One Thousand Cores，” Proc. of VLDB Endowment, Vol. 8，No. 3，pp. 209-220，2014.

（上段左から）小宮　輝之／石崎　晃朗
（下段左から）中園　　翔／内山　寛之

ハードウェアの進化とソフトウェアの革新の両輪によって既存のコンピューティングアーキテクチャのパラダイム変化を起こし，これまで困難であった大規模高速データ処理を高効率かつ省電力化することで，IOWN構想社会の実現をめざしていきます．

◆問い合わせ先
NTTソフトウェアイノベーションセンタ
企画担当
TEL　0422-59-2207
FAX　0422-59-2072
E-mail　sic＠hco.ntt.co.jp