URL　https://journal.ntt.co.jp/article/38176
DOI　https://doi.org/10.60249/26025006

主役登場

高性能・省電力・柔軟なIOWNコンピューティング基盤の実現に向けて

史　旭　Kyoku Shi

NTTソフトウェアイノベーションセンタ
AI基盤プロジェクト　主任研究員

AI（人工知能）やIoT（Internet of Things）をはじめとする大量データ処理が日常化する現代では，計算性能の向上と同時に電力消費の増大が進み，データセンタにおける電力や用地の不足，性能要件の高度化に伴うコストパフォーマンスの低下，そして顧客ニーズに応じたシステム構成の変更やサービス開発・運用の迅速化など，さまざまな課題が顕在化しています．私自身，こうした課題に日々向き合う中で，データ収集・活用の加速と環境に配慮したカーボンニュートラルの両立の難しさを強く実感してきました．その解決に向け，試行錯誤を重ねながら，アクセラレータを中心とした省電力・低遅延・柔軟性を兼ね備えた新しいコンピューティング基盤「Data-Centric Infrastructure（DCI）」の実現をめざしています．

私は，2025年の大阪・関西万博における映像AI分析を題材とした「万博DCIプロジェクト」に技術面での中核メンバとして参画しました．その過程では，設計段階から想定していなかった制約やトラブルに幾度となく直面しましたが，その都度，構成の見直しや設定の調整，現地での検証を重ねることで，アクセラレータ活用の高度化・柔軟化による電力効率改善の効果を実証できました．そして現在，これらの成果を踏まえて，DCIの中核を担う「DCIコントローラソフト」の商用開発を進めており，実用化・事業化に向けた具体的な検討を行っています．

開発中のDCIコントローラソフトは，大きく2つの技術で構成されます．1つは，動的ハードウェアリソース制御技術（Dynamic Hardware Resource Control：DHRC）です．DHRCは，アクセラレータの有効活用やハードウェア構成・アプリケーション配備の最適化を目的として，柔軟なハードウェアリソース割当とテナント間でのリソース共有，通信設定管理機能を提供します．従来，テナントや拠点ごとに個別設計・固定運用されていた基盤に対し，DHRCでは各テナントの利用状況をリアルタイムに監視し，動的かつ最適なリソース割当を実現することで利用率を大幅に向上させます．また，運用要件に応じた高度なスケジューリングを可能とし，リソースの片寄せや冗長化の担保にも対応します．さらに，Commercial Off-The-Shelf/Composable Disaggregated Infrastructureなど異なるサーバアーキテクチャを抽象化し，All-Photonics Network等で接続された遠隔拠点間のリソース共有を実現することで，拠点をまたいだ効率的なリソース活用を可能とします．

DCIコントローラソフトのもう1つの構成技術は，アプリケーションフレームワーク（Application Framework：APFW）です．アクセラレータを有効に活用し，高効率かつ低消費電力なAIアプリケーションを開発するためには，AI処理に適した技術の組込みが不可欠となります．しかし，SmartNICを介したCPU非介在のデータ転送（例：Remote Direct Memory Access）や，データ処理中におけるデータコピーの回避などには，センシティブなメモリ操作が必要となるため，高度な専門知識を要します．また，AI推論に関する一連の処理（デコード，前処理，行列演算，後処理）をバッチ化し，GPUへフルオフロードするためには，GPU高効率化に関する高度なオフロード技術の知識やOSSの使いこなしスキルが求められることから，開発難易度は非常に高いものとなります．APFWでは，GPUやSmartNICといったアクセラレータを効果的に扱うための技術・OSSの設定やチューニング等のノウハウを機能モジュールとして実装し，拡張しやすいかたちで提供します．これにより，高度な専門知識がなくても高性能かつ省電力なアプリケーション開発を容易にし，開発期間の短縮と品質の安定化を実現します．また，DHRCと連携する機能も提供することで，APFW上のAIアプリケーションが高いスケーラビリティと可用性を備えた分散構成を容易に実現できます．

私は，DCIコントローラソフトの事業導入により，高性能・省電力なアプリケーション開発と運用の効率化を実現し，AIサービス提供者におけるコストパフォーマンス向上を現場で実感してもらえる技術にしていきます．さらに，都心部データセンタで深刻化する用地・電力不足といった社会課題に対して，DCIとAll-Photonics Networkを組み合わせ，Location-freeなAI基盤の実現を通じて，技術者として貢献します．

今後は，さらなるビジネス拡大をめざし，大規模言語モデルなどを含むマルチAIエージェント分野における高性能化・省電力化，そして新たなアクセラレータへの適用による機能拡充を進めていく予定です．私は，DCIをIOWN時代におけるデータドリブン社会を支える中核として，持続可能で高性能・省電力かつ柔軟なIOWNコンピューティング基盤に発展させていきたいと考えています．