③ ABCI：世界初の大規模オープンAIインフラストラクチャ

小川宏高 | 国立研究開発法人 産業技術総合研究所

ABCIの概要

国立研究開発法人 産業技術総合研究所（以降，産総研という）は，経済産業省「人工知能に関するグローバル研究拠点整備事業」（平成28年度二次補正）の一環として，アルゴリズム（Algorithm），ビッグデータ（Big Data），計算能力（Computing Power）の協調による，高度な人工知能処理を可能にする大規模で省電力のクラウド型計算システム，「AI橋渡しクラウド（AI Bridging Cloud Infrastructure）」（以降，ABCIという）[1]を整備することとし，2017年9月に一般競争入札による調達を行い，富士通（株）（以降，富士通）の技術の採用を決定した．以降，富士通および関連各社の協力の下，東京大学柏IIキャンパス内の産総研柏センターに新たに建設された「AIデータセンター棟」[2]への搬入・設置・システム構築・運用に向けた準備等を進め，2018年8月より運用を開始した．

ABCIは，我が国の人工知能技術開発のためのオープンなリーディングインフラストラクチャとして，人工知能分野の最重要課題への挑戦，特に画像認識，音声認識，自然言語処理，種々の機械学習アルゴリズムやデータモデルの高度化，自動車／ロボットの自動運転／制御，創薬向け化合物推定，音声対話，自動翻訳等，幅広い分野での新たなアプリケーションの創出や，これらを支えるクラウド型計算システムの設計・運用ノウハウの民間への技術移転等，人工知能技術の社会実装の推進を目的としている．

ABCIのハードウェア上の一番の特徴は，「NVIDIA Tesla V100 SXM2」というGraphics Processing Unit（GPU）を4,352基搭載した国内最大規模のシステム，ということである．グラフィクス処理用プロセッサとして開発されたGPUは，ピクセル値など大量のデータ列・パラメータに対する積和演算などに特化した設計をとっている．数1,000基の演算コアと広帯域メモリをパッケージ化することで，Single Instruction Multiple Data（SIMD）型の処理の性能を追求する一方，仮想メモリの実現や複雑な分岐処理などの性能は犠牲にしている．これらの工夫により，大量のデータに対して同一の演算を行う並列性の高い処理においては，同世代のCPUの5～10倍の性能を示し，電力あたりの性能も高い．こうしたGPUの性質は，HPCアプリケーションのみならず，ABCIがターゲットとするディープラーニングの学習・推論にもよく適合する．

GPUベースのスパコンはすでにTOP500 Listの25%を占めるまで一般的となっている．特に2016年にチップ間インターコネクトNVLinkが登場して以降，NVLinkの高帯域幅（160～300GB/sec）を活用するため，マルチGPU構成によるファットノードをベースとしたシステムが増加している．米国オークリッジ国立研究所のSummitは，そうしたスパコンシステムの世界最大のインスタンスであり，ABCIの6倍以上のキャパシティを有する．国内の事例

では，産総研が ABCI のプロトタイプとして構築した AIST AI Cloud（AAIC）のほか，東工大 TSUBAME3，東大 Reedbush-H/L，九大 ITO サブシステム B，筑波大 Cygnus（PCIe 版を利用）等がある．また，民間でもさくらインターネットの事例などがある．

ABCI の実効性能・省電力性能は各種ベンチマーク結果が示す通りである．2019 年 7 月公表の MLPerf Training v0.6[3]（ディープラーニングの標準ベンチマーク）の Image Classification で，富士通研究所が ABCI を用いて 70 秒という世界最速記録を打ち立てた．2019 年 6 月公表の TOP500 List では 19.9 PFLOPS で世界 8 位かつ国内 1 位となった．同じく Green500 List でも 14.423GFLOPS/W で世界 3 位かつ国内 1 位となった．HPCG でも 0.509PFLOPS を記録し，世界 5 位に位置している．

図 -1　ABCI システム外観

ABCI システム

ABCI の外観を図 -1 に示す．ABCI は，AI データセンター棟が提供する電源，冷却システム，48U ラックを用いて導入された．

ABCI システムは，高性能計算システム，複数のストレージシステム，各種ネットワーク，それらの補助的機器から構成されるハードウェアと，システムを最大限活用するためのソフトウェア群からなる．ハードウェア構成の概要を図 -2 に示す．

以下では，主要な構成要素である高性能計算システムの計算ノードとインターコネクトにフォーカスして説明する．

計算ノード

ABCI の計算ノードは 1,088 台あり，そのすべての構成が同一である．FUJITSU Server PRIMERGY CX400 M4 をベースとした 2U シャーシに，PRIMERGY CX2570 をベースとした計算ノードを 2 台ずつ搭載した構成を基本とし，17 シャーシ 34 ノード，スイッチ 8 基，Cooling Distribution Unit（CDU）を 48U ラックに搭載する．ラックの重量は 1.1 トンを超え，消費電力は 67.33kW に及ぶ超高密度パッケージングを実現している．システム全体では 32 台のラックから構成される．

計算ノードの主要な構成は以下の通り．ブロック図は図 -3 に示す．

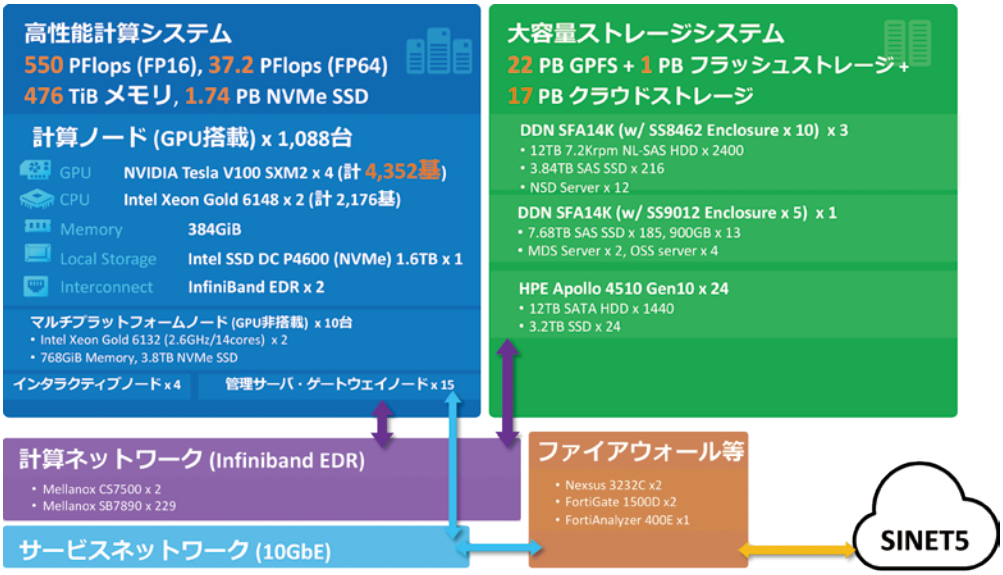

図 -2　ABCI ハードウェア構成

- CPU : Intel Xeon Gold 6148 Processor（Skylake-EP, 27.5MB Cache, 2.40GHz, 20 cores, 1.536TF@FP64）× 2
- メインメモリ : DDR4 2666MHz RDIMM（ECC）384GiB（32GiB x12），メモリバンド幅 : 128GB/s × 2
- GPU : NVIDIA Tesla V100 SXM2（5120 CUDA cores, 16GiB HBM2, 900GB/s, 7.8TF@FP64, 15.7TF@FP32, 125TF@FP16）× 4
- ローカル SSD : Intel SSD DC P4600 1.6TB u.2 × 1
- インターコネクト : InfiniBand EDR（100Gbps）× 2

全体では，Intel Xeon Gold 6148 Processor が 2,176 基，NVIDIA Tesla V100 SXM が 4,352 基，メモリ 476TiB，メモリバンド幅 4.19 PB/s，NVMe SSD 1.74PB となる．また，理論ピーク性能では 37.2 PFLOPS（FP64），75PFLOPS（FP32），550PFLOPS（FP16）に相当する．

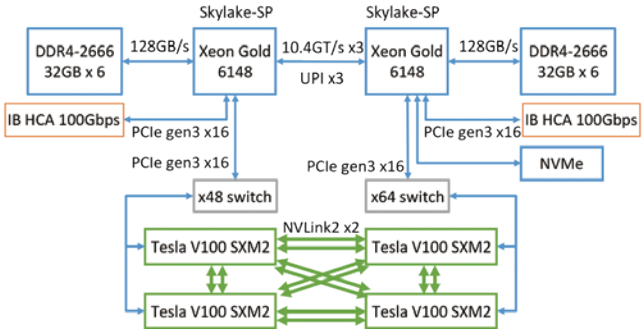

図 -3　ABCI 計算ノード構成

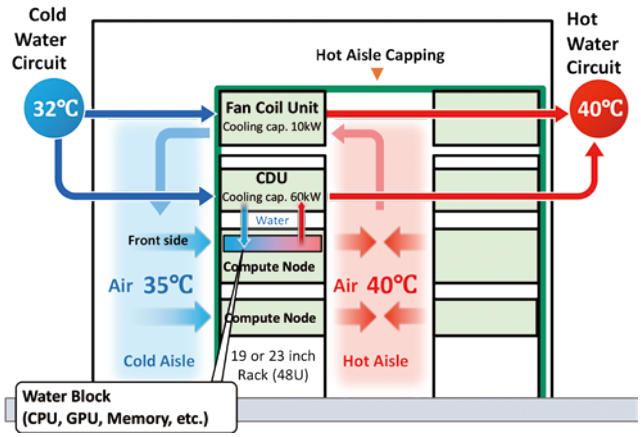

図 -4　AI データセンター棟の冷却システム

計算ノードの冷却

AI データセンター棟[4] では，冷却塔のみを用いて冷水を製造するフリークーリングシステムを採用しており，年間を通じて 32℃以下に冷却した 1 次冷却水をラックに供給する（図 -4）．32℃と比較的高温に設定しているのは，データセンターのある柏地区の最大湿球温度がおおむね 27℃であり，高温多湿の最悪ケースを想定しても 32℃の冷却水であれば常時安定的に製造できるからである．

この冷却水は CDU を介して 2 次冷却水としてラック内の計算ノードに分配される．計算ノードでは，高温になる CPU，GPU，メモリ等の基幹部品に取り付けられたコールドプレートを介して冷却する（図 -5）．

冷却しきれなかった熱はサーバ内のファンによりホットアイル（サーバ後方・排気側）に空気として排出される．ホットアイルの空気は，ラック上部に設置された簡易型空調機，Fan Coil Unit（FCU）によって冷却され，コールドアイル（サーバ前方・吸気側）に排出される．このとき FCU は CDU と同様に 32℃の冷却水を用いるため，コールドアイルの温度は 35℃となる．通常のデータセンターと比較すると高温だが，ほとんどの IT 機器は 10 ～ 35℃を稼働条件としており，問題とはならない．

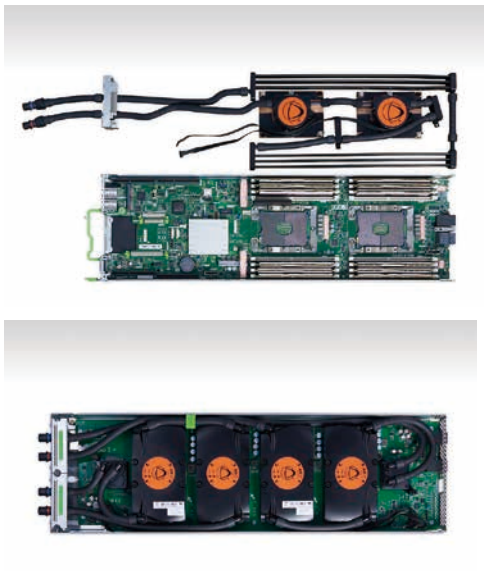

図 -5　ABCI 計算ノード

計算ノードのインターコネクト

ABCI では，図 -6 に示すように，ラック内のノードは，フルバイセクションバンド幅（任意の半数のノードが同時に残りの半数のノードに同時に通信しても競合が起きない構成）で相互に接続する．その一方，ラック間は，ノードの総インジェクションバンド幅 6,800Gbps（＝100Gbps × 2 ポート × 34 ノード）の約 1/3，2,400Gbps 相当でコアスイッチに結合する構成をとる．ABCI の規模であっても，コアスイッチは 2 基のみで済む．

こうした構成をとったのは，ABCI がターゲットとしているディープラーニングタスクの大半がノードまたは GPU 単体で実行されるためである．また，並列分散で実行されるタスクも Ring-AllReduce アルゴリズム[5] で通信するため，バイセクションバンド幅を使い切れない．また，全体的にインターコネクトが低コストで済み，システムの最小構成である「ラック」を増加させても主として配線コストしか増えないという利点もある．

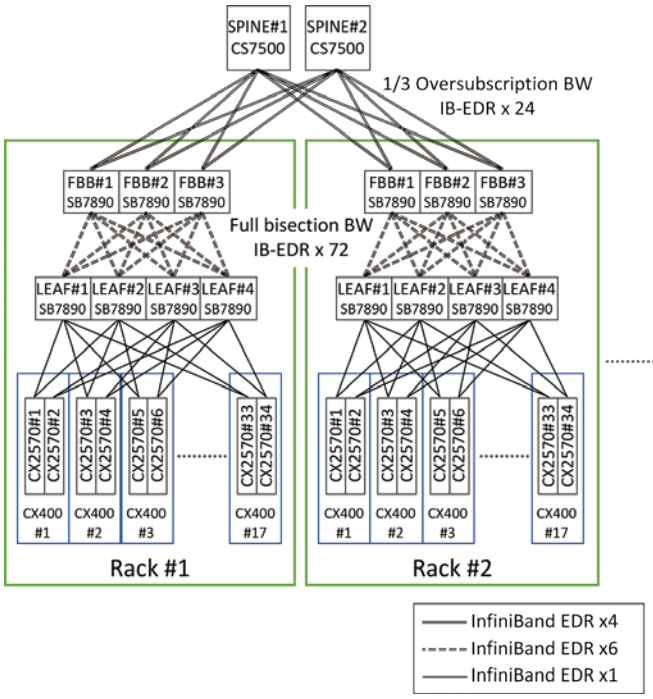

図 -6　ラック内・外インターコネクト

ABCI 利用を通じた研究成果創出に向けて

2018 年 8 月の運用開始から 1 年を経て，ABCI は 100 名を超えるプロジェクト，1,000 名を超えるユーザに利用されており，MLPerf Training のようなベンチマークのみならず，顕著な研究成果も生まれてきている．特に近年は，画像認識分野における ImageNet のような大規模な学習用データセットに対してディープラーニングを適用して汎用学習済みモデルを構築し，それをベースに転移学習によって個別課題のための学習済みモデルを構築して利用されるようになってきている．動画認識，自然言語理解の分野では，その有効性が示されつつある．このような大規模な学習用データを用いた学習も ABCI を利用することで可能となってきた．

ABCI では，さらなる研究成果の創出に向け，莫大な演算能力によりはじめて可能になる人工知能分野の最重要課題への挑戦を支援するため，グランドチャレンジプログラムを実施している．本プログラムは，ABCI の全 1,088 ノード（4,352GPU）を最大 24 時間，1 研究グループでの占有利用を可能にする公募型プログラムである．各年度とも 3 回の実施を予定しており，各回 2 課題程度を採択する予定としている．

参考文献

1) AI Bridging Cloud Infrastructure, https://abci.ai/
2) 小川宏高，松岡　聡，佐藤　仁，高野了成，滝澤真一朗，谷村勇輔，三浦信一，関口智嗣：世界最大規模のオープン AI インフラストラクチャ AI 橋渡しクラウド（ABCI）の概要，情報処理学会研究会報告（2018）．
3) https://mlperf.org/training-results-0-6/
4) 高野了成，三浦信一，杉田　正，小川宏高，松岡　聡：0.55 AI-EFLOPS の計算インフラストラクチャを支える超グリーン AI データセンタ，情報処理学会研究会報告（2018）．
5) Jeaugey, S. : Optimized Inter-GPU Collective Operations with NCCL, GTC (2017).

（2019 年 8 月 7 日受付）

小川宏高（正会員）　h-ogawa@aist.go.jp
1998 年東京大学大学院工学系研究科博士課程中退．博士（理学）．東京工業大学情報理工学研究科助手等を経て，現在，産業技術総合研究所 実社会ビッグデータ活用オープンイノベーションラボラトリラボ長 兼 人工知能研究センター人工知能クラウド研究チーム長．人工知能クラウド，ハイパフォーマンスコンピューティング等の研究に従事．ACM 等の会員．