IOWN Global Forumにおける次世代コンピューティング基盤の検討

インターネットやクラウドコンピューティングによってもたらされたさまざまな技術革新によって社会のあり方が大きく変わってきています．今後は，より一層のデータ活用とAI（人工知能）による高度なサービスが次々と登場することが期待されます．これらのトレンドを支えるためには，デナード則などのデータ処理能力の限界に対する抜本的な解決と，ESG/SDGsの要請によるエネルギー消費削減との両立が必要となります．このような背景の下，IOWN Global Forumでは，新しいコンピューティング基盤として，Data-Centric Infrastructure Functional Architectureを規定し，その初版となるドキュメントをリリースしました．本稿では，このドキュメントについて解説します．

ますたに　ひとし
益谷　仁士[†1]

シューマッハー　クリストフ[†2]

しみず　けんじ
清水　健司[†3]

NTTネットワークサービスシステム研究所[†1]
NTTソフトウェアイノベーションセンタ[†2]
NTT未来ねっと研究所[†3]

DCIがめざすコンピューティング基盤

IOWN Global Forum（IOWN GF）では，ネットワークやコンピューティングを含めた全体アーキテクチャが規定されています（**図1**）．このうちデータセントリックインフラストラクチャサブシステム（DCI）はオールフォトニクス・ネットワーク（APN）により実現される光ネットワーク上で構築される，コンピューティングとネットワークからなるサブシステムとして定義されており，さまざまなユースケースを実現するためのアプリケーションが動作する情報処理基盤となっています．DCIでは，分散配置された複数の異なるコンピューティングリソース（CPU，メモリ，FPGA，GPUなど）と，さまざまなQoSを提供するネットワークとを組み合わせて，アプリケーションを動作させるための環境を提供することを目標としています．DCIが想定する異種のコンピューティングリソースは，汎用的な計算処理に加え，AI（人工知能）処理専用などを包含し，これらのリソースは，後述するサービス事業者向けのAPIを通じてアプリケーションの動作環境として提供されるとしています．

現在のコンピューティングアーキテクチャとのギャップ

DCIのデザインゴールを設定するため，IOWN GFがめざすユースケースを実現するための要件に対して，既存技術とのギャップ分析を実施し，課題

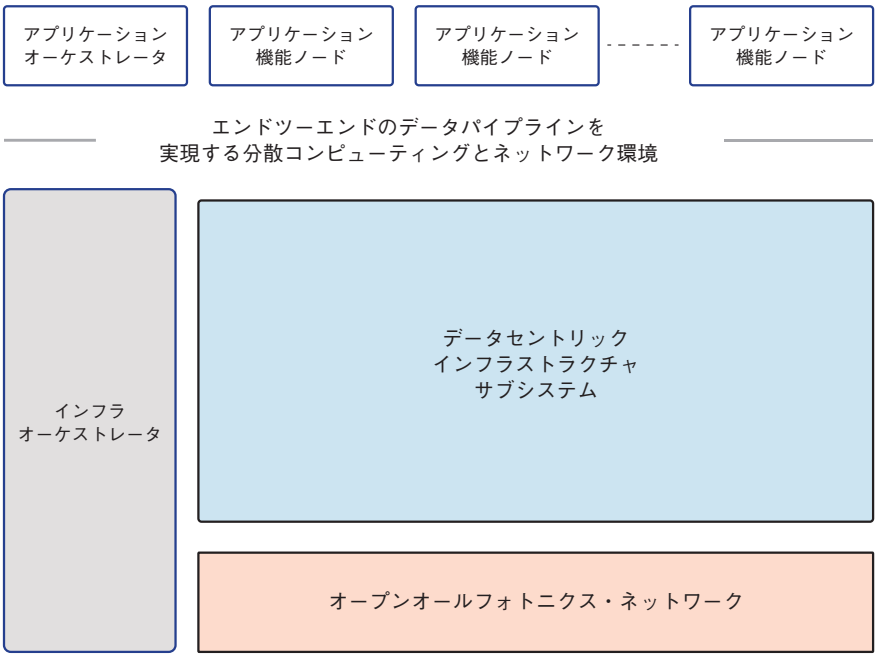


図1　IOWN Global Forumアーキテクチャ概要

が次のとおり記載されています．

■スケーラビリティに対する課題

さまざまなユースケースを実現する異なる要件を持つアプリケーションは，それぞれの情報処理に適したコンピューティングリソース，メモリ，入出力の要件があります．これらのアプリケーションを効率良く収容するためには，それら処理に適した各リソースを無駄なく割り当てていく必要がありますが，現在のサーバ筐体を中心とした従来型のコンピューティングでは異なるサーバごとのリソースをかき集めて効率的に結合することが難しいということが述べられています．

■パフォーマンスに対する課題

アプリケーションは多種多様に存在し，その中でもレイテンシやジッタに対する要件が厳しいアプリケーションがありますが，それらを既存のサーバ，ベストエフォート型のネットワークで実現する場合，それらの要件を満たすように設計されていないため，ユースケースを実現できないという問題があります．よって，このような要件の厳しいアプリケーションも最初から考慮したコンピューティングとネットワーキングに関するアーキテクチャ設計が求められるという課題が提起されています．

■エネルギー消費に関する課題

前述のスケーラビリティに関する課題に関連して，さまざまなリソースの割当て効率的にできなければ，無駄なリソースが使われないまま電力だけ消費されてしまいます．よって，さまざまなリソースの利用の効率化と併せて，IOWN GFで志向しているエネルギー消費の効率化を両立する新たなアーキテクチャ問題の検討が必要であることが述べられています．

デザインゴール

前述の既存アーキテクチャとのギャップ分析をふまえ，次のDCIのデザインゴールが規定されています．

① 端末からエッジクラウドに至るまでの環境においてスケーラビリティを考慮したアーキテクチャ
② CPU以外の異なるコンピューティングリソースが利用できるアーキテクチャ
③ 高速な光ネットワーク上でユースケースを構成するアプリケーション間のデータ転送を効率化するデータパイプラインの実現
④ GPUやFPGAなどの異なる種類のコンピューティングリソース間で不要なデータコピーを排除した効率的なデータ共有を可能とするアーキテクチャ
⑤ 高速・大容量に加え，帯域予約型，低レイテンシ，低ジッタといった異なるQoSを同時に提供可能なアーキテクチャ
⑥ IPと非IP間のデータ交換などのゲートウェイ機能を提供するアーキテクチャ

上記の6つのデザインゴールを元に，DCIアーキテクチャを設計する上で前提とする，DCI，APN，既存ネットワークとの関係が**図2**のように示されており，端末からDCIへのアクセスには2つの手段があることが分かります．つまり，APNにダイレクトに接続するパターンと，従来のネットワークスタックを有する端末もサポートすることを想定し，既存ネットワークを経由してDCIへアクセスするパターンの2つのパターンが記載されています．

DCIクラスタ

デザインゴールを元に，コンピューティングとネットワークリソースとを適切な単位でアプリケーションの実行環境としてユーザに提供するために，DCIアーキテクチャでは，Logical Service Node（LSN）が定義され

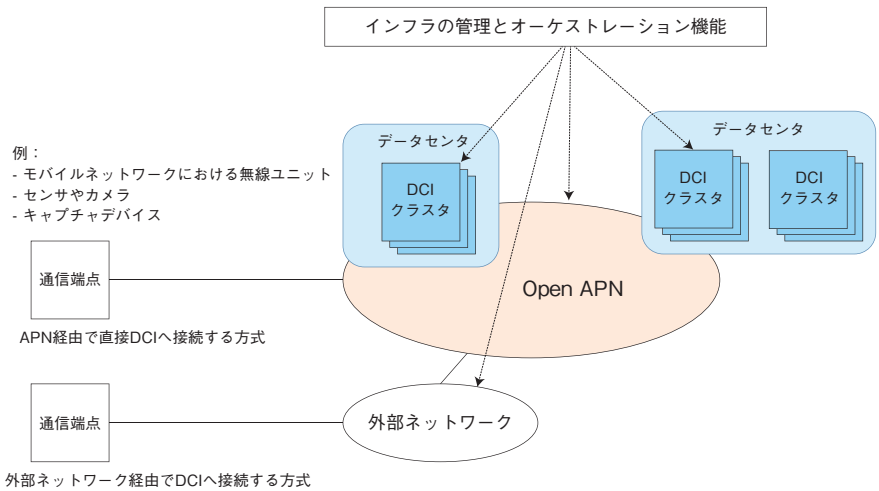


図2　DCIクラスタとOpenAPNの接続構成

ています．LSNではハードウェアレベルで論理分割されたリソースがユーザに提供されます．このLSNを提供するためには，複数のコンピューティングリソースとネットワークリソースから適切な組合せを選択して提供する必要があります．そこで，DCIアーキテクチャでは，DCIクラスタが定義されました．

■DCIクラスタの構成要素

構成要素は，DCIノード，ノード間インターコネクト，DCIゲートウェイからなります（**図3**）．

① DCIノード

・コンピューティングノードの基本単位を示します．このノードは従来のサーバに搭載されていたマザーボードに加え，FPGAやGPUといったさまざまな異種のコンピューティングリソースを提供するアクセラレータを複数搭載することを前提としたアーキテクチャとなっています．また，その異種コンピューティングリソース間でのデータのやり取りを実現するために，ノード内インターコネクトが定義されています．

・ノード内インターコネクト上では，異種コンピューティングリソース間で，同じデータを共有することになるため，そのデータの更新をそれらリソース間で同期をとる必要があります．そこで，従来のPCI expressバスに加え，将来的にはCompute Express Link（CXL）のようなキャッシュコヒーレンシ*1を考慮した新しい内部バス仕様も考慮し記載されています．

*1 キャッシュコヒーレンシ：メモリなどの共有リソースを複数のクライアントで読み書きする場合，各クライアントが持つデータの更新（キャッシュ）とその元のメモリ上のデータとの一貫性を保つことを指します．

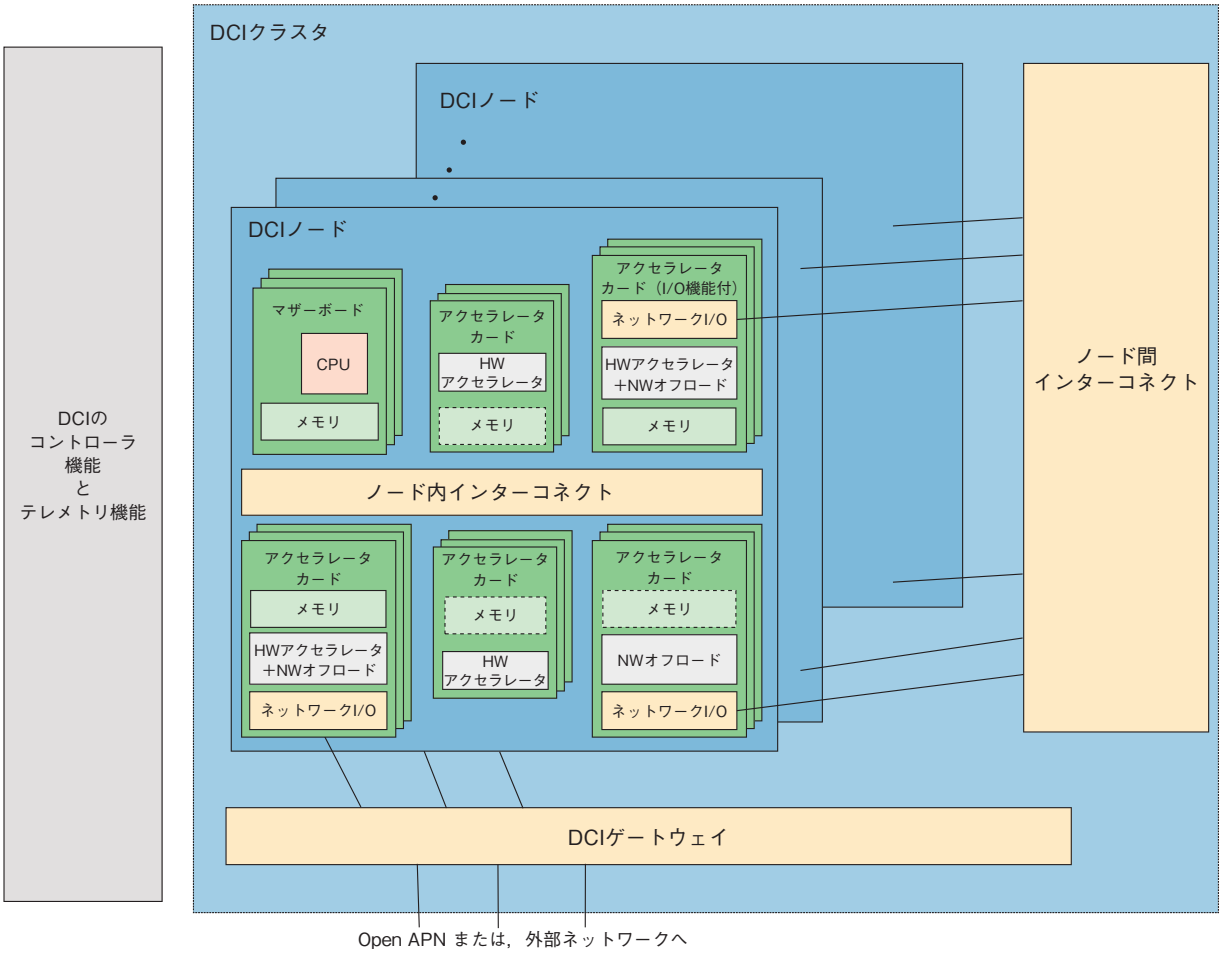


図3 DCIクラスタの構成例

② ノード間インターコネクト
・従来のTop Of Rack スイッチ相当に位置するネットワーク装置です．異なるQoSを持つネットワークを収容することを想定し記載されており，詳細な技術仕様は今後検討されることになっています．
③ DCIゲートウェイ
・APNとDCIクラスタを中継するネットワーク装置です．ノード間インターコネクト同様，異なるQoSに対応する必要があり，詳細な技術仕様は今後検討されることになっています．

■DCクラスタ向けコントローラ

上記のDCIクラスタの要素であるDCIノード，ノード間インターコネクト，DCIゲートウェイは，DCIクラスタコントローラに制御されます．DCIクラスタコントローラは，オーケストレータから依頼を受けて，必要な要件を満たすLSNの起動・停止などのライフサイクル制御を行います．DCIクラスタコントローラは，DCIクラスタの外に存在し，1つのDCIクラスタコントローラが制御するDCIクラスタの数に制約は設けられていません．

DCI Infrastructure as a Service（DCI IaaS）

DCIアーキテクチャで定義されたLSN，APNで定義される光ネットワーク，さらに今後議論される予定のFDN[*2]をまとめて，サービス事業者にIaaS（Infrastructure as a Service）のような形式でテナントとして提供されるDCI IaaSが定義されています（**図4**）．サービス事業者はテナントとして払い出されたLSNとその間を中継するネットワークの上に，サービス提供に必要なミドルウェア・アプリケーションをデプロイし，それらを介して最終的にエンドユーザにサービスを提供します．サービス事業者がDCIインフラを利用する際に必要となる，DCI IaaS向けのサービスAPIも定義されています．例えば，DCIアーキテクチャの主要コンポーネントであるLSNについて，生成，設定，起動，停止などを制御するAPIが定義されています．

IOWN GFユースケースを実現するデータプレーンの分析

さまざまなユースケースを分析すると，「データの流れ」と，「データを処理するコンポーネント」に整理することができます．IOWN GFにおいて先行ターゲットとするユースケースは，RIMのドキュメント[(1)]にまとめられており，サイバーフィジカルシステム（CPS）のユースケースでは，エリアマネジメント（AM）を実現するためのデータパイプラインが記載されています．例えば，各地域に設置された1000台のモニターカメラ映像をローカルアグリゲーションノードで一度集約し，さらに各地のローカルアグリゲーションノードから集約した膨大なリアルタイムデータを，地域エッジクラウドで常時AI分析し，モニターエリアにおける警告情報などを，即座に現地のセキュリティスタッフ等に通知するシナリオが定義されています（**図5**）．

各データフローは異なる転送性能の要件を持ちますが，異なるデータプレーンを通ることになるため，その分類を行ったうえで，データ転送の高速化が必要なデータプレーンを抽出する必要があります．例えば，DCIノード内であれば，ノード内インターコネクト，異なる拠点間のDCIノード間であれ

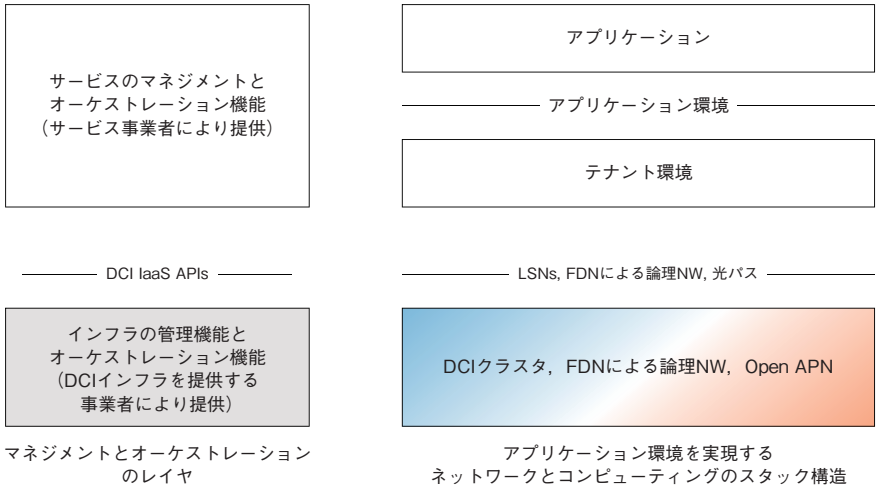


図4　DCI IaaS のサービスモデル

*2　FDN（Function Dedicated Network）：IOWN GFでコンセプトが提案されており，APNやその他の物理ネットワークで構成されたレイヤの上に構成される論理ネットワークという定義がドキュメントの中でされています．

CPSエリアマネジメントのユースケースにおけるデータパイプラインの例

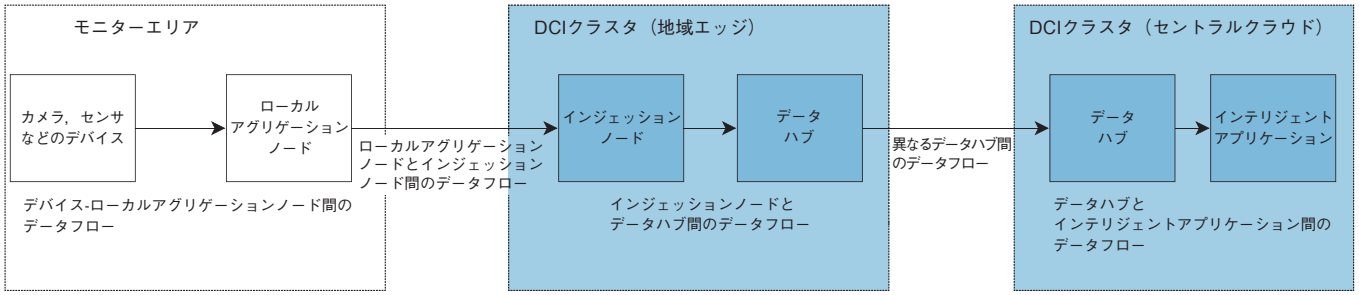


図 5　CPS AM データパイプライン

ば，DCIゲートウェイとAPNを経由したデータプレーンとなります．DCIドキュメントでは，データプレーンのパターンとして**図 6** のように分類されています．この分類を基に，図 5 の各データフローの性能要件を分析し，既存技術では実現が困難なデータプレーンのパターンに対し，次に示すデータプレーン高速化フレームワークが提案されています．

RDMAによるデータ転送高速化

広域分散されたコンピューティングリソースを接続するための，データプレーン高速化手法として，RDMA（Remote Direct Memory Access）を用いた長距離データ転送が提案されています．RDMAは，従来，データセンター内の数mから10 m程度の比較的短距離なHPCノード間の通信をターゲットに開発されてきましたが，遠隔地に離れたデータセンター内のリソースも，統一したリソースとしてシームレスに活用するためには，遠距離データ転送に適したRDMAが必要となります．

RDMAを活用する際は，データの完全性が保証され，高信頼なコネクションサービスに広く用いられるRC（Reliable Connection）トランスポートタイプが前提とされています．DCIドキュメントでは，RDMA RCを長距離に離れた拠点間に適用する際に生じるパフォーマンス劣化を解消するために，以下のような内容が提言としてまとめられています．

■キューの深さの最適化

RDMAにおいてデータの送受信には，送受信要求をWQE（Work Queue Element）のかたちで生成し，送信時であればWQEをSendキューへキューイングすることによって送信処理が開始されます．RDMA NIC（Network Interface Controller）が持つSendキューを深くすることにより，RCにおける送達確認（Ack，Acknowledgement）を待つことなく，一度に送り出すメッセージ数を増加させることができます．これにより，遠距離通信などのRTT（Round-Trip Time）が大きい通信でも，高いスループットを維持することができます．このように，RDMAを長距離通信へ適用する際は，次の式で導出されるように，RTTに応じたキューの深さの最適化が必要となると記載されています．

（RTT＊LineSpeed）/ MessageSize = Required QueueDepth

■RDMA用NICとさまざまなアクセラレータ間のデータ転送効率化

データパイプラインによっては，データコピーの回数を必要最小限にとどめ，RDMA用NICと，さまざまなアクセラレータ間で，メモリによるバッファを極力介さずにデータを直接転送する技術を利用可能とすることが提唱されています．

■長距離通信への適用時の信頼性

RDMA-RCでは，データロスが発生した際にも，再送によりデータの完全性を担保することができますが，特に長距離通信では再送が完了するまでにRTT分の時間を必要とするため，スループットの低下は避けられません．そこで，下位のネットワークレイヤであるAPNに対して，より品質の高い光パスを要求する機能や，より効率的な再送アルゴリズムを選択的に利用できる機能の提言が記載されています．

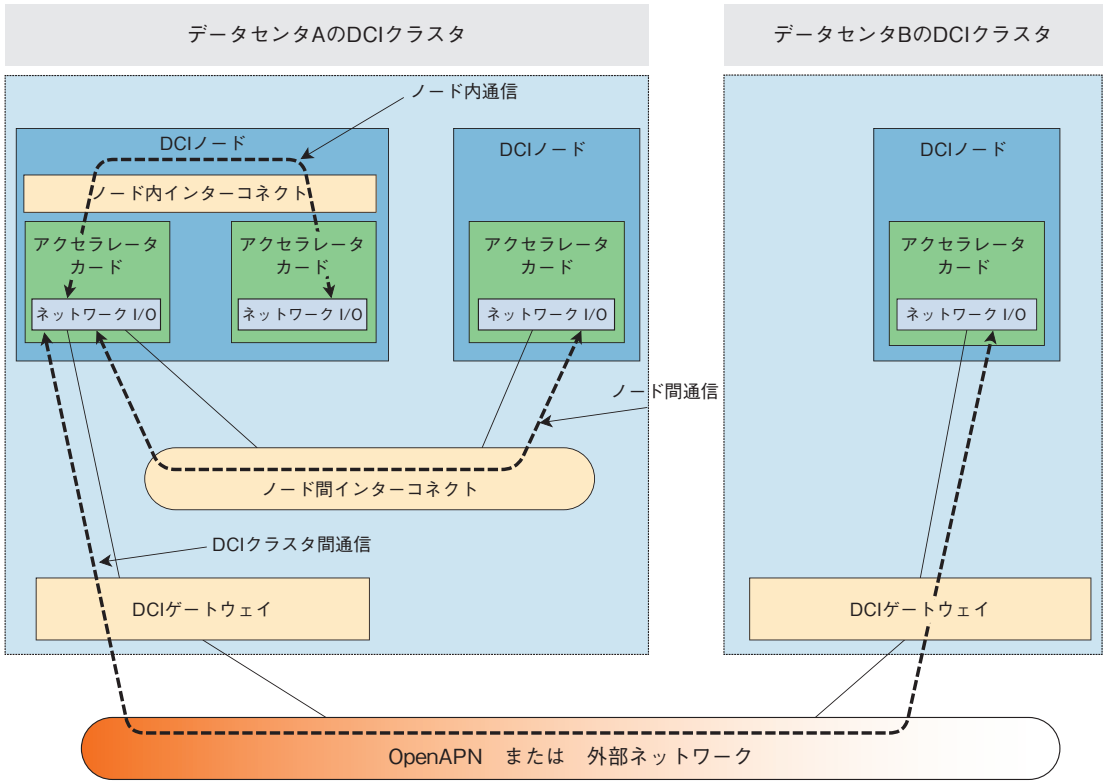


図6　3つのデータプレーンパターンの分類

■DCIのコントロール・マネジメントプレーンと連携したQoSの実現

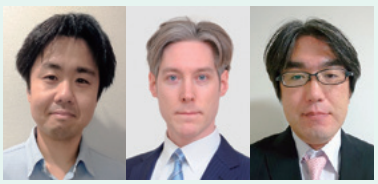
DCIクラスタ内においては，複数のデータフローが流れます．従来のクラウドコンピューティングの考えで，あらかじめリソースを確保せずに，ネットワーク装置とサーバでの輻輳制御アルゴリズムだけでは，十分なQoSを担保することが難しいことが記載されています．よって，異なるQoSクラスをあらかじめ確保したうえで，サービスを実現する考慮が必要であることが記載されています．

まとめ

DCIドキュメントでは，IOWN GFのユースケースを実現するためのデザインゴールを設定し，それを達成するためのコンピューティングアーキテクチャとしてDCIクラスタやそれを使ったDCI IaaSのサービスモデルを定義しています．また，ユースケース例の1つとしてCPS AMを分析し，DCIクラスタ間の長距離伝送において，長距離RDMAを用いたデータプレーン高速化について提言されています．今後は，実証等を通じて，その有効性と新たな課題を抽出・フィードバックし，さらなる技術革新を牽引することを想定しています．

■参考文献

(1) https://iowngf.org/wp-content/uploads/formidable/21/IOWN-GF-RD-RIM-for-AM-Use-Case-1.0.pdf



(左から) 益谷　仁士/
シューマッハー　クリストフ/
清水　健司

提案されている技術文書の有効性の確認には，技術評価の必要があり，また，その結果をフィードバックして，現在の技術文書をアップデートしていく必要があります．これらの活動は，さまざまなプレイヤーの参画が必要となりますので，ご興味のある方はぜひ活動にご参加いただきたいです．

◆問い合わせ先
NTTネットワークサービスシステム研究所
ネットワーク基盤技術研究プロジェクト
E-mail　iowngf-info＠ml.ntt.com