Vol.2017-HPC-160 No.28
2017/7/28

AI 橋渡しクラウド — AI Bridging Cloud Infrastructure (ABCI) — の構想

小川 宏高[†1] 松岡 聡[†1†2] 佐藤 仁[†1] 高野 了成[†1]
滝澤 真一朗[†1] 谷村 勇輔[†1] 三浦 信一[†2] 関口 智嗣[†1]

概要：国立研究開発法人産業技術総合研究所（以降，産総研）では，平成 28 年度二次補正「人工知能に関するグローバル研究拠点整備事業」の一環として，平成 29 年度末に，東京大学柏 II キャンパスに，「AI 橋渡しクラウド（AI Bridging Cloud Infrastructure)」（以降，ABCI という）の導入を計画している．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の実現を目指し，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かつ省電力なクラウド基盤である．本稿では，ABCI のサーバシステムにフォーカスしつつ，ABCI の概要と，システム設計上の論点と我々が採った方策について紹介する．

AI Bridging Cloud Infrastructure (ABCI): Concept

HIROTAKA OGAWA[†1] SATOSHI MATSUOKA[†1†2] HITOSHI SATOH[†1]
RYOSEI TAKANO[†1] SHINICHIRO TAKIZAWA[†1] YUSUKE TANIMURA[†1]
SHINICHI MIURA[†2] SATOSHI SEKIGUCHI[†1]

1. はじめに

国立研究開発法人産業技術総合研究所（以降，産総研）では，平成 28 年度二次補正「人工知能に関するグローバル研究拠点整備事業」の一環として，平成 29 年度末に，東京大学柏 II キャンパスに，「AI 橋渡しクラウド（AI Bridging Cloud Infrastructure)」（以降，ABCI という）の導入を計画している[a)]．

ABCI は，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かつ省電力なクラウド基盤である．本システムは，世界最高水準の機械学習処理能力，高性能計算能力，及び省電力性を備え，画像，音声，テキスト等の超大規模なデータセットを対象とした，ディープラーニングを含む高度な機械学習処理及びシミュレーション等を，超省電力で超高速に処理する必要がある．それればかりではなく，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，画像認識，音声認識，自然言語処理等，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基盤の設計・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を強力に支援することが期待される．ゆえに，著者らはこれらの理念を如何に実現するか，人工知能処理のリーディングインフラストラクチャの「ロールモデル」は如何にあるべきかを念頭において設計を進めてきた．

ABCI には既存のスパコン調達にはないユニークな点が数多くあるが，代表的なものを以下に挙げる．

- 補正予算成立から約一年半という超短期プロジェクトであること．東京大学柏 II キャンパス内にスクラッチから，サーバ棟，設備機器置場，外構の設計・施工，給電設備，冷却設備，サーバラック等を含む付帯設備の整備，サーバシステムの調達・製造・納品までを完了する必要がある．特にサーバシステムの調達では，産総研 AI クラウド（AAIC）[1][2]及び今年 8 月導入予定の東工大 TSUBAME3.0[1][3]の成果を取り込むことで基本設計フェーズの加速を図るとともに，最先端のコモディティハードウェアを効果的に取り込み，インテグレートする技量が問われる．
- HPC 向けではなく，人工知能処理のための，「はじめてのスパコン」であること[b)]．すなわち，人工知能処理においてシステムの絶対性能及びキャパシティを定めるメトリックや，システムの評価基準となるベンチマークセットを我々自身が新たに定義し，世に問うていく必要がある．

†1 産業技術総合研究所
National Institute of Advanced Industrial, Science and Technology
†2 東京工業大学
Tokyo Institute of Technology

a) サーバシステムの調達における呼称は，「人工知能処理向け大規模・省電力クラウド基盤」としている．

b) 人工知能技術開発にフォーカスした大規模システムとしては，産総研 AIRC が平成 28 年度末に導入した AAIC，及び理研 AIP が同時期に導入した RAIDEN が知られているが，これらは NVIDIA 社の最新 GPU を含むサーバを多数，迅速に整備することに主眼を置いており，本格的な大規模システム（いわゆるスパコン）ではない．

- ディープラーニングを含む高度な機械学習処理に代表される人工知能分野でのワークロードにおいて，高性能かつ費用対効果の高いシステム設計が未知であること．ベンチマークセットの定義と並行して，state-of-the-art なシステムである産総研 AAIC 及び東工大 TSUBAME3.0 を用いて基礎的なデータの取得を迅速に行い，その知見からシステム設計及び改善を行う必要がある．

本稿では，ABCI のサーバシステムにフォーカスしつつ，ABCI の概要と，システム設計上の論点と我々が採った方策について紹介する．

2. ABCI の概要

冒頭に述べた ABCI の導入の目的を達成するため，ABCI では実現されるべき機能・性能について，概念要件，技術要件を以下の通り定めている．

2.1 概念要件

(1) AI Infrastructure: 人工知能技術を支える機械学習の超高速処理

- ディープラーニングを含む超高速な機械学習処理を実現する 100 ペタ AI-FLOPS 超級の演算性能
- ディープラーニングの予測結果に基づく高度なシミュレーション解析や，高精度演算を必要とする機械学習アルゴリズム等，ビッグデータ処理と高性能計算の融合を可能にするマルチ PFLOPS 級の倍精度浮動小数点演算性能
- 上記を支えるペタバイト毎秒級の超高速な I/O，ペタビット毎秒級の超広帯域・超低遅延なネットワーク

(2) Bridging Infrastructure: 民間への技術移転のためのオープンプラットフォーム

- 機械学習の対象となるマルチペタバイト級のビッグデータを収集・蓄積・共用可能なストレージ
- 汎用製品により構成されたコストパフォーマンスが良く模倣しやすいアーキテクチャ
- 広範囲のオープンソースソフトウェア，商用アプリケーションが動作可能なソフトウェアエコシステムのサポート

(3) Cloud Infrastructure: TCO に優れた最新鋭のクラウド基盤・運用

- 資源のパーティショニングやプロビジョニング，動的な計算環境のデプロイメント等によるマルチテナントのサポート
- 自動的な障害回復等，少人数で運用可能なクラウド運用管理
- 温液冷却や高効率給電系を含む次世代省電力設計

2.2 技術要件の概要[c)]

ABCI のシステムは，高性能計算システム，大容量ストレージシステム，各種ネットワーク等から構成されるハードウェア（図 1）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．

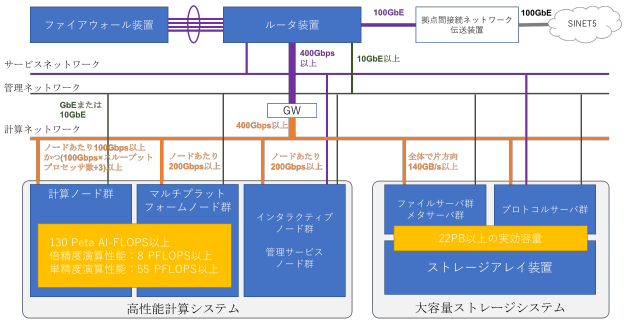

図 1　ABCI ハードウェア構成
Figure 1　ABCI System Architecture.

以下にシステムの機能及び性能に関する技術要件の概要を示す．

- 高性能計算システムの合算理論ピーク演算性能は 130 ペタ AI-FLOPS（3 節参照）以上であること．また，倍精度，単精度浮動小数点演算での合算理論ピーク性能はそれぞれ 8PFLOPS 以上，55PFLOPS 以上であること．
- 高性能計算システムのメモリの合算容量は 435TiB 以上，かつ合算理論ピークバンド幅は 3.8PB/sec 以上であること．
- 大容量ストレージシステムは，全体で 22PB 以上の実効容量を備え，高速かつ高信頼な並列ファイルシステムを提供すること．高性能計算システムのすべての計算ノードから利用できること．
- 計算ネットワークは，高性能計算システム及び大容量ストレージシステムを相互に接続すること．理論転送バンド幅 100Gbps 以上のネットワークを用いて，なるべく高いバイセクションバンド幅を有する構成をとること．なお，計算ノード間，及び計算ノード・大容量ストレージシステム間の理論転送バンド幅は 100Gbps 以上とすること．
- サービスネットワークは，高性能計算システム及び大容量ストレージシステムの外部アクセスを必要とする機器群を接続するとともに，計算ネットワーク，管理ネットワーク，及び SINET5 に接続すること．
- 産総研が準備する給電設備及び冷却設備を最大限活用した効率の良いシステムを構築すること．ただし，システムの有効総消費電力は 3,000kW 以下とすること．

c) 本稿で示すスペックは原稿執筆時点のものであり，入札公告時には変更になった可能性がある．

- システムを最大限活用し，クラウド運用，人工知能処理の高速化・高度化，ビッグデータの活用を実現するソフトウェア群を有すること．

3. AI-FLOPS

Caffe[4]，CNTK，TensorFlow，Chainer を始めとする多くのディープラーニング向けフレームワークでは，一般的に学習フェーズに単精度浮動小数点数（FP32）を用い，その演算カーネルはHPC分野でも一般的に用いられるSGEMMに相当する．

一方，チップメーカーはディープラーニングの高速化に資するべく，また自社製品のマーケットバリューを高めるべく，しのぎを削っている．NVIDIA 社は Pascal 世代で FP16 精度の演算コアを，Volta 世代で FP16/FP32 混合精度の演算コア（Tensor Core）を追加することで，より低精度の演算スループットを訴求してきた．Intel 社は Knights Mill（KNM）で，4 個の単精度 FMA 演算を 1 命令にパッキングする QFMA の導入により従来精度の演算の高速化を図る一方，INT16/INT32 の混合精度の積和演算 VNNI，4 個の VNNI 演算を 1 命令にパッキングする QVNNI などを追加することで低精度の演算スループットの改善も図ってきている．また，Google も TPU1 は FP16 演算にフォーカスしており，Google I/O 2017 で発表された TPU2 については詳細が明らかになっていないものの，おそらくは FP16 演算と推測される[5].

このような状況下で，ABCI のような人工知能向けの大規模システムで絶対性能を定めるには，メトリックをどう定めるのが適切であろうか．FP32 に固定すると，チップメーカーの state-of-the-art なエフォートを考慮しないことになり，旧製品でよいという意味になる．ましてや FP16 にすると，特定メーカーを利することになり公平性に欠き，演算精度や収束性能は二の次であるというメッセージになる．

我々は，ベンチマークを規定し，それに基づいて理論 AI-FLOPS 値，実効 AI-FLOPS 値，実行効率，検証指標，評価値という複数のメトリックでシステムを評価することで，この課題を解決する方法を提案する．

3.1 理論 AI-FLOPS 値

「参照精度」を FP32（または FP64）とする．「設定精度」は AI-FLOPS 値の算出根拠となる精度とし，参照精度と同一精度，参照精度を縮退または拡張した精度，混合精度のいずれであってもよい，すなわち state-of-the-art なプロセッサないしスループットプロセッサの計算精度に適合するように選んでよい．ただし，後述のベンチマークの検証指標が pass/fail criteria をパスするように選ばなくてはならないとする．

このとき，**理論 AI-FLOPS 値は，各プロセッサが設定精度で理論的に 1 秒間に実行できる演算の回数を指すものと定義**する．端的に言えば，理論 AI-FLOPS は，ベンダーの自己申告による理論的なピーク演算性能である．したがって，例えば NVIDIA の Tensor Core，あるいは Intel の QVNNI 命令のそれぞれ理論ピーク性能としてよい．

3.2 ベンチマーク

入力行列A（$M\times K$行列），B（$K\times N$行列），出力行列C（$M\times N$行列）に対する GEMM 計算を行い，性能を測定することを考える．簡単のため，$\alpha = 1, \beta = 0, M = N = K$とする．

$$C = \alpha AB + \beta C$$

行列のサイズは，参照精度での計算に必要なメモリサイズがプロセッサのオンチップメモリの 2 倍弱以上となるように選ぶ[d]．具体的にはオンチップメモリ量Memバイト，参照精度の表現バイト数bとするとき，行列のサイズは以下の条件を満たす値を選ぶとする．

$$M = N = K \geq sqrt\left(\frac{Mem\times\kappa}{b}\times\frac{1}{3}\right), where\ \kappa = 1.75$$

このときベンチマークは以下の手順の計算を行う．まず，参照精度に従う入力行列A^{ref}, B^{ref}を生成し，これらを用いて GEMM を計算し，参照精度の出力行列C^{ref}を得る．次に設定精度に従う入力行列A, Bを生成し，これらを用いて GEMM を計算し，出力行列Cを得る．この計算に要した時間をt秒とすると，実効 AI-FLOPS 値は以下のように算出できる．

$$\text{Effective AI-FLOPS} = MN(2K-1)/t$$

また，実行効率 efficiency は設定精度での理論 AI-FLOPS 値をPeak AI-FLOPSとすると，以下で求められる．

$$\text{efficiency} = \frac{\text{Effective AI-FLOPS}}{\text{Peak AI-FLOPS}}$$

参照精度と設定精度の計算結果を比較し，L2 ノルムを検証指標（validation metric）とする．

$$\text{validation metric} = \sqrt{\frac{\sum_{i=0,j=0}^{i<M,j<N}\left(C_{i,j}-C_{i,j}^{ref}\right)^2}{\sum_{i=0,j=0}^{i<M,j<N}\left(C_{i,j}^{ref}\right)^2}}$$

このとき，設定精度は下記の pass/fail criteria を満たさなければならないとする．3.1 において，あらかじめこの条件を満たせない精度を設定精度に選んではならないということである．

$$\text{validation metric} \leq [\text{pass/fail criteria}] = 0.1$$

実効 AI-FLOPS 値と検証指標から下記の評価値を得る．

$$\text{score} = (\text{Effective AI-FLOPS})\times e^{(-1\times validation\ metric\times\kappa)},$$
$$where\ \kappa = 75$$

d) 設定精度が参照精度の約 1/2 の表現バイト数と仮定したときに，設定精度での計算にオンチップメモリ容量の大半を使うようにするとともに，参照精度での計算はオンチップメモリ上だけでは実行できないようにする，という実用上の理由である．

この評価式の詳細な導出過程，根拠は本稿では省略するが，意図するところは，プロセッサの処理性能を CNN の処理スループットとその計算の質である精度の積として，モデル化することである（図 2）．この評価式では validation metric が大きく（悪く）なるにつれ，最初はなだらかに途中から急激に悪化する．ResNet[6]を用いて行った予備的実験では validation metric が一定以上悪化すると精度も急激に悪化するという傾向があり（validation metric が 0.001 では精度にほとんど影響がないが 0.01 では 1%程度悪化する），1%の精度向上に約 2 倍の計算量が必要であることからκを導出している．

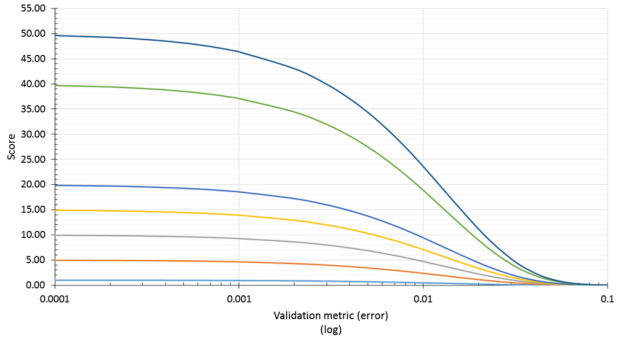

図 2　score と validation metric の関係

Figure 2　Relationship between score and validation metric.

4. ABCI のシステム設計

2.2 で述べた性能指標あるいはシステムキャパシティを実現する上で根幹となるのは，スループットプロセッサによる計算スループットの積み上げと，ノード間インターコネクト及びローカルストレージを含む I/O の設計である．

ここで前述の AAIC，TSUBAME3.0 に加え，Oakforest-PACS について簡単な諸元を表 1 に示す．

表 1　Specification 比較

Table 1　Comparing Specifications.

System	AAIC	TSUBAME 3.0	Oakforest-PACS	ABCI
# nodes	50	540	8208	未定
Throughput Proc	NVIDIA Tesla P100 SXM2	NVIDIA Tesla P100 SXM2	Intel Xeon Phi7250	未定
Inter T-Proc Connection	NVLink 160GB/s	NVLink 160GB/s	NA	未定
# T-Proc / node	8	4	1	未定
Network Interface / node	InfiniBand EDR 100Gbps x1	Omni-Path 100Gbps x4	Omni-Path 100Gbps x1	4.4 参照
Injection BW / T-Proc	12.5Gbps	100Gbps	100Gbps	4.4 参照
Interconnect Topology	Full bisection FatTree	Full bisection FatTree	Full bisection FatTree	4.4 参照
Local Storage	Intel DC S3510 480GB SATA SSD sr: 500MB/s, sw: 440MB/s	Intel DC P3500 2TB NVMe SSD sr: 2.7GB/s, sw: 1.8GB/s	なし 共有 BB あり	4.4 参照

AAIC の設計において我々は，主に単体ノードでディープラーニングを含めた機械学習処理を行うサービスをクラウド的にユーザに提供することを念頭においた．このため，Tesla P100（NVLink 版）を搭載した NVIDIA DGX-1 や Facebook Big Basin とほぼ同様のノード設計を行い，予算の許す限り GPU に投資するという方法を採った．結果として，ノードのインターコネクトは InfiniBand EDR 1 本とし，GPU ノード 50 台，GPU 非搭載ノード 68 台を単一のディレクタースイッチに繋ぐという大変分かりやすいが，しかしプアな構成となった．一方で，TSUBAME3.0 や Oakforest-PACS は，スループットプロセッサあたり 100Gbps のインジェクションバンド幅を有し，それらをフルバイセクションバンド幅で結合するというスパコン設計である．

では，ABCI の設計においても，スループットプロセッサへの投資を重視することは変わらないとして，また全体を利用してグランドチャレンジアプリを解く必要のない（要するにスパコンではない）システムであるとして，どのような全体設計が望ましいか検討した．

4.1 インジェクションバンド幅の重要性

図 4 は，AAIC を用いて，ImageNet の 1000 カテゴリ（学習イメージ約 128 万個，検証イメージ 5 万個）を用いた ResNet-50 による学習時間と検証精度のグラフである．ディープラーニングフレームワークとしては，CNTK 2.0 を用いている．学習用の各種設定の詳細については割愛する．

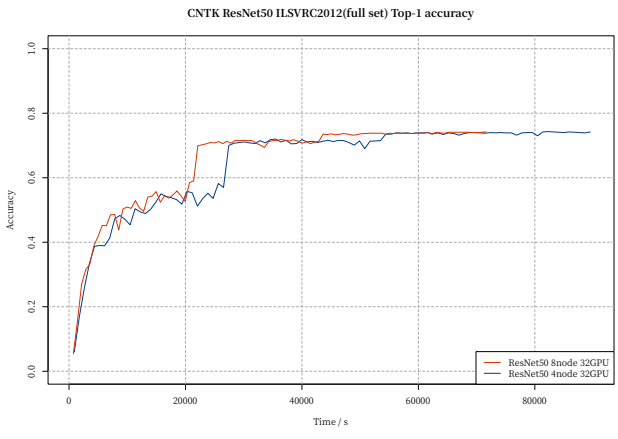

図 3　score と validation metric の関係

Figure 3　Relationship between score and validation metric.

赤線が各ノードの 8GPU のうち 4GPU のみを用い，8 ノード使った場合の，青線が各ノードの 8GPU を使い，4 ノード使った場合の結果である．つまり，どちらも 32GPU 使う．後者はノード内の NVLink を使った高速な集団通信の

恩恵を受けられるが，絶対性能としては前者が計算スループットにおいて 20%も上回るということである．これは，AAIC ではノードで 100Gbps のインジェクションバンド幅を共有するため，前者は GPU あたり 25Gbps，後者は 12.5Gbps であり，ノードをまたがるスループットプロセッサ間の通信がボトルネックとなったと推定される．

したがって，HPC システムと同様，CNN のワークロードにおいても，かつ比較的小規模な並列度の場合であっても，設計上ノードのインジェクションバンド幅を重視することは十分に利得が期待できる．

この実験に関して，追加の実験と詳細な分析が必要であるのは明白であり，発表時点でさらにベンチマーク結果を追加する予定である．ただし，AAIC はすでにランニングシステムであり，また GPU あたりのインジェクションバンド幅がより高い条件での評価等が不可能であるため，将来的には東工大 GSIC の協力の下，TSUBAME3.0 上などでさらなる調査を行う予定である．

4.2 ノードの高インジェクションバンド幅を前提としたインターコネクト設計

ノードあたりのインジェクションバンド幅が大きくなるにつれ，ノード間インターコネクトも巨大なものとなる．その一方で，スケーラビリティに関する研究が比較的先行している CNN の実装でも，state-of-the-art の結果で 256 GPU でのスケーラビリティを確認するに留まっている[7]．したがって，中規模の並列度にフォーカスしてインターコネクトを設計するのは経済的合理性がある．

我々はラック内の結合を密にし，ラック間の結合を疎にすることを前提に設計を行うこととした．具体的には，まず各ラックに N 個のリーフスイッチがあるとき，各スイッチは他のスイッチへの接続バンド幅が当該スイッチに直接接続するノードのインジェクションバンド幅の合計以上とした．N=1 のとき，すべてのノードは単一のディレクタースイッチに接続しているため，フルバイセクションで結合される．N=2 のとき，2 つのスイッチが全ノードのインジェクションバンド幅の 1/2 で結合され，フルバイセクションとなる．N≧3 の場合は，バイセクションバンド幅は抑制され，不均衡な接続も可能となる．

次にラック間の接続は，ラック内の全てのノードのインジェクションバンド幅の合計の 1/3 以上とした．また，上記の不均衡をなるべく解消するため，ラック間の接続に用いられるリンクがラック内のスイッチになるべく均等に分割されることとした．また，ラック間接続のホップ数も抑制した．

こうした設計から導き出される典型的な構成の一つは，各ラックに配されるディレクタースイッチとラック間を接続するスパインスイッチによる構成である．もう一つは，ハイパーキューブであってその低次元にリンクを多く割り当てる構成である．

4.3 ローカルストレージ I/O の重要性

ディープラーニングのワークロードにおいて，学習フェーズは計算インテンシブだが，学習モデルの書き出し時等は小サイズの大量のファイルを生成する．これは，Lustre，GPFS 等 HPC システムで標準的に使われる共有型の並列ファイルシステムが不得手とする I/O であって，しかもランダム読み書き性能の高いローカルストレージ I/O の恩恵を受けやすい．

また，Tesla P100 において，AlexNet の最適化を行った結果，2568 images/sec まで学習スループットが向上したとする NVIDIA による報告がある[7]．仮にイメージあたり 100KB（ImageNet の平均ファイルサイズは 168KB 程度）とすると，各 GPU が約 250MB/sec で読み込む必要があることを意味する．また，Caffe2 による ResNet-50 の学習において，8 基の P100 より 8 基の V100 の方が，約 3 倍スループットが高いという結果も示されている[9]．これらを勘案すると，新しい世代では，スループットプロセッサあたり 750MB/sec 以上の連続読み込み性能が必要だということになる．

これを現状満足するリーズナブルな解は NVMe SSD であり，また計算ノードが搭載するスループットプロセッサの個数に応じた設計が不可欠ということである．

4.4 ABCI の仕様における定義

以上の検討から以下の仕様を定義した．

ノードのネットワークインタフェース

① 理論インジェクションバンド幅は，100Gbps 以上，かつ（100×スループットプロセッサ数÷3）Gbps 以上であること．すべての計算ノードにおいて，（100×スループットプロセッサ数）Gbps 以上となる場合は加点する．

② 全スループットプロセッサが同時に通信する際，理論上①のバンド幅を達成できる構成をとること．

③ 各ネットワークインタフェースの理論バンド幅は，100Gbps 以上とすること．また，各スループットプロセッサは理論上このバンド幅で通信できること．

ラック内・ラック間インターコネクト

ノードを収容するラックにおけるラック内インターコネクト及びラック間インターコネクトは，以下の要件を満たすこと．

① ラック内を接続するネットワークを 2 基以上のスイッチで構成する場合，各スイッチのラック内の他のスイッチへの接続バンド幅の合計は，当該スイッチに直接接続する全てのノードのインジェクションバンド幅の合計以上であること．

② ラック間の接続バンド幅は，ラック内の全てのノードのインジェクションバンド幅の合計の 1/3 以上であること．①と同じく，ラック内ネットワークを 2 基以上のスイッチで構成する場合，ラック間の接続バンド幅を各スイッチでなるべく均等に分割すること．

③ ラック間の接続において，各ラックをスーパーノードとみなしたとき，任意の 2 つのスーパーノード間のネットワーク距離（ホップ数）の最大値が 6 以下となること．

ノードのローカルストレージ

以下の要件を満たす，NVMe SSD もしくはそれに相当するフラッシュストレージを備えること．

① 容量（400×スループットプロセッサ数）GB 以上．
② 連続読み込みバンド幅（500×スループットプロセッサ数）MB/s 以上．
③ 連続書き込みバンド幅（200×スループットプロセッサ数）MB/s 以上．
④ ランダム 4KB 読み込み（90×スループットプロセッサ数）kIOPS 以上．
⑤ ランダム 4KB 書き込み（9×スループットプロセッサ数）kIOPS 以上．
⑥ 書き込み保証値（0.5×スループットプロセッサ数）PBW 以上．

5. ベンチマーク

ABCI は，HPC 向けのシステムではなく，人工知能処理のための「はじめてのスパコン」である．したがって，人工知能処理においてシステムの評価基準となるベンチマークセットを我々自身が新たに定義し，世に問うていく必要がある．また，提案のために事業者が作成するベンチマークプログラムや既存のプログラムへの改変内容は，積極的にオープンにすることを求める．これらはコミュニティの共通知となるべきであると我々は考えている．なぜならば，ABCI に引き続く人工知能処理向けの計算インフラの構築の一助となるとともに，人工知能技術の発展にも寄与するからである．

以下では，我々が ABCI のために定義したベンチマークセットの概略を説明する．一部の基本ベンチマークを除くと，HPC システムに共通するベンチマークを廃して，ビッグデータ処理と人工知能処理に特化したベンチマークを設定していることがお分かりになるだろう．

SPEC CINT2006_rate 及び CFP2006_rate

ノードの OS プロセッサの基本性能を計測する．

Graph 500

複数ノードを用いて Graph 500 ベンチマークの幅優先探索の性能値を計測する．

並列分散ソート

複数ノードを用いて Sort Benchmark に準じたベンチマークを行い，単位時間あたりにソートできるレコード数を計測する．

ローカルストレージの実効性能

ノードに搭載するローカルストレージの読み書き性能を，fio を用いて計測する．

大容量ストレージシステムの全体読み書き性能

IOR を用いて，複数ノード・複数プロセスからの読み書き性能を計測する．

大容量ストレージシステムの単一クライアント読み書き性能

IOR を用いて，1 台のクライアント・1 プロセスからの読み書き性能の計測を行う．

人工知能処理向け数値演算カーネルの実行性能

人工知能処理のワークロードで頻出する行列積処理（GEMM）により AI-FLOPS の実測値を計測する．3.を参照．

畳み込みニューラルネットワーク（CNN）の処理性能

ImageNet の 1000 カテゴリ，訓練イメージ約 128 万個，検証イメージ 5 万個を用いて，ResNet-50 モデルで訓練と検証を行う．単一スループットプロセッサ及び複数ノードで実行し，処理スループットと検証精度で評価を行う．

Chainer の大規模メモリ実行性能

Chainer を使って記述された GoogLeNet V1 の拡張モデルを用いて，仮想的な大きなデータセットの，単一のスループットプロセッサによる学習時間を計測する．スループットプロセッサの広帯域メモリの容量を上回るデータを処理する場合の処理性能を評価する．

再帰型ニューラルネットワーク（RNN）の学習性能

複数のノードを用いて，RNN ベースの機械翻訳アプリケーションである OpenNMT のうち，学習処理の実行性能を計測する．データセットは，WMT15 で提供された英語-ドイツ語の対訳コーパス 3 種類（Europarl v7, Common Crawl corpus, News Commentary v10）を結合したものを用いる．Validation perplexity がしきい値以下になるまでの学習時間を評価する．

6. おわりに

国立研究開発法人産業技術総合研究所では，平成 28 年度二次補正「人工知能に関するグローバル研究拠点整備事業」の一環として，平成 29 年度末に，東京大学柏 II キャンパスに，「AI 橋渡しクラウド（AI Bridging Cloud Infrastructure）」の導入を計画している．本稿では，ABCI のサーバシステムにフォーカスしつつ，ABCI の概要と，システム設計上の論点と我々が採った方策について紹介した．

本発表は，技術提案期間中に行われるため，我々が調達のために用意した各種仕様の範囲内での説明しかできないが，無事導入された暁にはぜひご利用をご検討いただきたい．

謝辞 この研究の一部は，NEDO 次世代人工知能・ロボット中核技術開発の一環として実施した．また，この研究

の一部は，産総研がオープンイノベーションアリーナ構想の一環として平成 29 年 2 月に東工大大岡山キャンパスに設置した，産総研・東工大 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による研究協力の成果である．

ABCI の仕様策定にあたっては，産総研及び東工大のみならず，NICT ユニバーサルコミュニケーション研究所 鳥澤様，田仲様，藤原様，資料提供招請・意見招請にコメントいただいたベンダー各社様のご助力をいただきました．この場を借りて御礼申し上げます．

参考文献

[1] 東工大 TSUBAME3.0 と産総研 AAIC が省エネ性能スパコンランキングで世界 1 位・3 位を獲得！ http://www.aist.go.jp/aist_j/press_release/pr2017/pr20170619/pr20170619.html

[2] AAIC Computer Resource. http://www.airc.aist.go.jp/computer-resources/index.html .

[3] 東工大 TSUBAME3.0 の概要. http://www.gsic.titech.ac.jp/sites/default/files/tsubame3-pc-matsu.pdf .

[4] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the 22nd ACM international conference on Multimedia (MM '14). 2014.

[5] Norman P. Jouppi, et al. In-Datacenter Performance Analysis of a Tensor Processing Unit. The 44th International Symposium on Computer Architecture (ISCA), 2017. https://arxiv.org/abs/1704.04760

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition. CoRR. 2015.

[7] Priya Goyal, Piotr Dollar, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, Kaiming He, Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. https://research.fb.com/publications/imagenet1kin1h/

[8] Michael Houston. Deep Dive on DGX Deep Learning Frameworks: Engineered for Performance. GTC 2017. http://on-demand.gputechconf.com/gtc/2017/video/s7752-michael-o'connor-deep-dive-on-dgx-deep-learning-frameworks-engineered-for-performance.mp4 .

[9] The AI Revolution Is Eating Software: NVIDIA Is Powering It. https://blogs.nvidia.com/blog/2017/05/24/ai-revolution-eating-software/ .

[10] Sort Benchmark, http://sortbenchmark.org/

[11] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia and Marco Turchi, “Findings of the 2015 Workshop on Statistical Machine Translation”, in Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2015. http://aclweb.org/anthology/W15-3001