光ディスアグリゲーテッドコンピュータにおいて電力効率最大化を実現するパワーアウェア動的配置制御技術

IOWN（Innovative Optical and Wireless Network）により実現される光ディスアグリゲーテッドコンピュータは広域に分散した多様な計算デバイス（省電力CPU，アクセラレータ等）により構成されます．本稿ではネットワークワイドな処理のオフローディングや電力最適な計算デバイス選択等のコンピューティング基盤制御により，光ディスアグリゲーテッドコンピュータにおける省電力化・電力効率最大化を実現するパワーアウェア動的配置制御技術を紹介します．

金子 雅志　藤本 圭
岩佐 絵里子
NTTネットワークサービスシステム研究所

IOWN時代のサーバシステムにおける電力面の課題

近年IPトラフィックは指数関数的に増加しており2050年には現在の4000倍になると予測されています[(1)]．一方でそれらを処理するサーバについては，CPU（Central Processing Unit）の性能向上スピードが鈍化していることもあり，現状のサーバハードウェアの延長線上での電力効率で試算するとデータセンタ運用に要する電力も2050年に4000倍になると予想されています．このため，サーバの省電力化・電力効率向上は将来のネットワーク運用に向けた重要な課題になります．

IOWN（Innovative Optical and Wireless Network）においては，オール光ネットワークで接続された計算デバイスによるインネットワークでの高速かつ効率的な処理を可能とするネットワークワイドな新たなコンピュータアーキテクチャとなる光ディスアグリゲーテッドコンピュータの実現をめざしています．光ディスアグリゲーテッドコンピュータにおいては，光電融合デバイスやメモリセントリックコンピューティングといった新たなデバイスや処理方式の導入により，従来のサーバハードウェアの限界を超える電力性能を実現し，将来にわたる電力面の課題を解決することが期待されます．

光ディスアグリゲーテッドコンピュータとパワーアウェア動的配置制御技術

光ディスアグリゲーテッドコンピュータを構成する計算デバイスとしてはCPUやアクセラレータ〔GPU（Graphics Processing Unit），FPGA（Field Programmable Gate Array）等〕のような多様なデバイスが接続されることを想定しており，分散型のヘテロジニアスコンピューティングが実現されます．そのようなコンピューティング基盤上で，ネットワークを構成する仮想的な機能群であるVNF（Virtual Network Function）を稼動させる場合，VNFを構成するソフトウェアコンポーネント〔従来型のVNFであれば仮想マシン（VM）相当〕を各計算デバイスに対して割り当てていく制御機構が必要になります．従来型のVNFでは，アクセラレータを必要とするVMについてはアクセラレータが物理的に搭載されているサーバハードウェアに配置される必要があります．また，一般的なVMではアクセラレータをデバイス（GPUボード，FPGAボード等）単位で排他的に利用するのが一般的であるため，アクセラレータにオフロードする処理の割合が少ないアプリケーションがデプロイされたサーバにおいては，アクセラレータのハードウェアリソースが有効活用されません．光ディスアグリゲーテッドコンピュータにおいては，CPUやアクセラレータといった計算デバイスが光パスで接続されることで物理的に離れたデバイス間において高速低遅延でのインタラクションが可能となり，従来のようなサーバのフォームファクタによる構成の制約（例：ラックマウントサーバに搭載可能な拡張ボード数等）が緩和されます．また，

VNFとアクセラレータの関係を従来のように1対Nではなく，M対Nに柔軟化することができれば，1つのアクセラレータに対して複数VNFの処理を割り当てることが可能となり，アクセラレータのハードウェアリソースが最大限活用可能となります．

NTTネットワークサービスシステム研究所で研究しているパワーアウェア動的配置制御技術（**図1**）は，サーバ基盤に対してソフトウェア制御を行うことにより消費電力の低減に貢献する技術です．現在は，分散されたアクセラレータデバイスに対してネットワークワイドでの処理オフロードを可能にすること，複数CPU上で稼動するソフトウェアコンポーネントからアクセラレータへのアクセスが同時並列で発生した場合に従来のような排他的なアクセスではなく並列での処理受付を可能とすることでアクセラレータデバイスの共用化を実現し，アクセラレータの稼動率を向上させる「パワーアウェアアクセラレータ最適活用・高速低遅延技術」，並びに省電力CPUや各種アクセラレータ等の多様な計算デバイスが混在する環境において，デバイス単位・ソフトウェアコンポーネント単位での細やかな消費電力分析と，消費電力を最小化可能なソフトウェアコンポーネント配置および個々のデバイスの省電力制御により電力低減を図るとともに，将来的には地域ごとの再生可能エネルギー発電状況なども考慮したかたちでサーバ基盤の消費電力を調整することで電力需要をコントロールし，不安定な再生可能エネルギー発電から得られる電力を最大限活用可能な「再生可能エネルギー活用型サーバ基盤技術」の技術確立に取り組んでいます．

パワーアウェアアクセラレータ最適活用・高速低遅延技術

近年，データ分析や機械学習のような複雑な計算処理に対してCPUだけでなく，GPUやFPGAのようなアクセラレータを活用するケースが増えてきています．アクセラレータは一般的に処理に対する汎用性が低い代わりに，得意な処理に対してはCPUに対し100倍以上の効率で処理が可能である場合があります．現状，アクセラレータを利用する際にはCPUで実行されているプログラムの一部をOpenCL等のAPI（Application Programming Interface）を通してアクセラレータにオフロードする利用方法が一般的です．この場合，プログラムを処理しているCPUと同一サーバハードウェア上に空き状態のアクセラレータが存在しなければ処理をオフロードさせることができません．

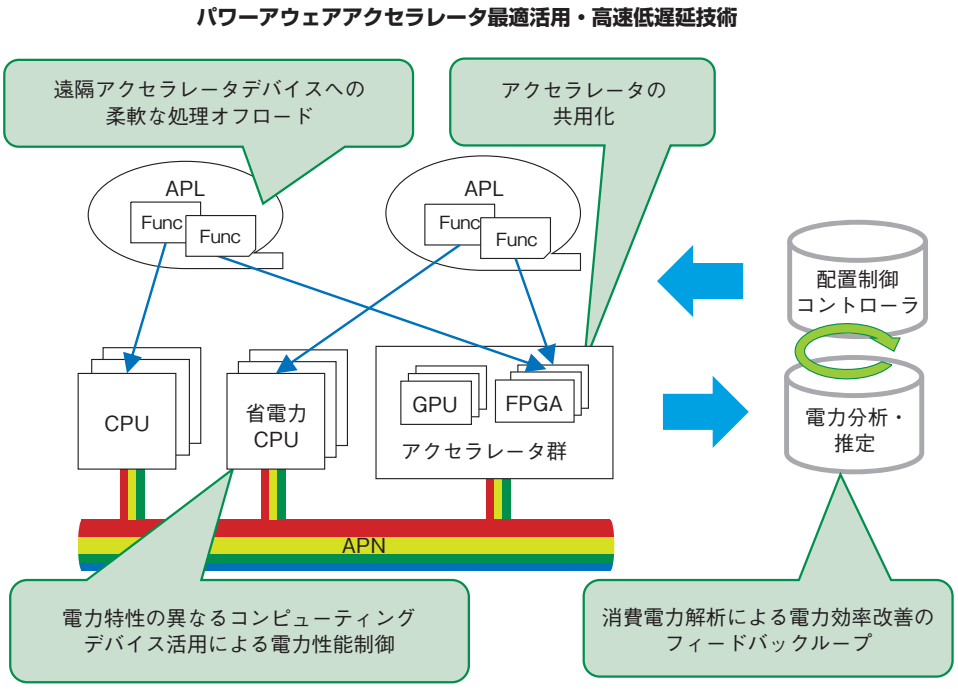

図1　パワーアウェア動的配置制御技術

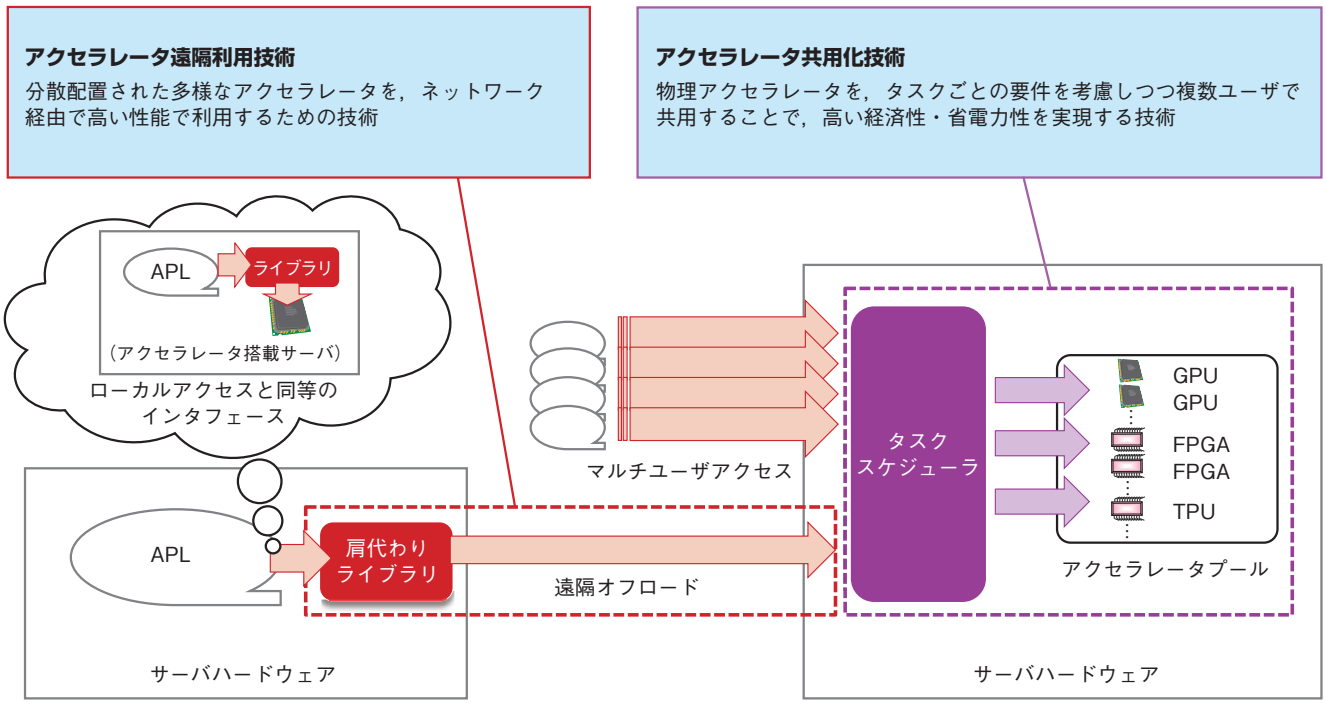

図 2　パワーアウェアアクセラレータ最適活用・高速低遅延技術

パワーアウェアアクセラレータ最適活用・高速低遅延技術では，従来技術のように物理的な接続構成に縛られず，ネットワーク上に分散配置された遠隔のアクセラレータを，まるでローカルに配置されているように位置透過的かつオーバヘッドが少なく高速・低遅延に利用可能とすること，および複数ユーザからアクセラレータを共用利用可能とすることでアクセラレータの利用効率をさらに高める技術です（**図2**）．遠隔のアクセラレータへの処理オフロードにおいては，既存のオフロードAPI（例：OpenCL API）を提供する肩代わりライブラリをオフロード元に組み込み，アプリケーションには従来のローカルアクセスと同等のインタフェースを提供することで位置透過性を実現するとともに，既存アプリケーションのポータビリティ向上を図ることが可能です．

再生可能エネルギー活用型サーバ基盤技術

現在のサーバハードウェアは常時安定的な電力が給電されることを前提に設計されているため，多数のサーバを稼動させるデータセンタでは，常に大容量の電力を安定的に供給できる必要があります．近年，データセンタへの給電に再生可能エネルギーを活用する動きがありますが，前述のとおり現状のサーバは常に安定的な電力供給が必要であるため，太陽光発電のような発電量が不安定な再生可能エネルギーのみでの長時間安定稼動は困難です．再生可能エネルギー活用型サーバ基盤技術ではサーバが消費する電力の分析・推定により，計算デバイス単位での電力制御やシステムを構成するソフトウェアコンポーネントの配置制御により，サーバ基盤トータルでの省電力化および給電量に応じた稼動制御を実現します．

一般的なサーバハードウェアは低負荷時でも高負荷時の6～7割程度の電力を消費するといわれていますが，スマートフォンに搭載されるSoC（System-on-Chip）においては低負荷時の消費電力を極力低下させるため，負荷に応じたCPUの電力ステート制御や周波数制御に加え，低電力稼動に適した専用CPUコアを搭載するなどして，スマートフォンにおける“省電力モード”のような電力供給（バッテリ蓄電）状況に応じた稼動制御を実現しています．サーバ基盤においても，電力供給量に応じて性能を変動させることで電力の需要を調整することができれば，時々刻々と変動する再生可能エネルギー発電を無駄なく利用可能なサーバ基盤を実現することができます．また別のアプローチとして，サーバクラスタにおいて電力性能の異なるサーバを複数種類組み合わせたヘテロクラスタを構成し，稼動状況により最適なサーバを動的に選択することで電

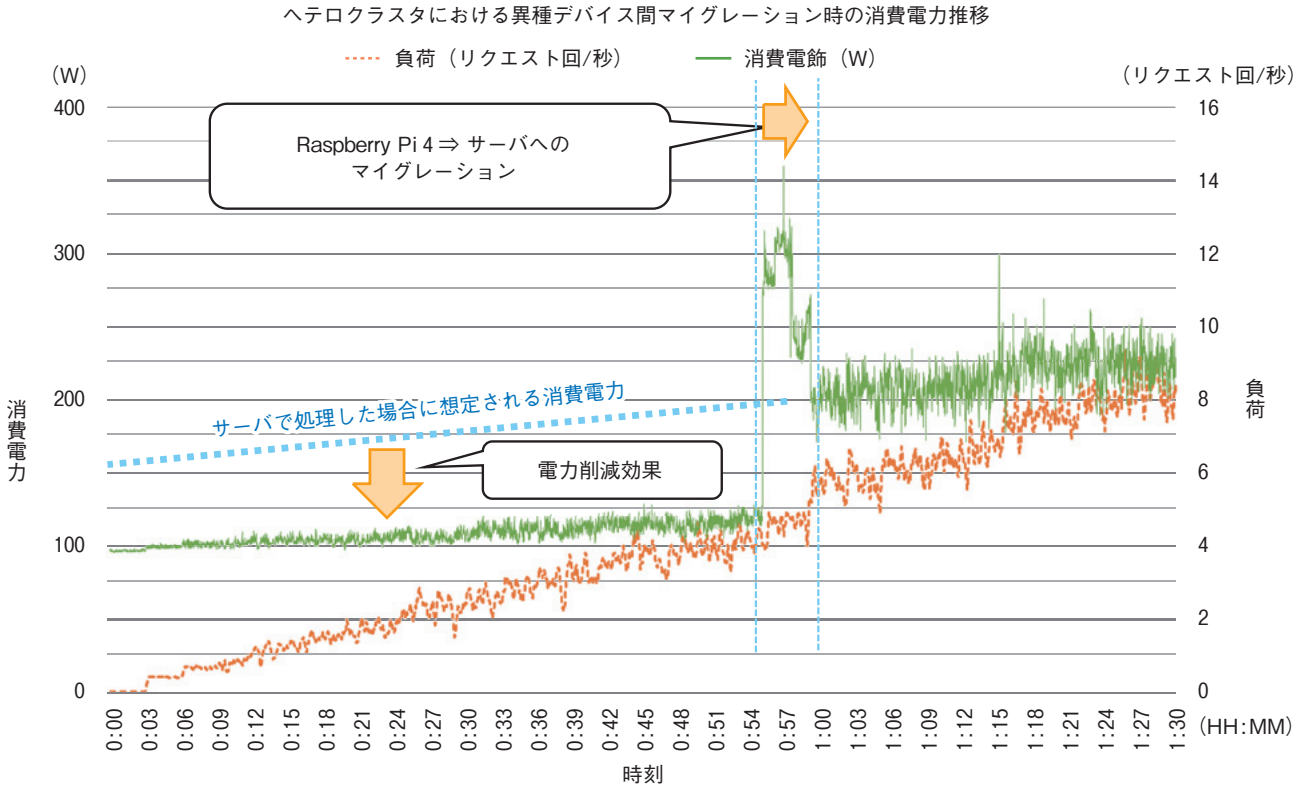

図 3　ヘテロクラスタにおける省電力化制御の一例

力効率の向上を図ることが可能となります．一例として，一般的なサーバハードウェアとRaspberry Piのような最大性能は低いものの電力性能に優れるコンピューティングデバイスとを組み合わせたヘテロクラスタを実現し，負荷状態に応じて最適なサーバを選択し，動的にシステムマイグレーションすることで低負荷時の電力向上を図ることができます．**図 3** のグラフは15台のRaspberry Pi 4と 1 台の1Uラックマウントサーバでサーバクラスタを構成し，負荷に応じてソフトウェアコンポーネントマイグレーションさせたときの消費電力を表しています．Raspberry Piクラスタで処理を行っている低負荷状態（1Uラックマウントサーバは OFF 状態）では，サーバのアイドル時の電力（およそ150 w）よりも低い電力で処理を実施できていることが分かります．このように，電力的に特性の異なる複数のコンピューティングデバイスを組み合わせ，負荷に応じたシステムを構成するソフトウェアコンポーネントの配置制御を実施することで，幅広い負荷レンジにおいて電力効率を向上させることが可能となります．

今後の展開

私たちは今後，深刻化が懸念されるITサービスが消費する電力増大の問題に対して，ディスアグリゲーテッドコンピュータを効率的に制御することにより解決をめざすパワーアウェア動的配置制御技術の検討を進めています．従来のような性能やコストだけでなく，電力効率や再生エネルギー活用も併せて追求することで社会が求める低炭素社会の実現に貢献できるよう技術開発を進めます．

■参考文献
(1) https://www.jst.go.jp/lcs/pdf/fy2018-pp-15.pdf

（左から）金子 雅志/ 藤本 圭/
岩佐 絵里子

光ディスアグリゲーテッドコンピュータの実現を通してオール光ネットワークによる新時代のサービスを実現するサーバ基盤を実現するとともに，電力面の課題を解決することでIOWNが掲げる電力100分の 1 目標の達成をめざします．

◆問い合わせ先
NTTネットワークサービスシステム研究所
ネットワーク制御基盤プロジェクト
TEL　0422-59-7718
E-mail　e2serv-p-ml @ hco.ntt.co.jp