NEDO委託事業

「次世代人工知能・ロボット中核技術開発」

（人工知能分野）中間成果発表会

－人間と相互理解できる人工知能に向けて－

次世代人工知能フレームワーク・テストベッドの研究開発

平成29年３月29日

国立研究開発法人 産業技術総合研究所

小川 宏高

国立研究開発法人 産業技術総合研究所

国立研究開発法人 新エネルギー・産業技術総合開発機構

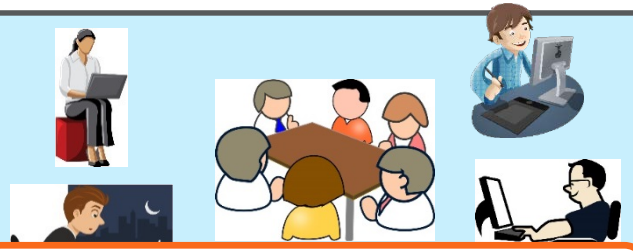
AIクラウドのエコシステム

人工知能の競争力強化には、大規模データの集約と活用、及び要素技術の研究開発と応用実証を可能にするエコシステムの構築が必要

自動運転　インフラ　防災　健康・医療　農業　セキュリティ　・・・
製造現場　流通

①研究開発用途

2）世界最先端のAI研究を産学官連携して推進



シーズプロバイダ
・エンジニア

②ビジネス用途

3）AIの社会実装を促進

ネット上のパブリックデータ
企業保持クローズドデータ
企業保持オープンデータ
府省・自治体オープン

データ収集・クレンジング → モデリング → 学習

学習用データ

社会実装
学習済モデル
学習済モデル
再利用

新たな価値創造・提供(知の循環)

リソースプロバイダ
フレームワーク
ワークフロー管理
リソース管理
クラウドストレージ
AIクラウドハードウェア
AIクラウドライブラリ

データプロバイダ

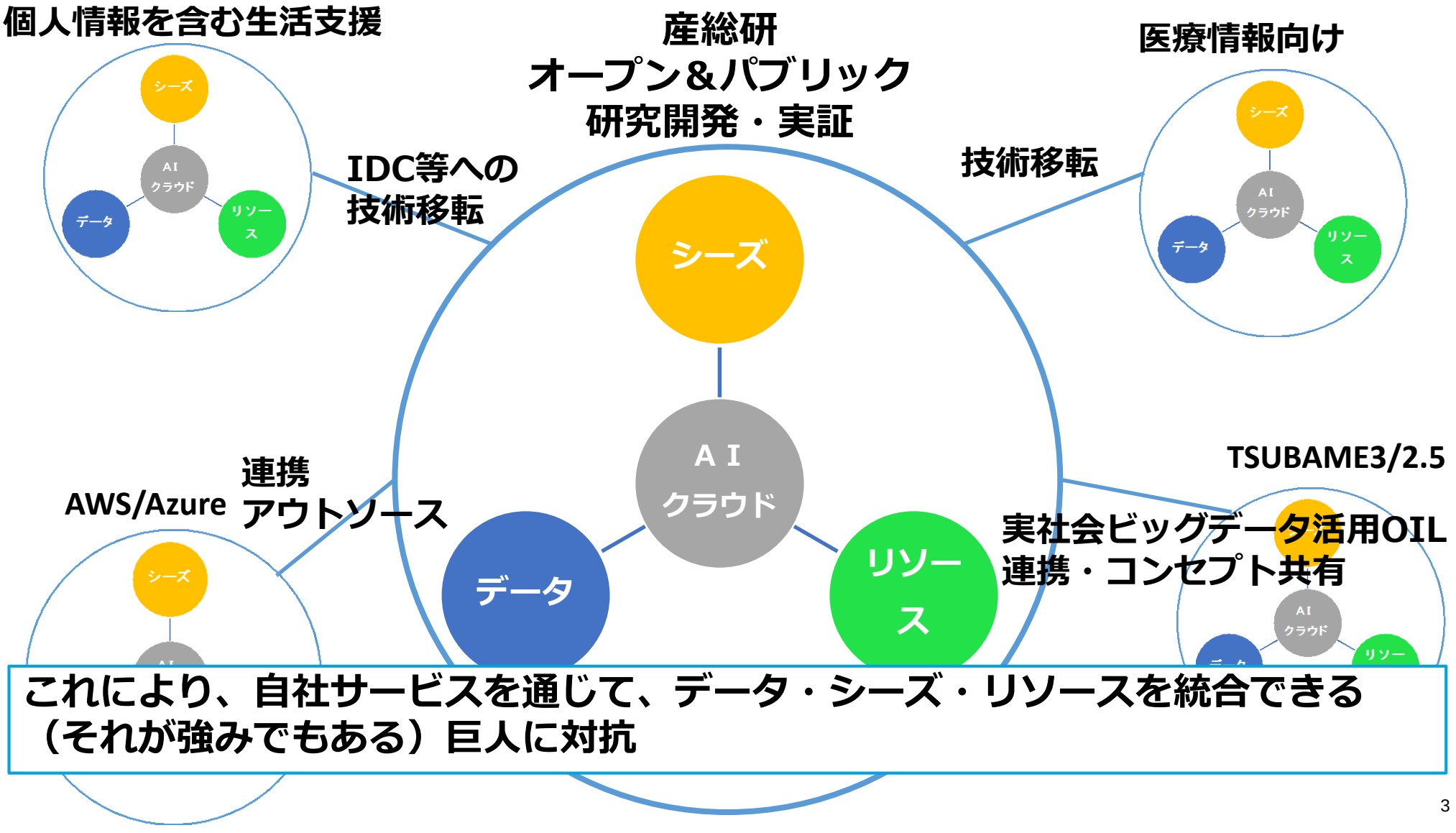
1）世界最先端のAIクラウド計算資源の構築

- 既存のクラウドにも膨大なデータが蓄積、データがある場所で処理した方が合理的

- オープン＆パブリックな衛星画像と、個人情報である医療データ等を、シングルポリシー、シングルシステムで取り扱うのは困難

AIクラウドのエコシステム（続）

- オープン&パブリックな「参照モデル」を構築し、連携や技術移転、あるいは「模倣」を容易に
- AIクラウドの構成要素はなるべくコモディティHW、オープンソース化



これにより、自社サービスを通じて、データ・シーズ・リソースを統合できる（それが強みでもある）巨人に対抗

AIクラウドを実現する計算インフラ

既存クラウド

- **コモディティハードウェアで低密度実装**
- 主に**レイテンシ重視**だが、最近はGPU、SSDもサポート
- **小規模な並列処理**（数十～数百）を提供
- ネットワーク、バイセクションバンド幅、ストレージ**I/O性能が弱い**
- **オープン、パブリックデータセット**が集約
- 高機能、高SLA、相互運用が容易
- TCOは小さいが、深層学習用途には高コスト

既存スパコン・HPC

- **専用ハードウェアで高密度実装**（TSUBAME等は例外）、最新のマルチコアCPU、GPU
- **スループット重視**（倍精度演算のスコア命）
- **大規模並列**（数千～数百万）による高速計算
- ネットワーク、バイセクションバンド幅、高速ストレージなど**I/O性能がリッチ**
- **クローズドな利用環境**、特にデータセットへのアクセスが煩雑
- 高コスト・高TCO、模倣しづらい

amazon web services™ Windows Azure

パブリッククラウド

両者のいいとこどりをしたい

- **AIワークロードにおいて投資対効果の高いアーキテクチャ、特にGPU、マルチコア、FPGAを含む最新の人工知能技術開発に必要なリソースの提供**
- **安価で、模倣しやすいコモディティハードウェアによる高密度実装**
- **ただし、AIに特化した計算インフラの調達、運用組織は世界的に前例がほとんどない**

AIクラウドを実現するSWスタック(1)

既存クラウド

ユーザプログラム

機械学習ライブラリ MLlib/ Mahout

グラフ処理ライブラリ GraphX/ Giraph

SQLクエリエンジン Hive/Pig

Java・Scala・Python＋統合開発環境

MapReduceフレームワーク Spark/Hadoop

RDB PostgresQL

CloudDB/NoSQL Hbase/Cassandra/MondoDB

分散ファイルシステム HDFS

コーディネーションエンジン ZooKeeper

仮想マシン・コンテナ・クラウドサービス

Linux OS

Ethernet ネットワーク

ローカルノードストレージ

x86 CPU

アプリケーション層

- クラウドはプログラムの実行に対話的な操作が必要
- スパコンはバッチジョブによる実行のため不要

システムソフトウェア層

- クラウドは利便性が高いプログラム言語を採用するも高速化には向かない。データ解析等頻繁にプログラムを書き換える利用に特化
- スパコンはマシンの性能を活かせるプログラム言語を採用するも、プログラムが難しく生産性が低い。数値演算などコアな処理はあまりプログラムを書き換える必要がないため
- クラウドはデータベース利用が多い
- スパコンは数千・数万台の計算機向けにデバッグ・性能チューニングが必要
- クラウドは用途に応じた環境構築が可能
- スパコンは高速処理のための環境が主

OS層

ハードウェア層

- スパコンは超広帯域・低遅延ネットワーク、共有ストレージ、GPUなどを採用、高速処理に特化
- クラウドはWebサーバ由来の技術を採用、分散されたストレージ

既存スパコン・HPC

ユーザプログラム

数値計算ライブラリ BLAS

ドメイン固有言語

Fortran・C・C+++統合開発環境

MPI・OpenMP・CUDA/OpenCL

デバッグ・ 性能プロファイル

並列ファイルシステム

バッチジョブスケジューラ

Linux OS

InfiniBabd ネットワーク

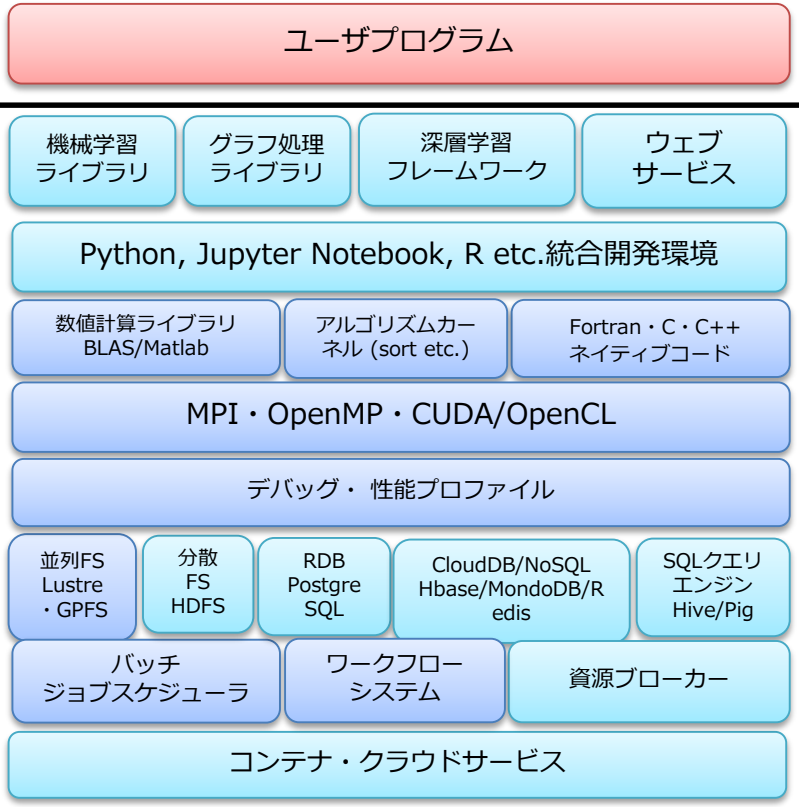
SAN＋ローカルストレージ

X86＋GPU/アクセラレーター

AIクラウドでは既存クラウド・スパコンの両方の技術要素が必要だがそれだけは不十分

AIクラウドを実現するSWスタック(2)

AIクラウド



アプリケーション層

- ✓ Python, Jupyter Notebook, Rなどからの**各種フレームワークの簡便な利用**
- ✓ ウェブを介した**アプリ・サービスの提供**

システムソフトウェア層

- ✓ HPC由来の**数値計算/アルゴリズムカーネルの高速化、特に深層学習の高速化**
- ✓ 学習のための**長時間実行、モジュールベースのワークフロー実行**のサポート
- ✓ コンテナ技術による**ユーザカスタマイズされた複雑なモジュールの簡便な構築・再現性の担保**
- ✓ **大規模データセットへの高速なアクセス**、秘匿データへのセキュリティ
- ✓ 人工知能応用で重要な**時空間データ、機械学習モデルの収集(生成)・蓄積・利用、標準化**

OS層

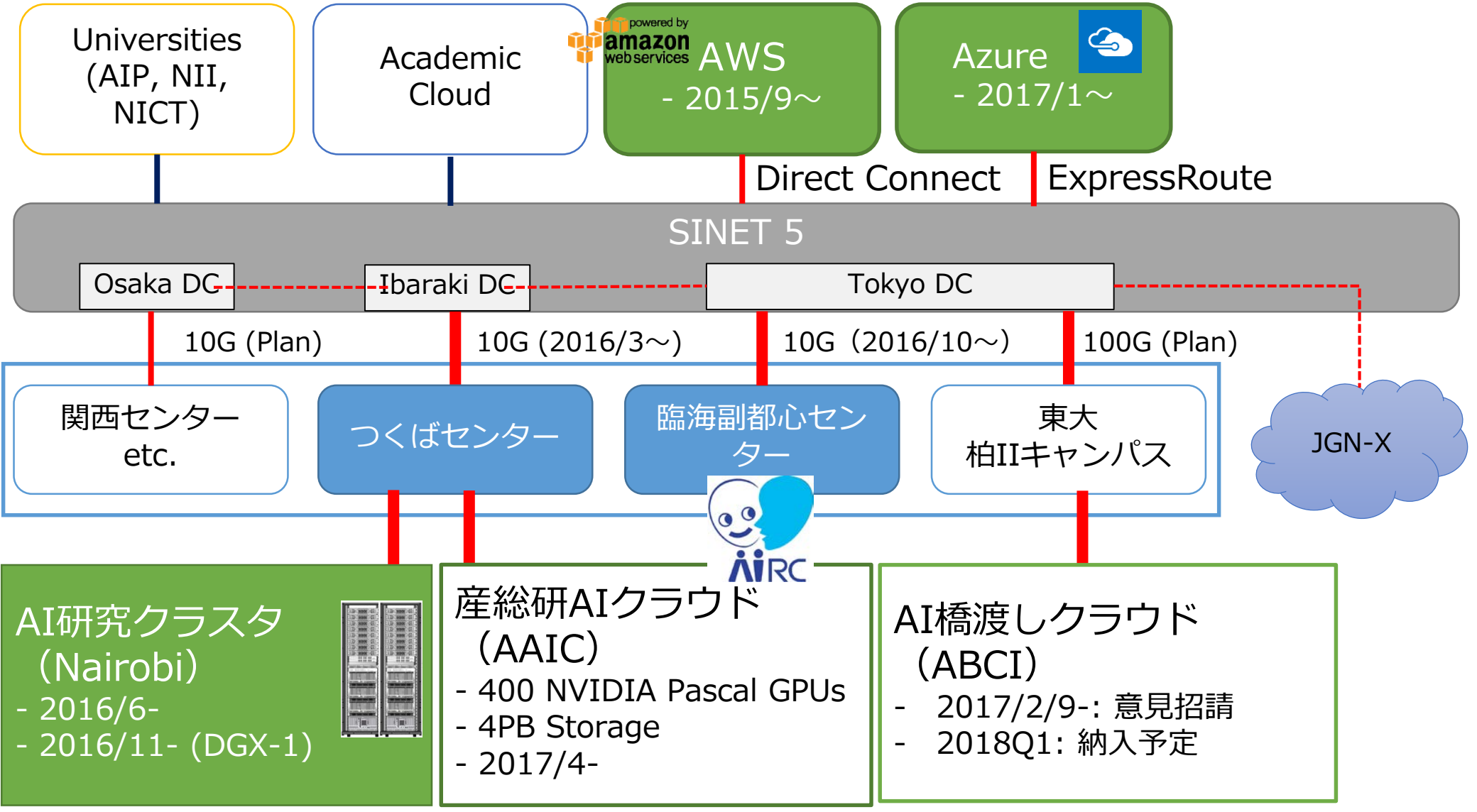
ハードウェア層

- ✓ スパコン由来の**最先端のハードウェア性能を最大限活用するシステムソフトウェア**

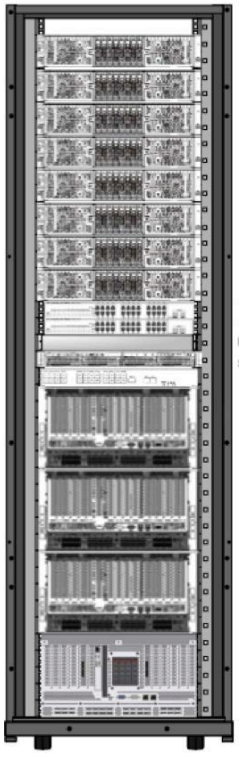
実施内容（テストベッド）

- **人工知能クラウドの構築・運用とエコシステム検討**
  - スパコンとクラウドが融合したAIクラウド（テストベッド）を構築
  - AIクラウドの運営・運用方針の検討と実運用
  - 人工知能技術開発のためのAIクラウドエコシステム、オープンプラットフォームのあり方を検討
- **SW/HW協調によるグランドチャレンジアプリ創出**
  - 人工知能処理向け計算インフラの「性能」を規定可能とするベンチマークAI500の開発
    - 先導研究：AI-FLOPSの定義、後述のABCIの調達仕様に一部ベンチマークを包含
  - スマートシティ、セキュリティ、ヘルスケア、保険、金融、地質調査等ターゲットとなるグランドチャレンジアプリを開拓

テストベッド概要


Universities (AIP, NII, NICT)
Academic Cloud
powered by amazon web services
AWS
- 2015/9～
Azure
- 2017/1～
Direct Connect
ExpressRoute
SINET 5
Osaka DC
Ibaraki DC
Tokyo DC
10G (Plan)
10G (2016/3～)
10G（2016/10～）
100G (Plan)
関西センター etc.
つくばセンター
臨海副都心センター
東大 柏IIキャンパス
JGN-X
AIRC
AI研究クラスタ（Nairobi）
- 2016/6-
- 2016/11- (DGX-1)
産総研AIクラウド（AAIC）
- 400 NVIDIA Pascal GPUs
- 4PB Storage
- 2017/4-
AI橋渡しクラウド（ABCI）
- 2017/2/9-: 意見招請
- 2018Q1: 納入予定

AI研究クラスタ（Nairobi）



GPGPUサーバ × 8台
- 2ソケット, 28コア
- 512GBメモリ
- GPGPU数値演算アクセラレータ × 4
  - 3,072 CUDAコア
  - 12GB GDDR5メモリ
  - 7TFlops (単精度)

大容量メモリサーバ
- 16ソケット, 256コア
- 対称型マルチプロセッシング
- 12TBの単一メモリ空間

GPGPUサーバ × 2台
- 2ソケット, 40コア
- 512GBメモリ
- GPGPU数値演算アクセラレータ × 8
  - 3,584 CUDAコア
  - 16GB HBM2メモリ
  - 21TFlops (半精度)

- NEDOプロジェクト参加者が拠点で共同利用
- 最新の数値演算アクセラレータTesla M40を計32基搭載し、高速なディープラーニング等を支援
- 計16TBの主記憶を搭載し、大容量データのリアルタイムな解析処理、科学技術シミュレーション等を支援
- 2016年6月より稼働

- NVIDIA GDX-1を2台追加導入
- 理研AIPに今月入るものと同じ
- 2016年11月より稼働

産総研の人工知能計算インフラ

NEDO

NEDO次世代人工知能中核技術開発PJ

Nairobiクラスタ

H28.6-

AI研究開発・実証のための研究テストベッド

FY27補正
人工知能・IoT研究開発加速のための環境整備事業の一環

産総研AIクラウド

H29.4-

産総研と連携機関による
AI実証のための共用PF

FY28二次補正
人工知能に関するグローバル研究拠点整備事業の一環

AI橋渡しクラウド

H30.3末以降

複数の産学官による
オープンイノベーション
プラットフォーム
最初からIDCへの技術移転を見越した設計・運用

DL性能	**0.5 PFlops**	約16倍	**8.6 PFlops**	約15倍以上	**>130 PFlops**
HPC性能	**0.2 PFlops**	約10倍	**2.1 PFlops**		**>12 PFlops**
ストレージ	**23 TiB**	約200倍	**4.5 PiB**	約10倍	**>40 PiB**

産総研AIクラウド（AAIC）

FY27補正「人工知能・IoT研究開発加速のための環境整備事業」の一環
４月中旬サービス開始予定（本日これから納品検収）
４月上旬ベンチマーク実施（Top500/Green500）
→ISC17（2017/6）で公表予定

主なスペック

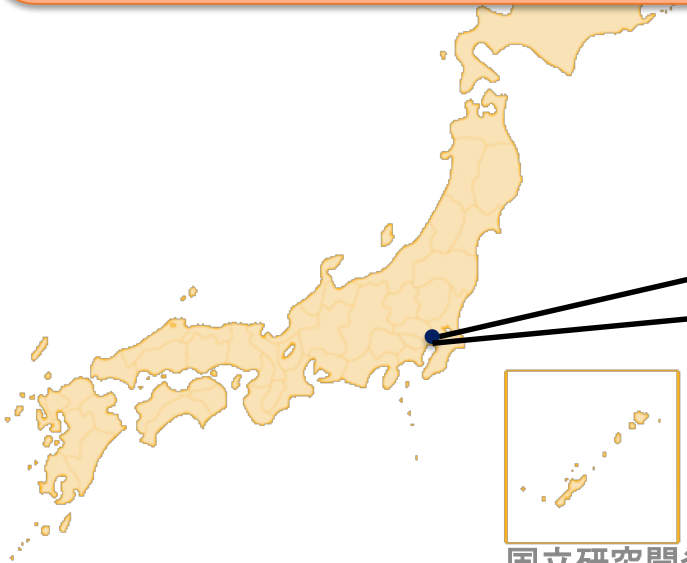
- GPUサーバ 50台
  + CPUサーバ 68台
- GPUサーバはDGX-1の廉価版
- NVIDIA Tesla P100
  NVLinkを計400基搭載
- 4.5PiB GPFSストレージ
  (DDN SFA14K)
- IB EDR 100Gbpsでフルバイセクション構成

AI橋渡しクラウド（ABCI）

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ABCI: AI Bridging Cloud Infrastructure

- **トップスパコン級の計算・データ処理能力（130～ AI-Petaflops）**
- **アルゴリズム・ビッグデータ・計算を集約するオープンな共通基盤**
- **産学官の連携によるAI研究開発の推進**
- **AIワークロードに特化したベンチマークを策定し、評価**

- 130~200 AI-Petaflops
- 消費電力：3MW以下
- 年間平均PUE：1.1以下
- 稼働開始：2018第1四半期以降

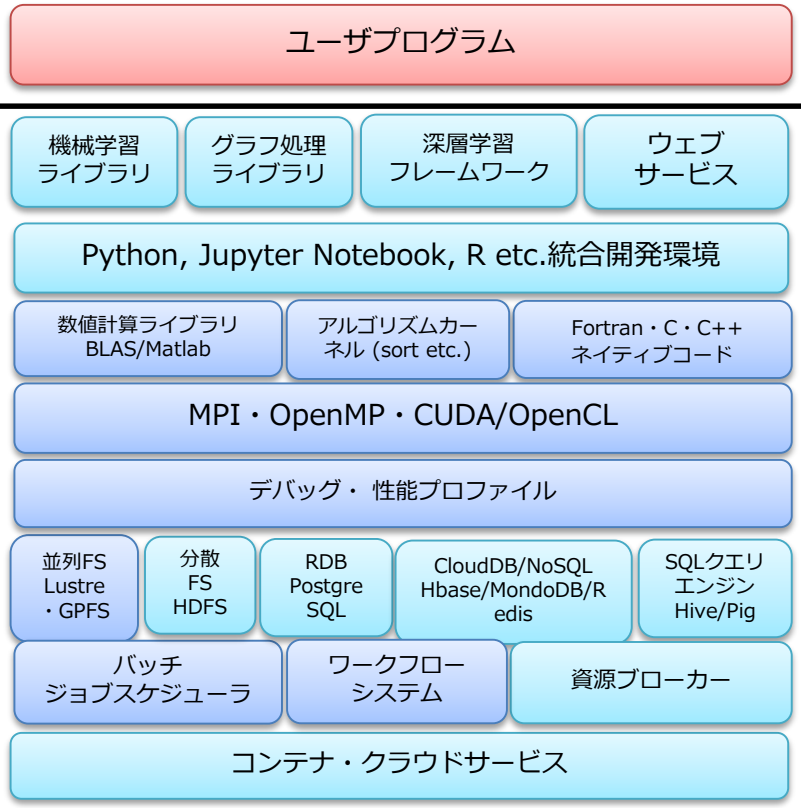


東京大学
THE UNIVERSITY OF TOKYO

AIST
NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

東京大学柏キャンパスに設置

AIクラウドを実現するSWスタック(再掲)



アプリケーション層

- ✓ Python, Jupyter Notebook, Rなどからの**各種フレームワークの簡便な利用**
- ✓ ウェブを介した**アプリ・サービスの提供**

システムソフトウェア層

- ✓ HPC由来の**数値計算/アルゴリズムカーネルの高速化、特に深層学習の高速化**
- ✓ 学習のための**長時間実行、モジュールベースのワークフロー実行**のサポート
- ✓ コンテナ技術による**ユーザカスタマイズされた複雑なモジュールの簡便な構築・再現性の担保**
- ✓ **大規模データセットへの高速なアクセス**、秘匿データへのセキュリティ
- ✓ 人工知能応用で重要な**時空間データ、機械学習モデルの収集(生成)・蓄積・利用、標準化**

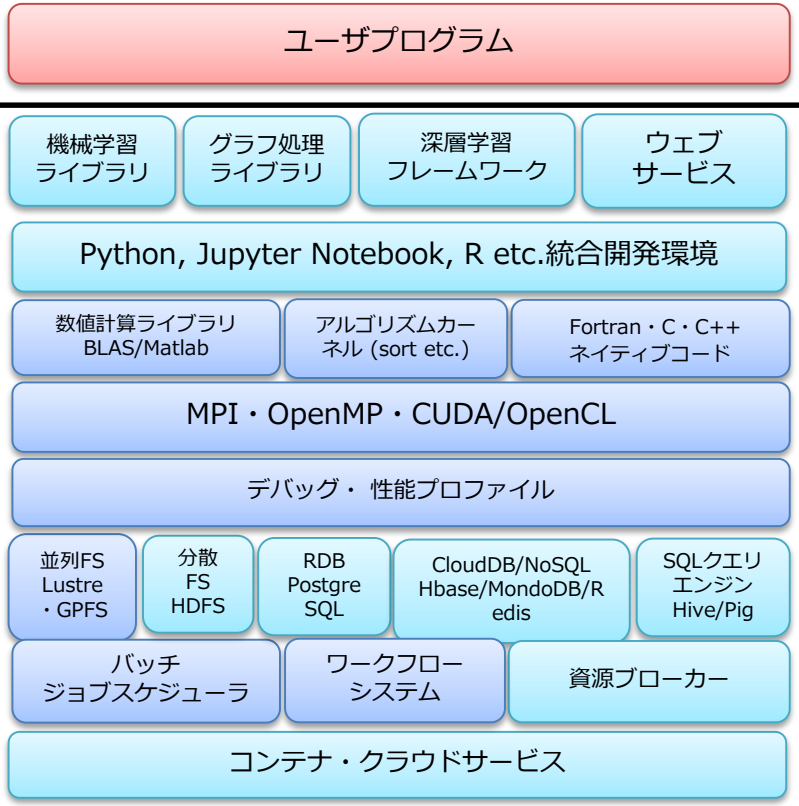
OS層

ハードウェア層

- ✓ スパコン由来の**最先端のハードウェア性能を最大限活用するシステムソフトウェア**

実施内容（フレームワーク）

AIクラウド



アプリケーション層

- ✓ Python, Jupyter Notebook, Rなどからの**各種フレームワークの簡便な利用**
- ✓ ウェブを介した**アプリ・サービスの提供**

システムソフトウェア層

- ✓ HPC由来の**数値計算/アルゴリズムカーネルの高速化、特に深層学習の高速化**
- ✓ 学習のための**長時間実行、モジュールベースのワークフロー実行**のサポート
- ✓ コンテナ技術による**ユーザカスタマイズされた複雑な実行環境の簡便な構築・再現性の担保**
- ✓ **大規模データセットへの高速なアクセス**、秘匿データへのセキュリティ
- ✓ 人工知能応用で重要な**時空間データ、機械学習モデルの収集(生成)・蓄積・利用、標準化**

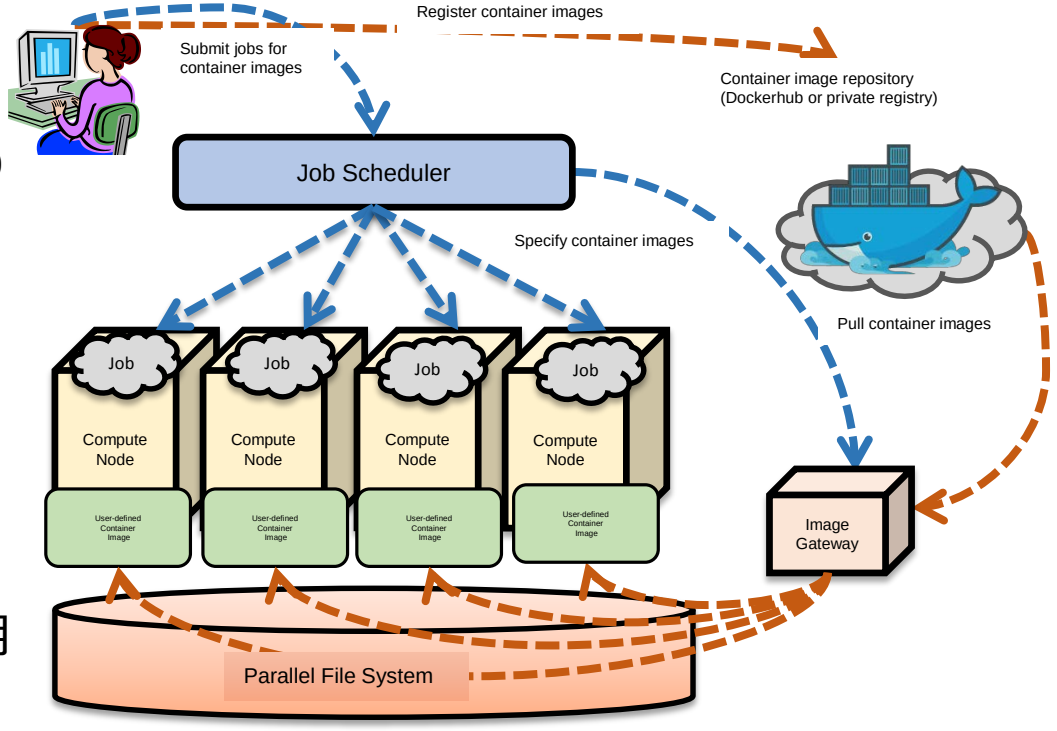
OS層

ハードウェア層

- ✓ スパコン由来の**最先端のハードウェア性能を最大限活用するシステムソフトウェア**

ShifterによるAIワークロード配備システムの実現

- **米国NERSCで開発中のHPC向けコンテナShifter**
  - 世界のトップスパコンでの利用事例: NERSC Cori (Top500 #4), CSCS Piz Daint(Top500 #8), LHC ATLAS(CERN) etc.
- **深層学習を含むAIワークロード向けに適用、実験**
  - ジョブスケジューラと連携してコンテナのイメージを動的に配備
  - Docker Hubなどレポジトリと連携
  - コンテナイメージに対してchrootを適用
- **AIクラウドコンフォーマント**
  - ユーザ権限でプログラムを実行、ストレージへアクセス
  - HPC系のソフトウェアスタック(MPI, CUDA etc.)のサポート
  - 大容量共有ストレージへの非rootアクセス



Nairobi上でプロトタイプ実現

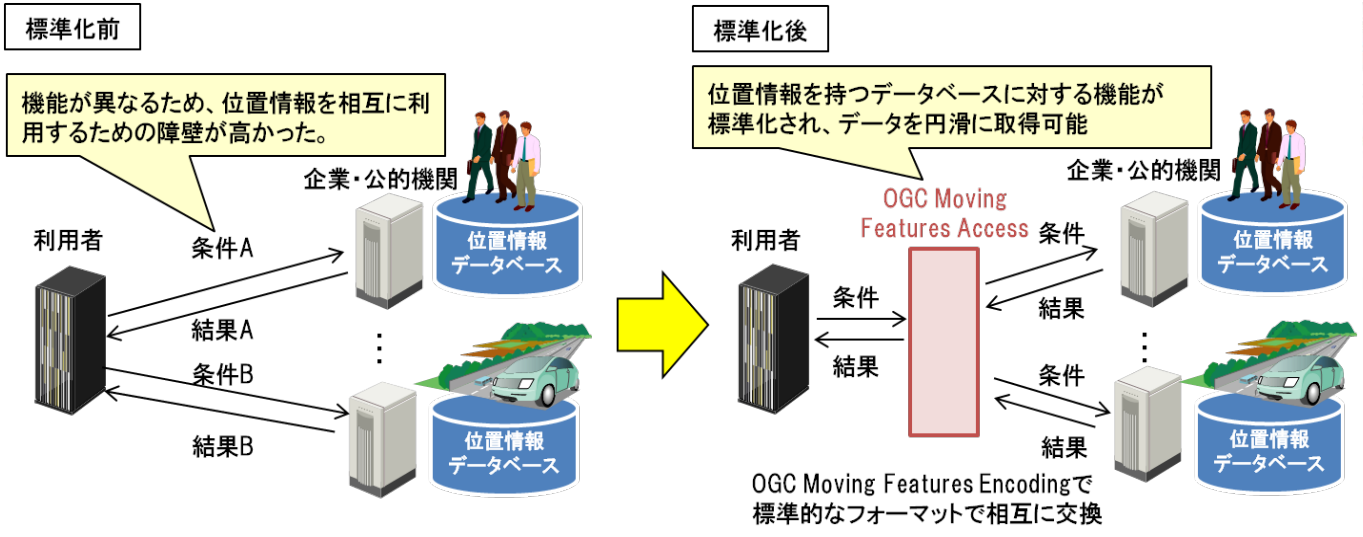
SG以降、早期に産総研AIクラウドでサービス化を図り、共有タスク等での利用を促進

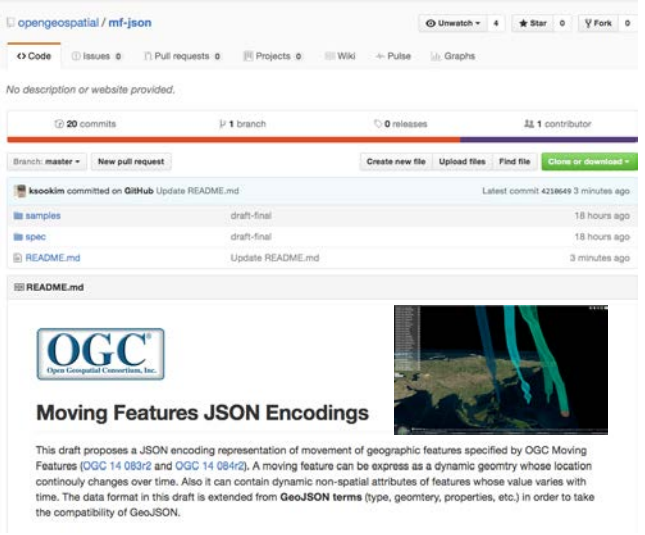
時空間データの利活用を広げる国際標準化の取り組み

- 異種の時空間データを人工知能応用に利活用するための、データ管理・分析データプラットフォームをプロトタイプ実装し、その一部成果を国際標準化
- 人や車など移動物体の位置情報データを横断的に検索・分析する機能仕様を、地理空間情報の国際標準化団体Open Geospatial Consortium(OGC)の標準仕様として提案、採択
- 移動物体の位置情報の軽量なデータ交換形式と、それに基づくデータサービスのAPI仕様を国際標準ベストプラクティスとして提案

【OGC Moving Features Access】
- OGC標準仕様
- 移動物体の位置情報に関する検索・分析機能を標準化することで、迅速かつ横断的に人や車等の位置情報の検索可能に。

【OGC Moving Features JSON Encoding】
- OGCベストプラクティス
- 既存のXMLより簡潔なデータ形式とすることで、処理効率と可読性を向上。





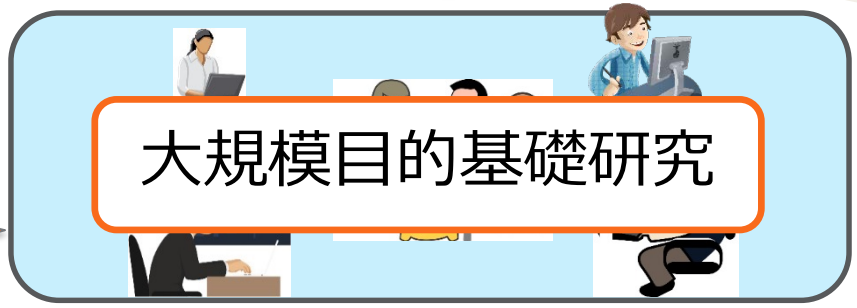
SG以降は、OGC Moving Features AccessとJSON Encodingを活用した
共有タスク等を通じたインターオペラビリティ実証や人工知能応用分野を広げる

ステージゲート後の取り組み

AIクラウドプラットフォームのエコシステム

自動運転　インフラ　防災　健康・医療　農業　セキュリティ　・・・
製造現場　流通

①研究開発用途



②ビジネス用途

2）世界最先端のAI研究を産学官連携して推進

3）AIの社会実装を促進

AIリサーチャー・エンジニア

ネット上の

社会実装

AI for 科学技術研究

AI for ロボット

AI for 生活支援：人間行動モデリングタスク

AI for 地理空間情報：地理空間情報画像解析タスク

AIクラウドストレージ　AIクラウドハードウェア　AIクラウドライブラリ

1）世界最先端のAIクラウド計算資源の構築

補足資料：AI橋渡しクラウド

ABCI: the world's first large-scale OPEN AI Infrastructure

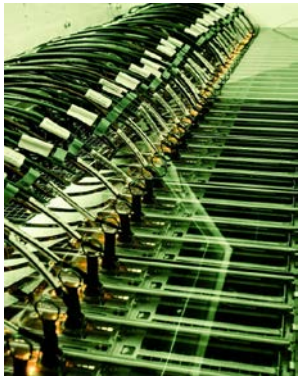
- ABCI: AI Bridging Cloud Infrastructure
  - Top-Level SC compute & data capability: 130~200 AI-Petaflops
  - Open Public & Dedicated infrastructure for Al & Big Data Algorithms, Software and Applications
  - Platform to accelerate joint academic-industry R&D for AI in Japan

Univ. Tokyo Kashiwa Campus

- 130~200 AI-Petaflops
- < 3MW Power
- < 1.1 Avg. PUE
- Operational 2018Q1

ABCI: Overview

- **Extreme computing power**
  - w/ **130～200 AI-PFlops** for AI, ML, DL
  - **x1 million speedup** over high-end PC: 1 Day training for 3000-Year DNN training job
  - TSUBAME-KFC (1.4 AI-Pflops) x 90 users (T2 avg)
- **Big Data and HPC converged modern design**
  - For advanced data analytics (Big Data) and scientific simulation (HPC), etc.
  - Leverage Tokyo Tech's "TSUBAME3" design, **but differences/enhancements being AI/BD centric**
- **Ultra high bandwidth and low latency in memory, network, and storage**
  - For accelerating various AI/BD workloads
  - Data-centric architecture, optimizes data movement
- **Big Data/AI and HPC SW Stack Convergence**
  - Incl. results from JST-CREST EBD
  - **Wide contributions from the PC Cluster community desirable**.
- **RFC just out, includes 10 BD/ML benchmarks**
  - **No HPC benchmarks**



ABCI-IDC: Design

- **Ultra-dense IDC design from ground-up**
  - Custom inexpensive lightweight "warehouse" building w/ substantial earthquake tolerance
  - **x20 thermal density of standard IDC**
- **Extreme green**
  - Ambient warm liquid cooling, large Li-ion battery storage, and high-efficiency power supplies, etc.
  - **Commoditizing supercomputer cooling technologies to Clouds (60KW/rack)**
- **Cloud ecosystem**
  - Wide-ranging Big Data and HPC standard software stacks
- **Advanced cloud-based operation**
  - Incl. dynamic deployment, container-based virtualized provisioning, multitenant partitioning, and automatic failure recovery, etc.
  - Joining HPC and Cloud Software stack for real

CG Image

Reference Image

ABCI Benchmarks

- **Basic performance**
  - Baseline Performance: SPEC CINT2006_rate, CFP2006_rate
  - Local Storage IO: Fio (Flexible IO Tester)
  - Global Storage IO: IOR
- **Big Data workloads**
  - **Graph 500**: breadth-first search in a large undirected graph
  - **MinuteSort**: amount of data that can be sorted in 60.00 seconds or less
- **AI/DNN workloads**
  - GEMM: numerical kernel performance for **DNN-oriented matrix distributions**
  - **Single-node Caffe** performance for AlexNet & GoogLeNet V1
  - **Multiple-nodes Caffe** performance for AlexNet & GoogLeNet V1
  - Chainer performance for GoogLeNet V1 w/ **extra large memory usage**
  - RNN (Recurrent Neural Network) performance

TSUBAME3.0 & ABCI Comparison Chart

	TSUBAME3 (2017/7)	ABCI (2018/3)	K Computer (2012)
AI-FLOPS Peak AI Performance	**47.2 Pflops (DFP 12.1 PFlops) 3.1 PFlops/rack**	**130~200 Pflops (DFP 12~ PFlops) 3~4 PFlops/rack**	11.3 PFlops 12.3 TFlops/rack
System Packaging	**Custom SC (ICE-XA), Liquid Cool**	**19 inch rack (LC), ABCI-IDC**	Custom SC (LC)
Operational Power incl. Cooling	Below 1MW	Approx. 2MW	Over 15MW
Max Rack Thermals & PUE	**61KW, 1.033**	**50-60KW, below 1.1**	~20KW, ~1.3
Node Hardware Architecture	Many-Core (NVIDIA Pascal P100) + Multi-Core (Intel Xeon)	Many-Core AI/DL oriented processor (incl. GPUs)	Heavyweight Multi-Core
Memory Technology	HBM2 + DDR4	On Die Memory + DDR4	DDR3
Network Technology	**Intel OmniPath, 4 x 100Gbps / node, full bisection, optical NW**	**Injection/bisection scaled down c.f. to save cost & IDC friendly**	Copper Tofu 6-D torus custom NW
Per-node non volatile memory	2TeraByte NVMe/node	> 400GB NVMe/node	None
Power monitoring and control	Detailed node / whole system power monitoring & control	Detailed node / whole system power monitoring & control	Whole system monitoring only
Cloud and Virtualization, AI	All nodes container virtualization, horizontal node splits, Cloud API dynamic provisioning, ML Stack	All nodes container virtualization, horizontal node splits, Cloud API dynamic provisioning, ML Stack	None
Procurement Benchmarks	**HPC-Oriented Benchmarks**	**BD & DNN Benchmarks**	HPC Benchmarks