URL　https://journal.ntt.co.jp/article/23718
DOI　https://doi.org/10.60249/23115103

IOWN時代のデータ処理を支えるデータセントリック基盤とそのコンセプト実証

データセントリック基盤は，広域に散在するデータを高効率に処理するためのICT基盤であり，データ駆動型社会における大規模サイバーフィジカルシステム（CPS）での活用が期待されています．本稿では，同基盤のコンセプトである，アクセラレータやオールフォトニクス・ネットワークを効果的に用いたデータ処理パイプラインの提供について概説します．次に，CPSにおける映像解析をユースケースとした，同コンセプトの段階的な実証について紹介します．

キーワード：#データセントリック基盤，#ディスアグリゲーテッドコンピューティング，#アクセラレータ

樽林（くればやし）　亮介（りょうすけ）　/石崎（いしざき）　晃朗（てるあき）
Sampath Priyankara
Christoph Schumacher
水野（みずの）　伸太郎（しんたろう）
NTTソフトウェアイノベーションセンタ

データ駆動型社会の実現に向けて

近年のセンシング技術の向上とネットワーク化，デジタルトランスフォーメーション，そしてAI（人工知能）技術の進展により，データ駆動型社会が到来しようとしています．データ駆動型社会では，フィジカル空間（現実世界）・サイバー空間（コンピュータ）上のさまざまなデータを，業界や分野の枠を超えて幅広く流通・掛け合わせていくことで，社会課題を解決したり，新たな価値の創出をめざします．このデータ駆動型社会における中心技術の1つがサイバーフィジカルシステム（CPS：Cyber-Physical System）です．CPSとは，フィジカル空間から得られる膨大なデータを，サイバー空間上で分析し，その結果をフィードバックすることで現実世界の最適な制御をします．これまでに工場・プラントのスマート化，交通の最適化など，特定分野での活用が始まっています．データ駆動社会では，このCPSを大規模化させ，あらゆる分野に適用したり，相互に連携させることが求められます．

CPSの大規模化に向けてはさまざまな技術課題が存在しますが，本稿では特に，データ処理を担うICT基盤に注目します．これまでのCPSでは，目的・処理方法ごとにサイロ化された個別システム内でのデータ処理をすれば十分でした．これが，現状よりはるかに多く，地理的にも分散した主体（データを流通する人間・システム・デバイス等）間でデータを高速に流通させ，より多種大量のデータを分析していくことになります．そのためのICT基盤には，地理的に分散する主体と分析を担う計算リソースとの間を高いQoS（Quality of Service）でネットワーク接続し，膨大な演算コストの要求に対してボトルネックのない高効率なデータ処理パイプラインを提供していくことが求められます．

データセントリック基盤

大規模CPSの要件を充足可能なICT基盤として，NTTは，IOWN（Innovative Optical and Wireless Network）技術を活用したDCI（Data-Centric Infrastructure subsystem：データセントリック基盤）の検討を進めています．DCIは，フィジカル空間・サイバー空間において広域に散在するデータに対して，同じく広域に分散する計算リソースを最適に組み合わせながら，効率良くデータ処理していくためのICT基盤です．**図1**にDCIを用いた大規模CPSの実現イメージを示します．

本稿では，DCIの以下の特徴について概説します．

- アクセラレータを活用した高効率データ処理パイプライン
- APN（All-Photonics Network：オールフォトニクス・ネットワーク）との統合
- オープンなエコシステムの形成

■アクセラレータを活用した高効率データ処理パイプライン

DCIの特徴として，まず，アクセラレータを活用した高効率データ処理パイプラインが挙げられます．従来のデータ処理はホストCPU上でのソフトウェア処理が中心でした．一方で，近年では，計算コストが高い領域において，さまざまなアクセラレータが活用されるようになっています．AI・メディア処理におけるGPU（Graphics Processing Unit），ネットワーク処理に対するSmart NIC（Network Interface Card）/IPU（Infrastructure Processing Unit）/DPU（Data Processing Unit）がその典型です．これらのアクセラレータでは，特定領域における並列処理を高効率化したり，ASIC（Application-Specific Integrated Circuit）やFPGA（Field-Programmable Gate Array）等のハードウェアを用いて高速化を図ります．DCIでは，これらのアクセラレータを積極的に活用することで，データ処理パイプラインを高効率化します．

一方で，従来技術の延長でのアクセラレータ活用では，ホストCPUを起因としたボトルネックの軽減が課題となります．**図2**のとおり，従来は，アクセラレータを用いたデータ処理であってもホストCPUが間に介在します（図2(a)）．このため，ホストCPUがボトルネックとなり，実効的に扱えるアクセラレータ数も制限される結果となっています．

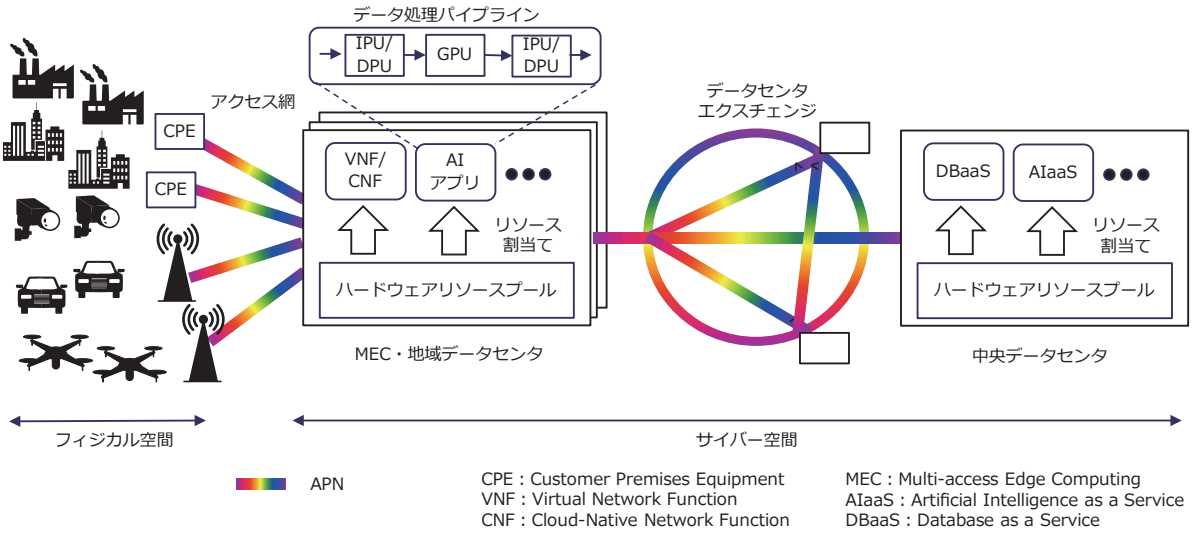


図 1　DCI を用いた CPS の実現イメージ

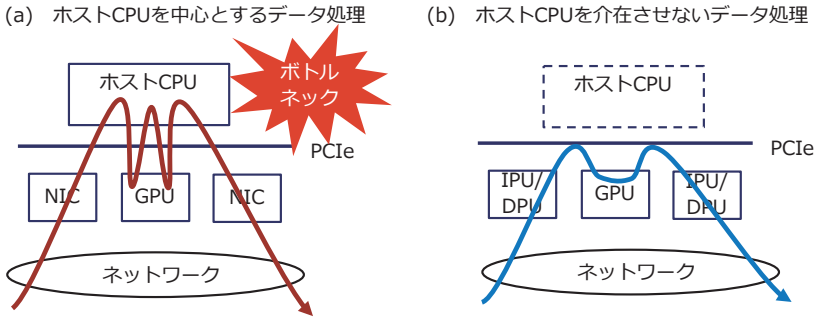


図 2　ホスト CPU を介在させないデータ処理

そこで，DCI では，ホスト CPU を介するデータ処理パイプラインから，アクセラレータ間でより自律的にデータを転送し処理を進める方式（図 2 (b)）への転換を図ります．そのために，データ転送に際してホスト CPU を介さないでリモートのメモリ上にデータを展開できる RDMA（Remote Direct Memory Access），ASIC・FPGA 等を用いた通信プロトコル処理のハードウェア化，アクセラレータ間の直接データ転送，等を検討しています．さらには，データ到着の検知や，処理の開始指示といった実行制御についても，アクセラレータ自体にオフロードする技術を検討しています．

アクセラレータを活用したデータ処理パイプラインでは，そのワークロードにあったアクセラレータを適切に選択・組み合わせることが不可欠です．一方で，ワークロードに応じて必要となるアクセラレータの種類・数は大きく異なります．この結果，サーバを基本単位とする従来の基盤構築では，アクセラレータに多くの無駄が生じます．すなわち，画一的なアクセラレータ構成のサーバを並べる場合，ワークロードによって利用されるアクセラレータと利用されないアクセラレータが生じます．一方で，あらゆるワークロードに対応できるよう，さまざまなパターンのアクセラレータ構成を持った多種類のサーバを事前に用意することも現実的ではありません．

このため，DCI では，**図 3** に示す，ハードウェアリソースプールを活用します．ハードウェアリソースプールでは，従来サーバを構成していた各種デバイス（アクセラレータを含む）を，高速なインターコネクトを介して接続しプール化します．そして，与えられたワークロードに対して，最適なデバイスを最適な数だけプールから選択し割り当てます．また，不要となったデバイスは，プールに戻して再利用可能な状態にしたり，電源オフの対象とします．このように，ハードウェアリソースプールを用いることで，アクセラレータの柔軟な選択や再利用が可能となり，従来のサーバ単位の基盤と比較し，デバイスの利用率を飛躍的に高めることができます．

ハードウェアリソースプールは，いくつかの異なるレイヤで実現できます．第一に，ハードウェアレベルでベアメタルサーバとして切り出す方法があります（図 3 ①）．市中技術としても，インターコネクトとして PCIe（Peripheral Component Interconnect-Express）を拡張し，ホスト CPU に対して任意のデバイスを紐付ける CDI（Composable Disaggregated Infrastructure）製品が提供され始めています[(1)]．また，今後，次世代インターコネク

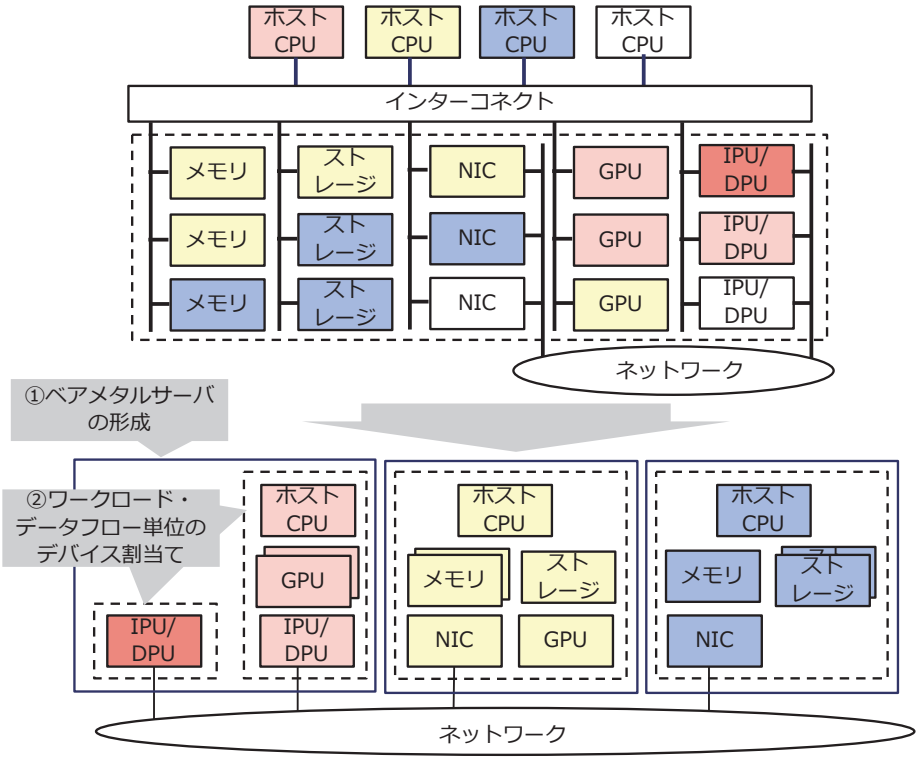


図3　ハードウェアリソースプールの概念

ト規格であるCXL（Compute Express Link）[2]が普及していくことで，規格の統一や機能の拡充が期待されています．第二に，特定の用途やテナント向けに必要なデバイスをまとめて接続したベアメタルサーバに対して，ソフトウェアレベルで適切にリソース管理をしながら，より細粒度なワークロード・データフロー単位でデバイスを割り当てる方法があります（図3②）．特に，ホストCPUを介在させないデータ処理パイプラインが主流になると，1つのホストCPUが実効的に扱えるアクセラレータの規模も拡大します．この結果，図3②のようなソフトウェアレベルでのリソース管理の重要性が増していくと考えられます．DCIの実現にあたっては，ユースケースや技術の普及をみながら，これらの手法を柔軟に組み合わせていきます．

■APNとの統合

DCIでは，フィジカル空間とサイバー空間をつなぐアクセス網，そして各データセンタ間の接続（データセンタエクスチェンジ）に，高速・高品質なAPNを適用します．この特徴は，デバイス～データセンタ間の広域にまたがったデータ処理パイプラインをボトルネックなく構成することに役立ちます．また，将来的には，APNのような光ネットワークの技術を，広域ネットワークだけでなく，データセンタ内ネットワークや，ハードウェアリソースプールのインターコネクトまで適用していくことが想定されています．

■オープンなエコシステムの形成

DCIの実現には，ネットワーキングからコンピューティングまで，そしてハードウェアからソフトウェアまでの幅広い分野にまたがった技術の再構築が必要です．その実現に向けては，多くの企業が参画しそれぞれの知見を結集できるオープンなエコシステムを確立していくことが重要です．このため，DCIは，IOWN GF（Global Forum）といったグローバルコミュニティの中でも議論され，その具体化と合意形成が進められています．IOWN GFにおけるDCIの議論では，主にハードウェアリソースを活用した効率化や管理に力点が置かれています．すなわち，特定用途向けに割り当てられたデバイスの集合である論理サービスノード（典型的には図3①のベアメタルサーバに相当）の形成と論理サービスノード間のQoSを考慮したネットワークリソースの割当てが検討されています．また，論理サービスノード間のデータ転送の高速化に向けて，IOWN GFにて検討を進めるOpen APN越しでのRDMA技術の適用が議論されています．これらの成果をまとめるかたちで，DCI機能アーキテクチャ文書の第二版[3]が発行されています．

コンセプト実証

NTTソフトウェアイノベーションセンタでは，DCIの特徴である，アクセラレー

タを用いた高効率データ処理パイプラインを実現する技術の1つとして，ディスアグリゲーテッドコンピューティングを研究開発しています．そして，CPSの大規模化に向けて，ディスアグリゲーテッドコンピューティングを適用したPoC（Proof of Concept）を段階的に進めています．本稿では特に，CPSにおけるリアルタイム映像解析を対象とした，以下の2つのPoCについて紹介します（**図4**）．

・PoC-1：多様なアクセラレータを用いた映像解析パイプライン
・PoC-2：as a Service化を実現するディスアグリゲーテッドコンピュータコントローラ

■PoC-1：多様なアクセラレータを用いた映像解析パイプライン

本PoCでは，多様なアクセラレータを組み合わせたデータ処理パイプラインの効果を実証します．具体的には，図4の映像解析部に着目し，8台の4Kカメラを用いて人物検出を行います．そして，昼・夜の人流の変化に従って，人物検知をするデータ処理パイプラインを最適構成することで，その消費電力を削減します．本PoCは，DCIに関するPoC計画書[4]に基づき実施されたPoCとして，IOWN GFに公式に認められています．

本PoCにおける映像解析部の構成を**図5**に示します．図5に示すとおり，カメラ

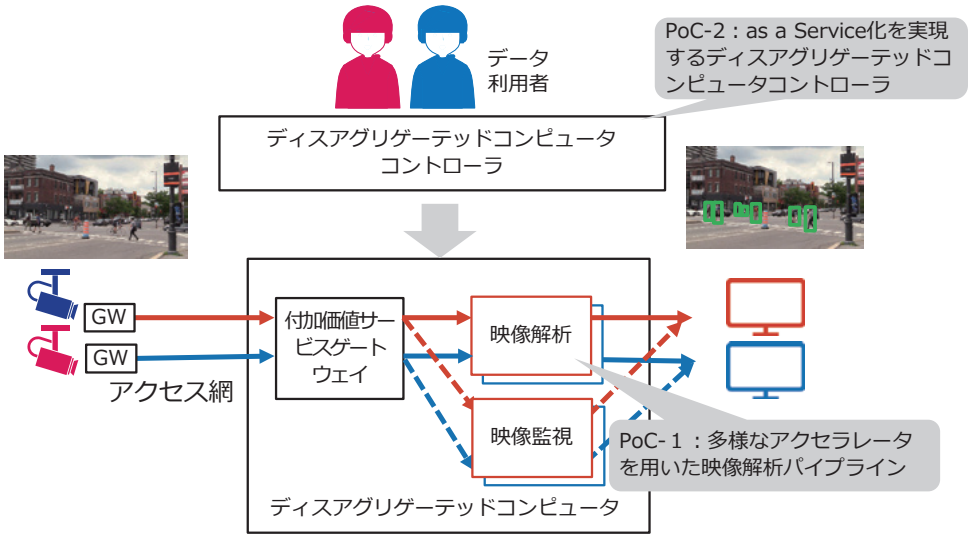


図4　ディスアグリゲーテッドコンピュータを用いたコンセプト実証

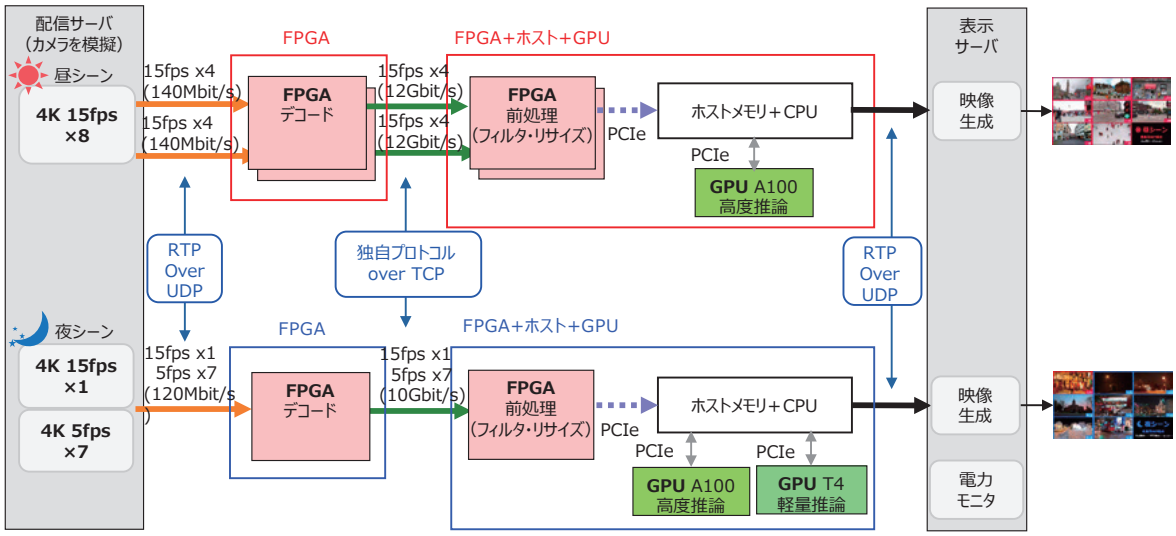


図5　映像解析部の構成図

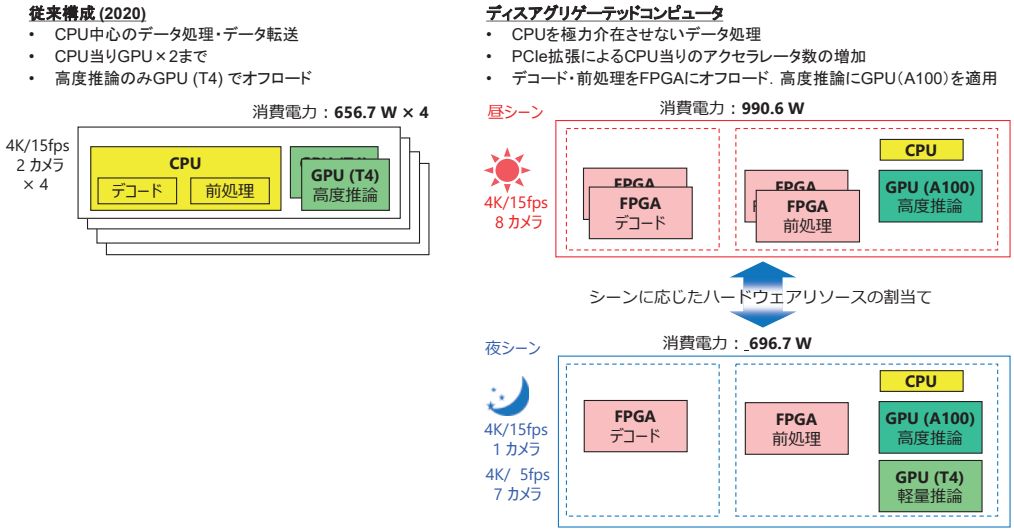


図 6　従来構成との電力比較

映像を入力として受け付け，デコード，前処理としてフィルタ・リサイズを施した後，映像推論として人物を検出します．また，映像推論では，以下のとおり，高度推論と軽量推論とを使い分けることを想定しています．

- 高度推論：人が映っているカメラに対して，高精度な人物検知を行います．検知精度を優先し，高フレームレートで高画質な画像を基に推論します．
- 軽量推論：人が映っていないカメラに対して，少ない電力により粗確認レベルの人物検知を行います．電力効率を優先し，低フレームレートで低画質な画像を基に推論します．

昼間は人が多いので，結果的に高度推論をするカメラが多くなり，一方で，夜間は人が少ないので，軽量推論をするカメラが多くなります．このように，昼間と夜間では求められるワークロードの比率が異なることから，各シーンに必要な処理や演算負荷に合わせてデータ処理パイプラインを構成します．**図 6** の上部が昼シーン用，下部が夜シーン用のデータ処理パイプラインとなります．また，本映像解析部は，多様なアクセラレータの活用というコンセプトを実証するため，FPGA と GPU を用いて処理パイプラインを構成しています．また，FPGA は，前述の CPU を介在させないデータ処理を行うため，NIC の機能およびデータ転送に特化した独自回路を搭載しています．すなわち，デコードと前処理には FPGA が利用され，推論処理には GPU が適用されます．また，GPU も，高度推論に対しては高パフォーマンスの GPU（NVIDIA A100）を，軽量推論に対しては省電力な GPU（NVIDIA T4）を選択します．図 6 に示すとおり，昼シーン用のデータ処理パイプラインは，より多くのカメラに対して高フレームレートの高度推論を実施するため，より多くの FPGA が割り当てられます．

次に，アクセラレータ間の通信方法について述べます．カメラからの映像符号化ストリームは RTP over UDP を用いてデコード処理部に入力されます．そして，後段の前処理部に対して，復号化した映像を TCP/IP 上の独自プロトコルで転送します．RTP over UDP，および TCP/IP 上の独自プロトコルは，それぞれ FPGA 上で終端され，必要なデータが，FPGA のユーザ回路向けのメモリ上に展開されます．これにより，プロトコル処理のハードウェアオフロードの効果を得ます．特に，デコード処理を行う FPGA では，映像の受信～デコード～送信までの一連の処理を FPGA 内で完結させており，ホスト CPU を介さない自律的なデータ処理を実現しています．また，前処理を行う FPGA から推論処理を行う GPU へのデータ転送も，ホストメモリを経由するものの，ホスト CPU のオーバヘッドの少ない，独自 DMA ベースでのデータ転送を行っています．

本映像解析部の消費電力の測定結果を示します．比較対象とする従来構成（2020年の典型的な構成を想定）と本映像解析部の構成，およびそれぞれの消費電力を図 6 に示します．本映像解析部の昼シーンと従来構成とを比べると，約62%の電力が削減されています．これは，本映像解析部では，ハードウェアの進化（NVIDIA T4から A100への変更）を含む最適なアクセラレータの選択，デコード・前処理を含むより広い範囲でのアクセラレータの活用，アクセラレータ間の高効率なデータ転送，といった高効率化が図られているためです．さらに，昼シーンと夜シーンの電力を比較すると，昼シーンが990.6 W であることに対して，夜シーンの電力は696.7 W に抑えられることが確認できました．このことから，シーンに応じてパイプライン構成を柔軟に

変更することで，従来構成と比べて約73%の電力を削減できるといえます．

なお，本映像推論部はPoCを目的とし，以下の制限があります．

- FPGA上の独自回路はプロトタイプ実装であり，実装の改善によりさらなる電力削減が期待できます．
- 昼・夜シーンの切り替えは静的にオフラインで実施しています．実用化に際しては動的な切り替えに向けた拡張が必要となります．

■PoC-2：as a Service化を実現するディスアグリゲーテッドコンピュータコントローラ

PoC-2では，ディスアグリゲーテッドコンピュータコントローラを用いることで，高度な専門知識を持たないデータ利用者であっても，アクセラレータを活用したデータ処理パイプラインを容易に構成可能であることを示します．本PoCは「NTT R&Dフォーラム 2023」においても展示されます．

本PoCでは，図4に示すデータ処理パイプラインを，データ利用者からの要求に応じて生成したり，構成変更したりします．ここで，本PoCの映像解析部として，PoC-1の昼シーンを用います．また，映像監視部では，監視者向けに映像を適切に転送します．本PoCでは，複数のテナントの存在を仮定します．そして，映像を集約分配する付加価値サービスゲートウェイ[5]部を用いて，テナントごとにカメラと映像解析・監視部との間を結ぶセキュアな接続を提供します．すなわち，テナントごとに映像を分離できるように，カメラとの接続にGRE（Generic Routing Encapsulation）トンネルが張られます．そのうえで，各テナントのデータ利用者からの要求に応じて，映像データを適切に複製し，映像解析，映像監視といった後段の処理部に分配します．この分配を，GREトンネルからVLAN（Virtual Local Area Network）への変換によって実現します．

本PoCでは，ディスアグリゲーテッドコンピュータコントローラが重要な役割を果たします．ディスアグリゲーテッドコンピュータコントローラは，データ利用者からのデータ処理パイプラインの生成や構成変更の要求を受け付けると，映像解析部・監視部のデータ処理パイプラインの構築，付加価値サービスゲートウェイ部を含む設定変更までを自動的に実施します．このとき，データ利用者は，要求するデータ処理パイプラインの設計図を，アクセラレータの詳細が抽象化されたYAMLファイルで記述できます．また，ディスアグリゲーテッドコンピュータコントローラは，図3②に示したソフトウェアレベルのリソース管理を担います．すなわち，すでにベアメタルサーバには多くのアクセラレータが接続されている状態を仮定します．そのうえで，要求に応じて，データ処理パイプラインに対して必要なアクセラレータを割り当てます．ディスアグリゲーテッドコンピュータコントローラは，デファクトのコンテナオーケストレータであるKubernetesを拡張して実装されています．そして，Kubernetesのカスタムリソースを用いて，ワーカーノード（前述のベアメタルサーバ）上のアクセラレータやその間の接続を管理します．

従来，アクセラレータを活用したデータ処理パイプラインを実際に構築・運用する場合，高度な専門知識と多くの時間を必要としていました．これをディスアグリゲーテッドコンピュータコントローラが適切に隠蔽し，as a Serviceとして提供することを可能にしています．これにより，データ利用者はデータ解析の機能面のみに集中したまま，アクセラレータを活用したデータ処理パプラインのメリットを享受できるようになります．

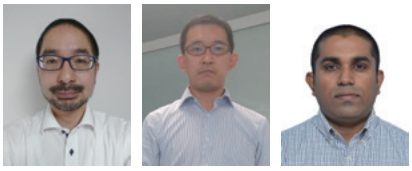
今後の展開

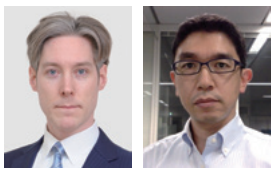
本稿では，DCIの概要と，ディスアグリゲーテッドコンピューティングを活用したコンセプト実証について説明しました．そして，CPSにおける映像解析を例として，アクセラレータを活用したデータ処理パイプラインの効果やas a Service化の有用性について示しました．今後は，DCIのもう1つの特徴であるAPNとの連携を進めつつ，各種オペレーション機能の拡充による実用化を推進します．また，映像を活用する他ユースケースや，映像処理以外のユースケースへの展開を図ります．

本成果は富士通株式会社との共同研究開発で研究中の技術を活用しています[6]．

■参考文献

(1) 水野，畠山，福田，松井，松田：“通信キャリアにおけるComposable disaggregated infrastructure，”電子情報通信学会ソサエティ大会，2022.
(2) https://www.computeexpresslink.org/_files/ugd/0c1418_a8713008916044ae9604405d10a7773b.pdf
(3) https://iowngf.org/wp-content/uploads/2023/04/IOWN-GF-RD-DCI_Functional_Architecture-2.0.pdf
(4) https://iowngf.org/wp-content/uploads/formidable/21/IOWN-GF-RD-DCIaaS_PoC_Reference_1.0.pdf
(5) https://journal.ntt.co.jp/article/20101
(6) https://pr.fujitsu.com/jp/news/2021/04/26.html





（上段左から）樽林 亮介／石崎 晃朗／Sampath Priyankara
（下段左から）Christoph Schumacher／水野 伸太郎

IOWNは，ネットワークだけでなく，コンピューティング基盤の変革をもたらす構想です．その構想の実現に向けて，IOWN GF等を通じて他の企業と協力・連携しつつ，積極的に取り組んでいます．

◆問い合わせ先
NTTソフトウェアイノベーションセンタ
システムソフトウェアプロジェクト
E-mail　scg-p @ ntt.com