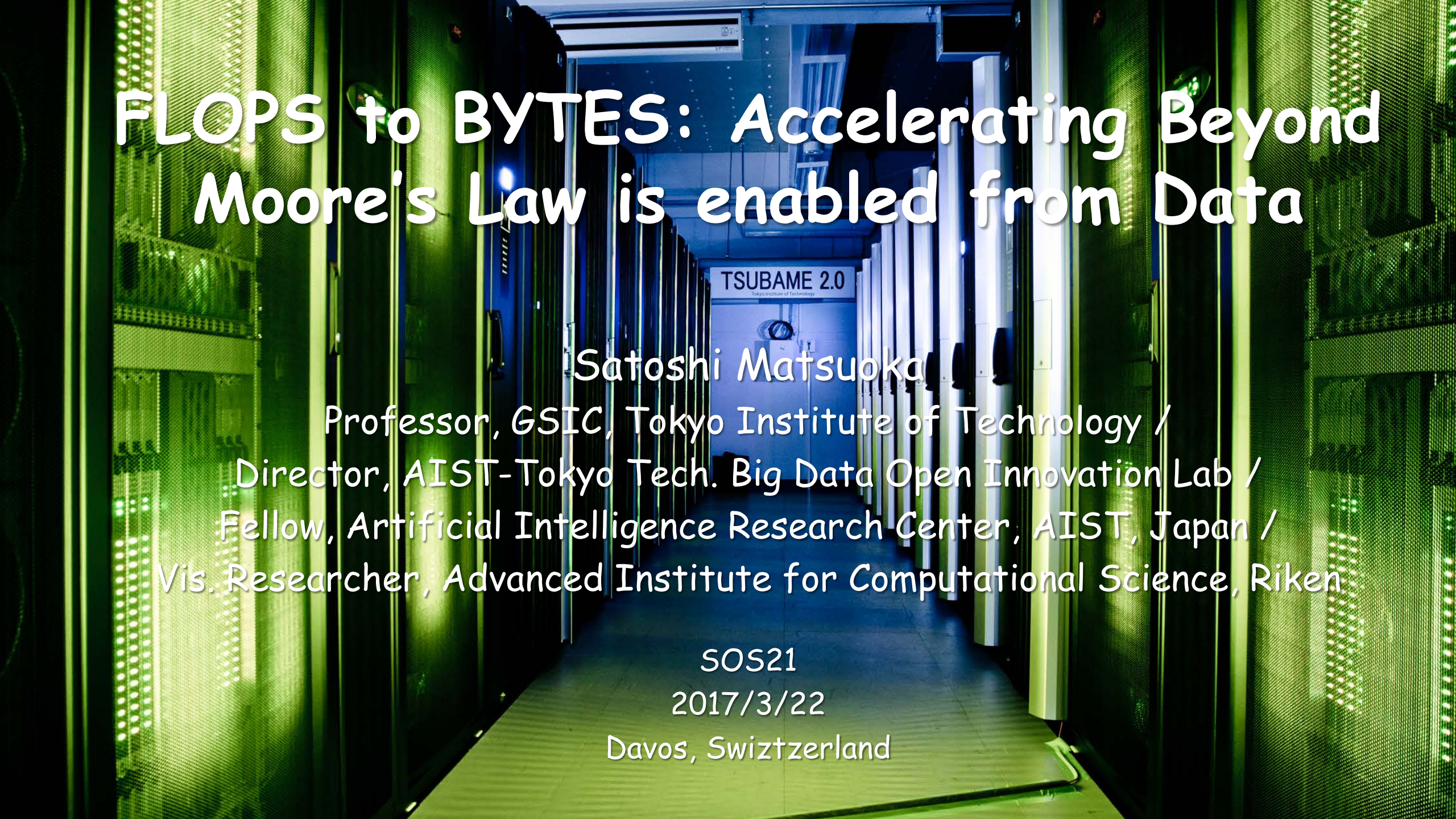


FLOPS to BYTES: Accelerating Beyond Moore's Law is enabled from Data

Satoshi Matsuoka
Professor, GSIC, Tokyo Institute of Technology /
Director, AIST-Tokyo Tech. Big Data Open Innovation Lab /
Fellow, Artificial Intelligence Research Center, AIST, Japan /
Vis. Researcher, Advanced Institute for Computational Science, Riken

SOS21
2017/3/22
Davos, Swiztzerland

Tremendous Recent Rise in Interest by the Japanese Government on Big Data, DL, AI, and IoT

- Three national centers on Big Data and AI launched by three competing Ministries for FY 2016 (Apr 2015-)
  - METI – AIRC (Artificial Intelligence Research Center): AIST (AIST internal budget + > $200 million FY 2017), April 2015
    - Broad AI/BD/IoT, industry focus
  - MEXT – AIP (Artificial Intelligence Platform): Riken and other institutions ($~50 mil), April 2016
    - A separate Post-K related AI funding as well.
    - Narrowly focused on DNN
  - MOST – Universal Communication Lab: NICT ($50~55 mil)
    - Brain –related AI
  - $1 billion commitment on inter-ministry AI research over 10 years

Vice Minsiter Tsuchiya@MEXT Annoucing AIP estabishment

Estimated Compute Resource Requirements for Deep Learning [Source: Preferred Network Japan Inc.]

To complete the learning phase in one day

P:Peta
E:Exa
F:Flops

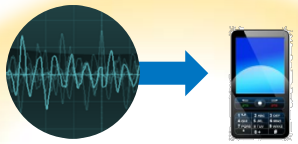
Image/Video Recognition

10P（Image）～ 10E（Video） Flops
学習データ：1億枚の画像 10000クラス分類
数千ノードで6ヶ月 [Google 2015]

Bio / Healthcare

100P ～ 1E Flops
一人あたりゲノム解析で約10M個のSNPs
100万人で100PFlops、1億人で1EFlops

Image Recognition

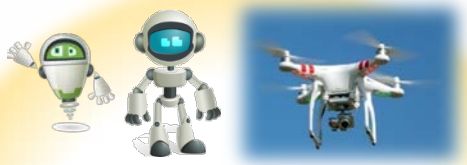


10P～ Flops
1万人の5000時間分の音声データ
人工的に生成された10万時間の
音声データを基に学習 [Baidu 2015]

Auto Driving

1E～100E Flops
自動運転車１台あたり1日 1TB
10台～1000台, 100日分の走行データの学習

Robots / Drones



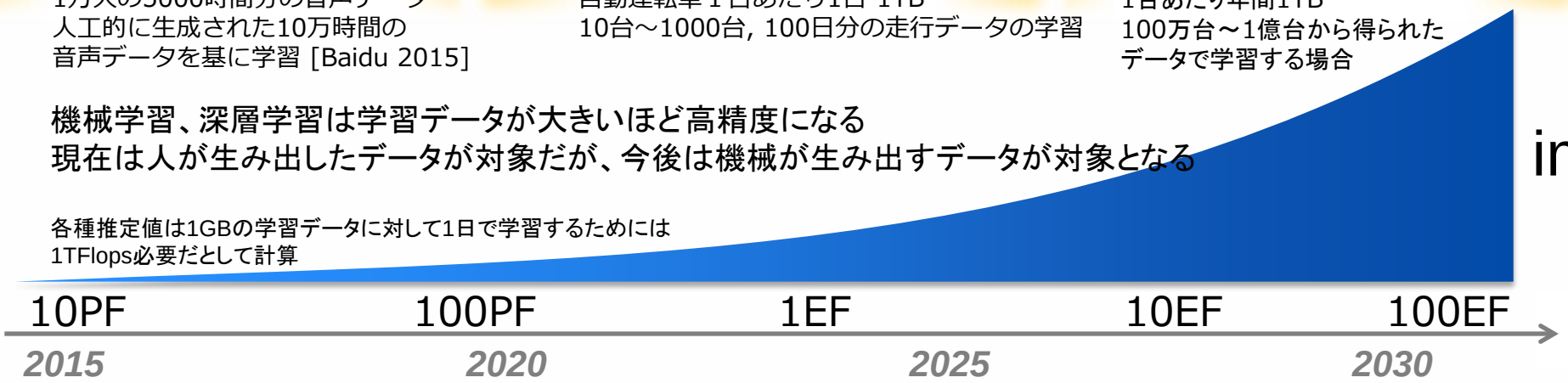
1E～100E Flops
1台あたり年間1TB
100万台～1億台から得られた
データで学習する場合

It’s the FLOPS (in reduced precision) and BW!

So both are important in the infrastructure

機械学習、深層学習は学習データが大きいほど高精度になる
現在は人が生み出したデータが対象だが、今後は機械が生み出すデータが対象となる

各種推定値は1GBの学習データに対して1日で学習するためには
1TFlops必要だとして計算



The current status of AI & Big Data in Japan

We need the triage of advanced algorithms/infrastructure/data but we lack the cutting edge infrastructure dedicated to AI & Big Data (c.f. HPC)

Tokyo Tech

AIRC

Joint RWBC Open Innov. Lab (OIL) (Director: Matsuoka)

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

東京工業大学 Tokyo Institute of Technology

Riken -AIP

NICT- UCRI

AI Venture Startups

Preferred Networks

UEI

AI/BD Centers & Labs in National Labs & Universities

R&D ML Algorithms & SW

Big Companies AI/BD R&D (also Science)

IT LAB DENSO IT LABORATORY, INC.

Panasonic

ABEJA

MIZUHO みずほ情報総研

DeNA

ZMP

Seeking Innovative Application of AI & Data

Use of Massive Scale Data now Wasted

DENSO Petabytes of Drive Recording Video

FANUC FA&ロボット&ロボマシン FA&Robots

TOYOTA

JARI 一般財団法人 日本自動車研究所

Web access and merchandice

Y! YAHOO! JAPAN

SoftBank

NTT

IoT Communication, location & other data

"Big" Data

Massive Rise in Computing Requirements (1 AI-PF/person?)

SAKURA Internet

Google Cloud Platform

Windows Azure

amazon web services

In HPC, Cloud continues to be insufficient for cutting edge research => dedicated SCs dominate & racing to Exascale

Over $1B Govt. AI investment over 10 years

AI&Data Infrastructures

Massive "Big" Data in Training

Example: Tokyo Tech IT-Drug Discovery Factory
Simulation & Big Data & AI at Top HPC Scale

(Tonomachi, Kawasaki-city: planned 2017, PI Yutaka Akiyama)

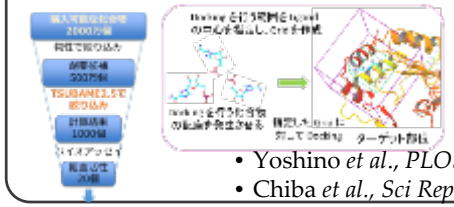
Tokyo Tech's research seeds

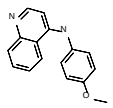
①Drug Target selection system

Minister of Health, Labour and Welfare Award of the 11th annual Merit Awards for Industry-Academia-Government Collaboration

②Glide-based Virtual Screening

TSUBAME's GPU-environment allows **World's top-tier Virtual Screening**

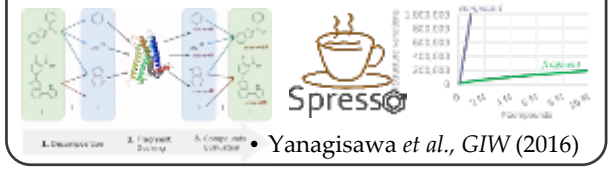




- Yoshino *et al.*, *PLOS ONE* (2015)
- Chiba *et al.*, *Sci Rep* (2015)

③Novel Algorithms for fast virtual screening against huge databases

Fragment-based efficient algorithm designed for **100-millions cmpds data**

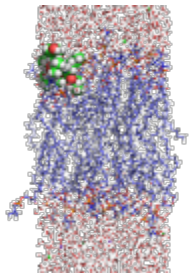


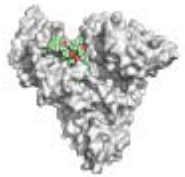
- Yanagisawa *et al.*, *GIW* (2016)

Drug Discovery platform powered by Supercomputing and Machine Learning

Application projects

New Drug Discovery platform especially for specialty peptide and nucl. acids.





Plasma binding (ML-based)

Membrane penetration (Mol. Dynamics simulation)

PeptiDream

Catalyst Inc. (株)カタリスト

DiscoveReSource ディスカヴァリソース(株)

SCHRÖDINGER シュレーディンガー(株)

(株)情報数理バイオ

KING SKYFRONT
Kawasaki Innovation Gateway at SKYFRONT

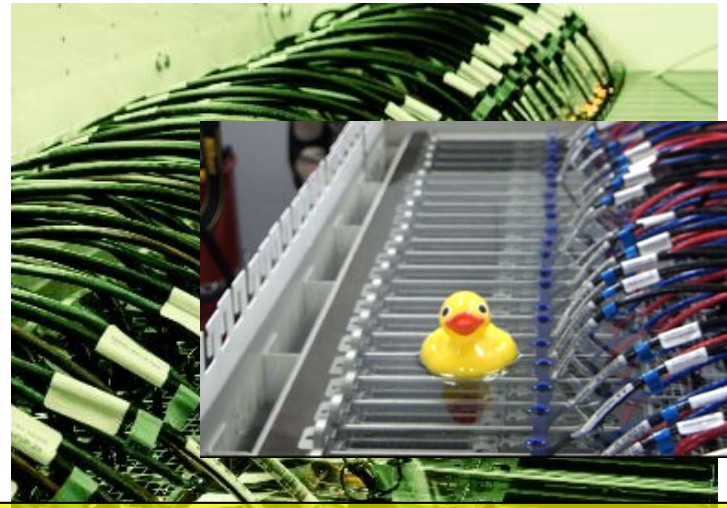
Multi-Petaflops Compute
Peta~Exabytes Data Processing Continuously

Cutting Edge, Large-Scale HPC & BD/AI Infrastructure Absolutely Necessary

Investments from JP Govt., Tokyo Tech. (TSUBAME SC) Muninciple Govt (Kawasaki), JP & US Pharma

TSUBAME-KFC/DL: TSUBAME3 Prototype [ICPADS2014]

Oil Immersive Cooling＋ Hot Water Cooling + High Density Packaging + Fine-Grained Power Monitoring and Control, *upgrade to /DL Oct. 2015*



High Temperature Cooling
Oil Loop 35~45℃
⇒ Water Loop 25~35℃
(c.f. TSUBAME2: 7~17℃)

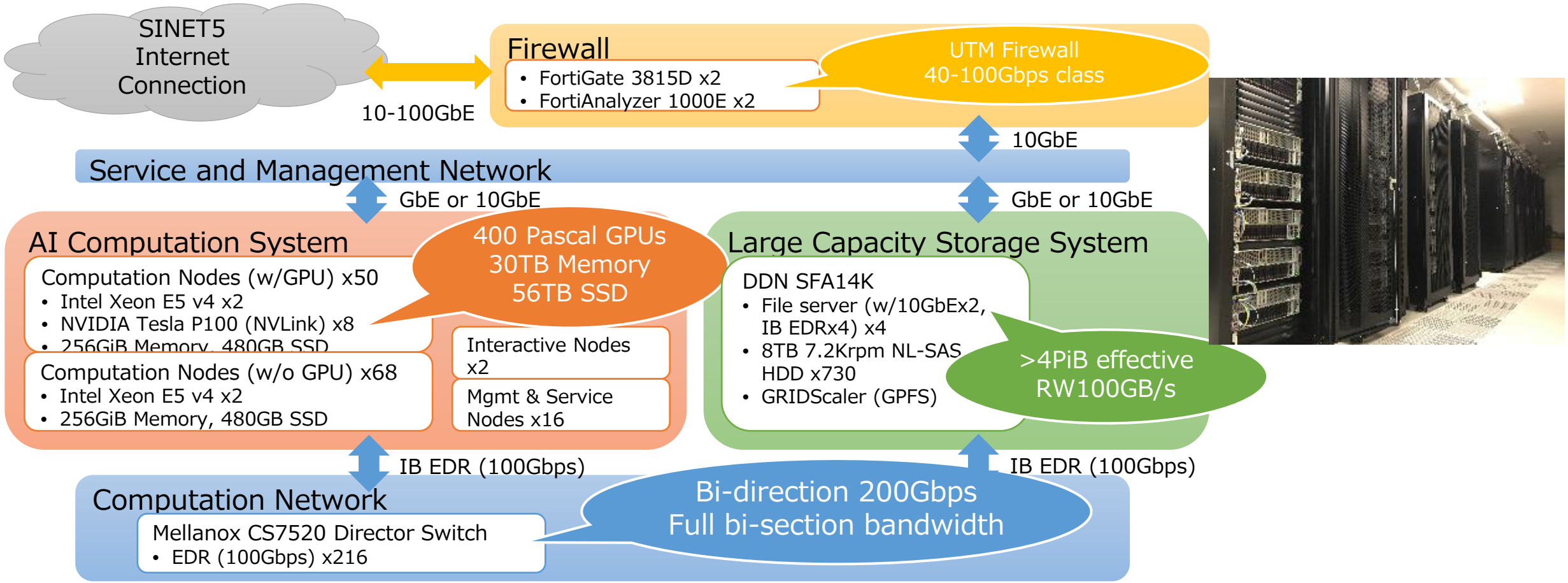
Cooling Tower: Water 25~35℃
⇒ To Ambient Air

Single Rack High Density Oil Immersion
168 NVIDIA K80 GPUs + Xeon
413+TFlops (DFP)
1.5PFlops (SFP)
~60KW/rack

Container Facility
20 feet container (16m²)
Fully Unmanned Operation

ABCI Prototype: AIST AI Cloud (AAIC) March 2017 (System Vendor: NEC)

- **400x NVIDIA Tesla P100s and Infiniband EDR** accelerate various AI workloads including ML (Machine Learning) and DL (Deep Learning).
- Advanced data analytics leveraged by **4PiB shared Big Data Storage and Apache Spark** w/ its ecosystem.



2017 Q2 TSUBAME3.0 Leading Machine Towards Exa & Big Data

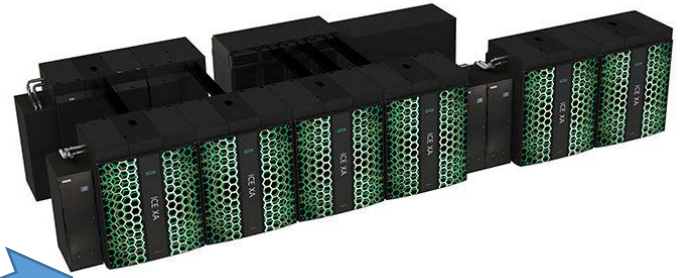
1. "Everybody's Supercomputer" - High Performance (12~24 DP Petaflops, 125~325TB/s Mem, 55~185Tbit/s NW), innovative high cost/performance packaging & design, in mere 180m^2…
2. "Extreme Green" – ~10GFlops/W power-efficient architecture, system-wide power control, advanced cooling, future energy reservoir load leveling & energy recovery
3. "Big Data Convergence" – Extreme high BW &capacity, deep memory hierarchy, extreme I/O acceleration, Big Data SW Stack for machine learning, graph processing, …
4. "Cloud SC" – dynamic deployment, container-based node co-location & dynamic configuration, resource elasticity, assimilation of public clouds…
5. "Transparency" - full monitoring & user visibility of machine & job state, accountability via reproducibility

2006 TSUBAME1.0
80 Teraflops, #1 Asia #7 World
"Everybody's Supercomputer"

2010 TSUBAME2.0
2.4 Petaflops #4 World
"Greenest Production SC"

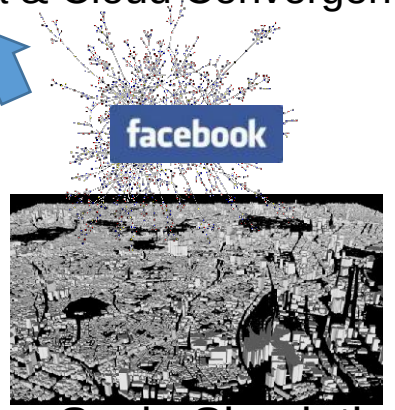
2011 ACM Gordon Bell Prize

2013
TSUBAME2.5
upgrade
5.7PF DFP
/17.1PF SFP
20% power
reduction



2016 TSUBAME3.0+2.5
~20PF(DFP) 4~5PB/s Mem BW
10GFlops/W power efficiency
Big Data & Cloud Convergence

2013 TSUBAME-KFC
#1 Green 500



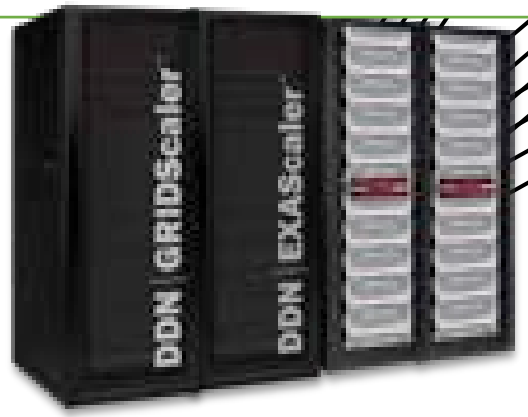
Large Scale Simulation
Big Data Analytics
Industrial Apps

Overview of TSUBAME3.0

Full Operations
Aug. 2017

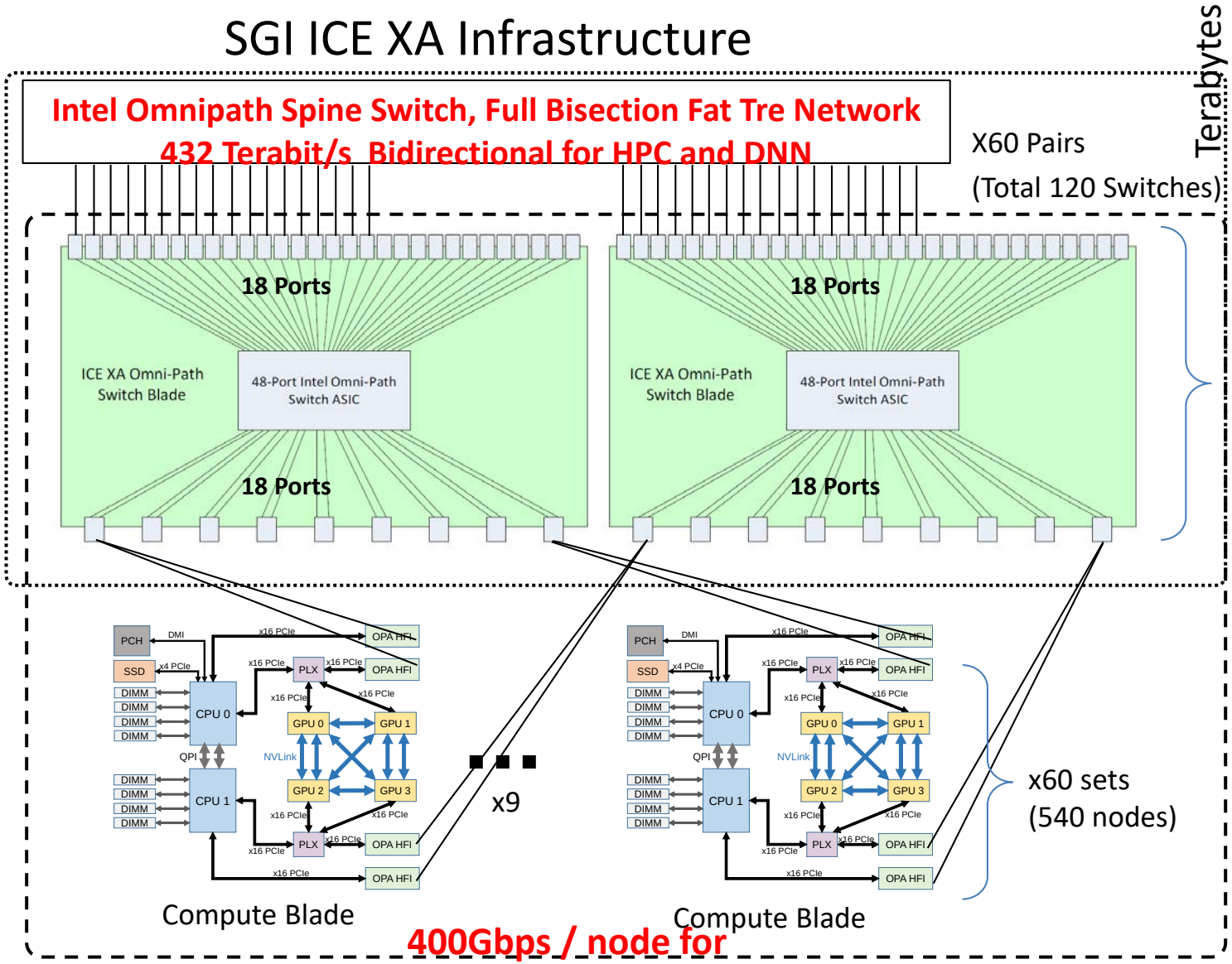
Full Bisection Bandwidgh
Intel Omni-Path Interconnect. 4 ports/node
Full Bisection / 432 Terabits/s bidirectional
~x2 BW of entire Internet backbone traffic

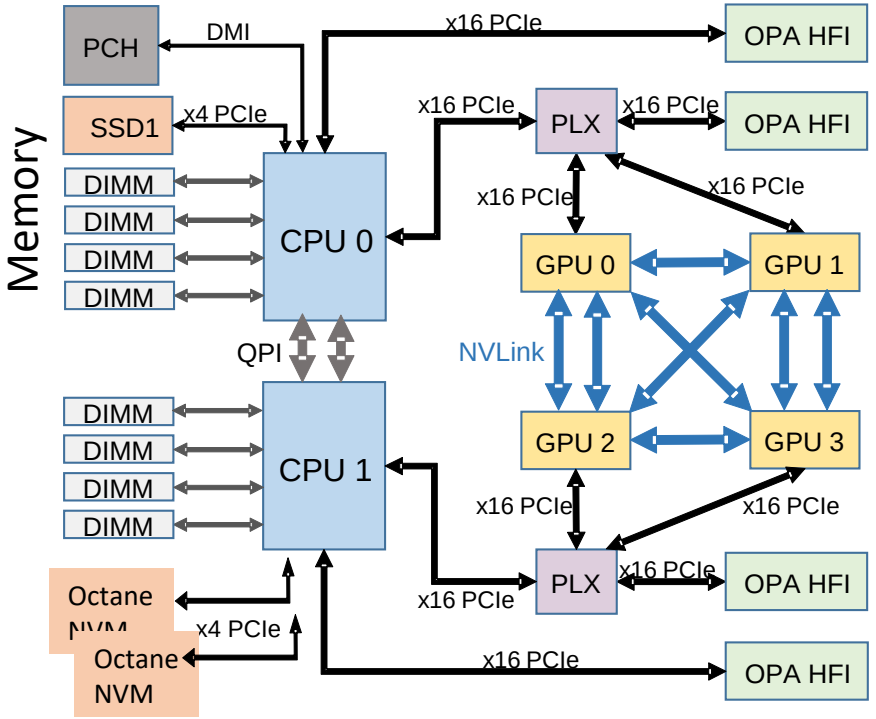
DDN Storage
(Lustre FS 15.9PB+Home 45TB)



540 Compute Nodes SGI ICE XA + New Blade
Intel Xeon CPU x 2+NVIDIA Pascal GPUx4 (NV-Link)
256GB memory 2TB Intel NVMe SSD
47.2 AI-Petaflops, 12.1 Petaflops

TSUBAME3.0 Compute Node SGI ICE-XA, a New GPU Compute Blade Co-Designed by SGI and Tokyo Tech GSIC

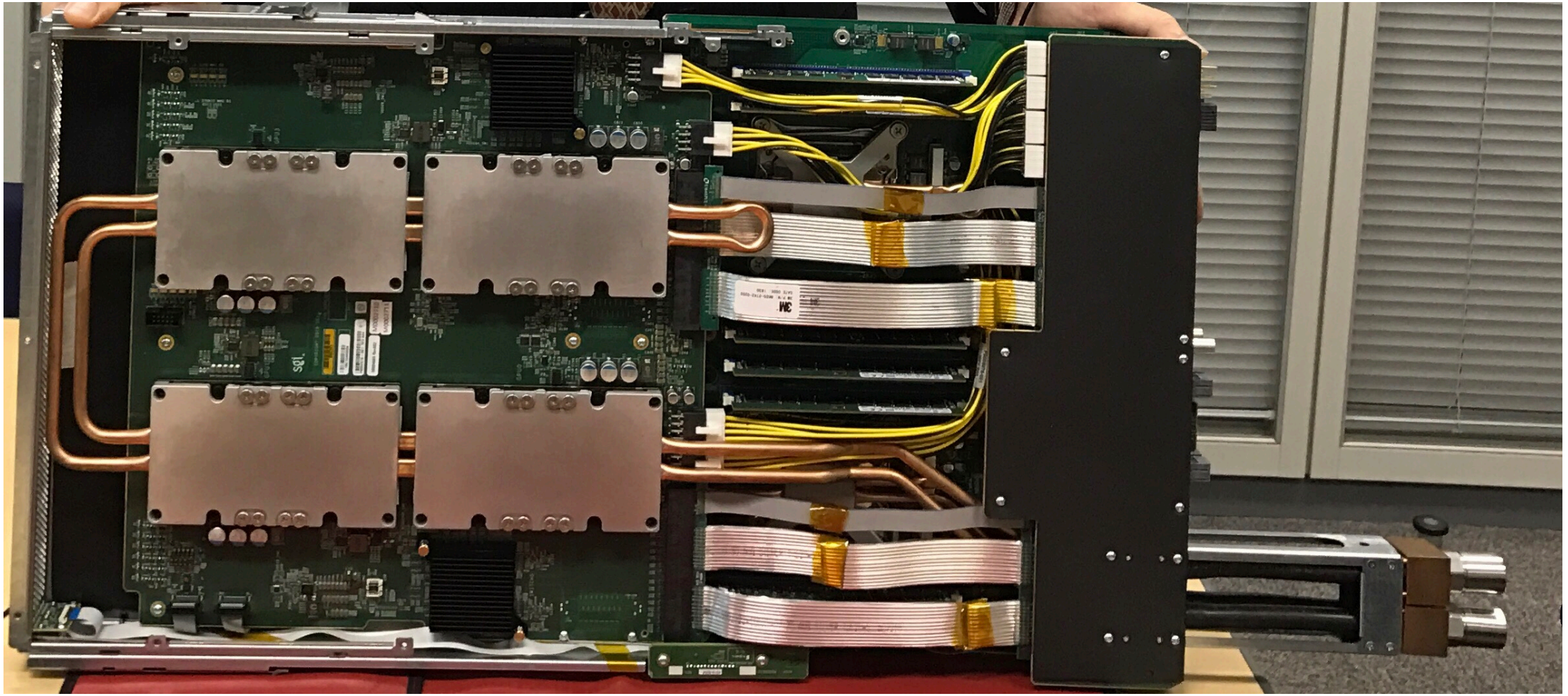




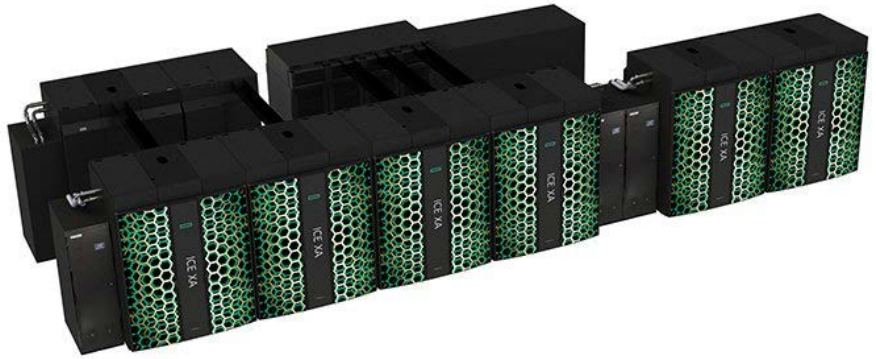
Ultra high performance & bandwidth "Fat Node"

- High Performance: 4 SXM2(NVLink) NVIDIA Pascal P100 GPU + 2 Intel Xeon 84 AI-TFLops
- High Network Bandwidth – Intel Omnipath 100GBps x 4 = 400Gbps (100Gbps per GPU)
- High I/O Bandwidth - Intel 2 TeraByte NVMe
  - > 1PB & 1.5~2TB/s system total
  - Future Octane 3D-Xpoint memory Petabyte or more directly accessible
- Ultra High Density, Hot Water Cooled Blades
  - 36 blades / rack = 144 GPU + 72 CPU, 50-60KW, x10 thermals c.f. IDC

TSUBAME3.0 SGI ICE-XA Blade (new) - Plan to become a future HPE product



TSUBAME3.0 Datacenter

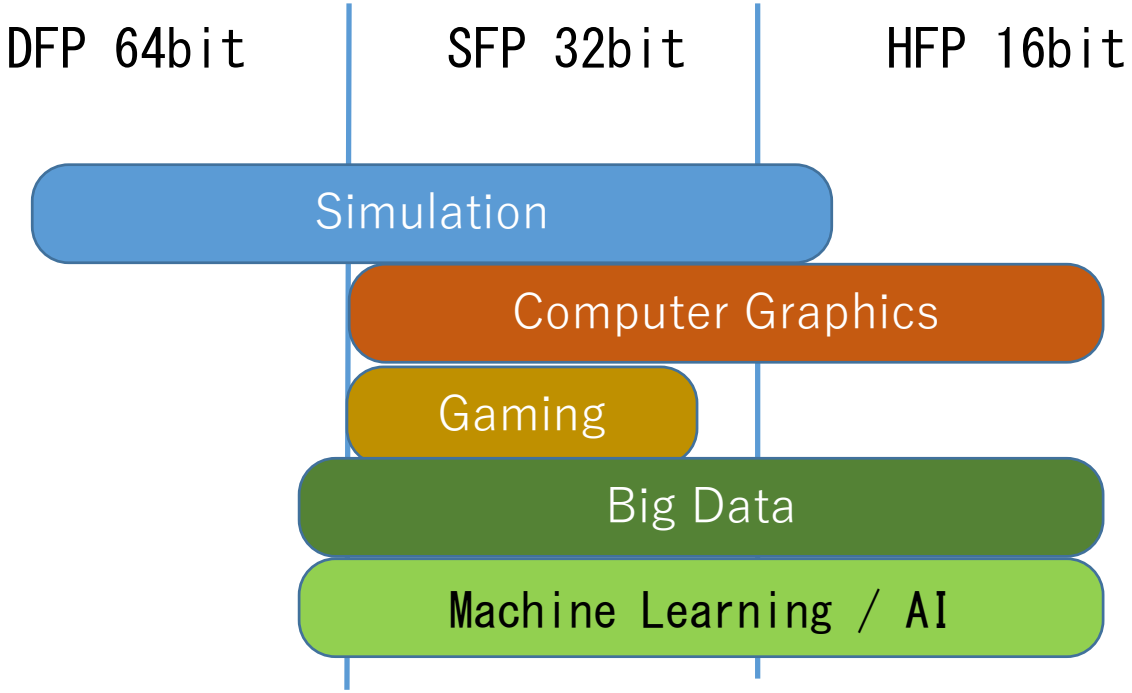


15 SGI ICE-XA Racks
2 Network Racks
3 DDN Storage Racks
20 Total Racks

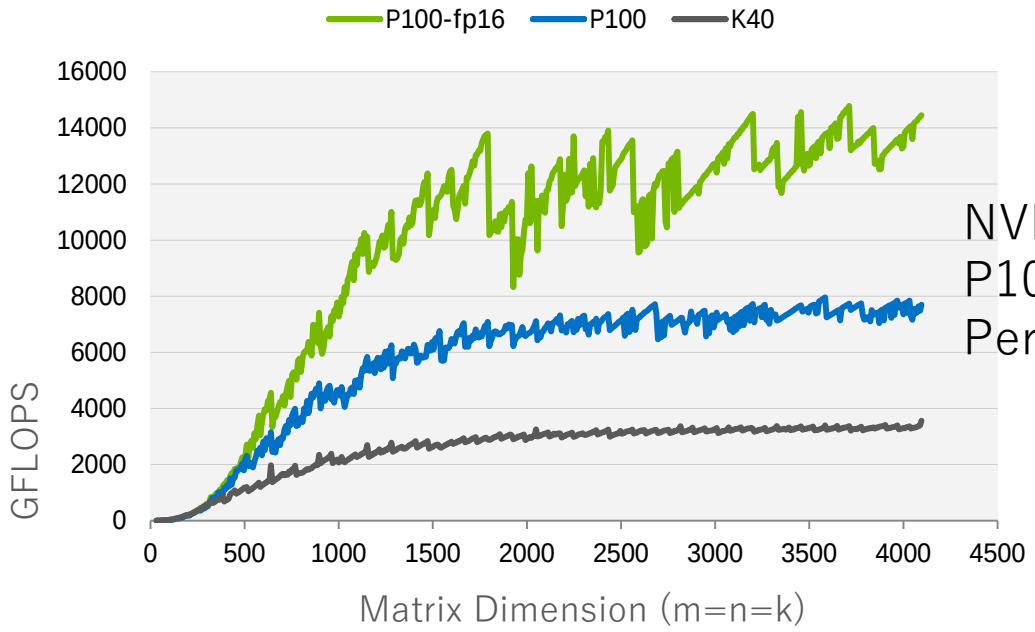
Compute racks cooled with
32 degrees warm water,
yearound ambient cooling
PUE = 1.033

Japanese Open Supercomputing Sites Aug. 2017 (pink=HPCI Sites)

Peak Rank	Institution	System	Double FP Rpeak	Nov. 2016 Top500
1	U-Tokyo/Tsukuba U JCAHP	Oakforest-PACS - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path	24.9	6
2	Tokyo Institute of Technology GSIC	TSUBAME 3.0 - HPE/SGI ICE-XA custom NVIDIA Pascal P100 + Intel Xeon, Intel OmniPath	12.1	NA
3	Riken AICS	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	11.3	7
4	Tokyo Institute of Technology GSIC	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.93GHz, Infiniband QDR, NVIDIA K20x NEC/HPE	5.71	40
5	Kyoto University	Camphor 2 – Cray XC40 Intel Xeon Phi 68C 1.4Ghz	5.48	33
6	Japan Aerospace eXploration Agency	SORA-MA - Fujitsu PRIMEHPC FX100, SPARC64 XIfx 32C 1.98GHz, Tofu interconnect 2	3.48	30
7	Information Tech. Center, Nagoya U	Fujitsu PRIMEHPC FX100, SPARC64 XIfx 32C 2.2GHz, Tofu interconnect 2	3.24	35
8	National Inst. for Fusion Science(NIFS)	Plasma Simulator - Fujitsu PRIMEHPC FX100, SPARC64 XIfx 32C 1.98GHz, Tofu interconnect 2	2.62	48
9	Japan Atomic Energy Agency (JAEA)	SGI ICE X, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR	2.41	54
10	AIST AI Research Center (AIRC)	AAIC (AIST AI Cloud) – NEC/SMC Cluster, NVIDIA Pascal P100 + Intel Xeon, Infiniband EDR	2.2	NA



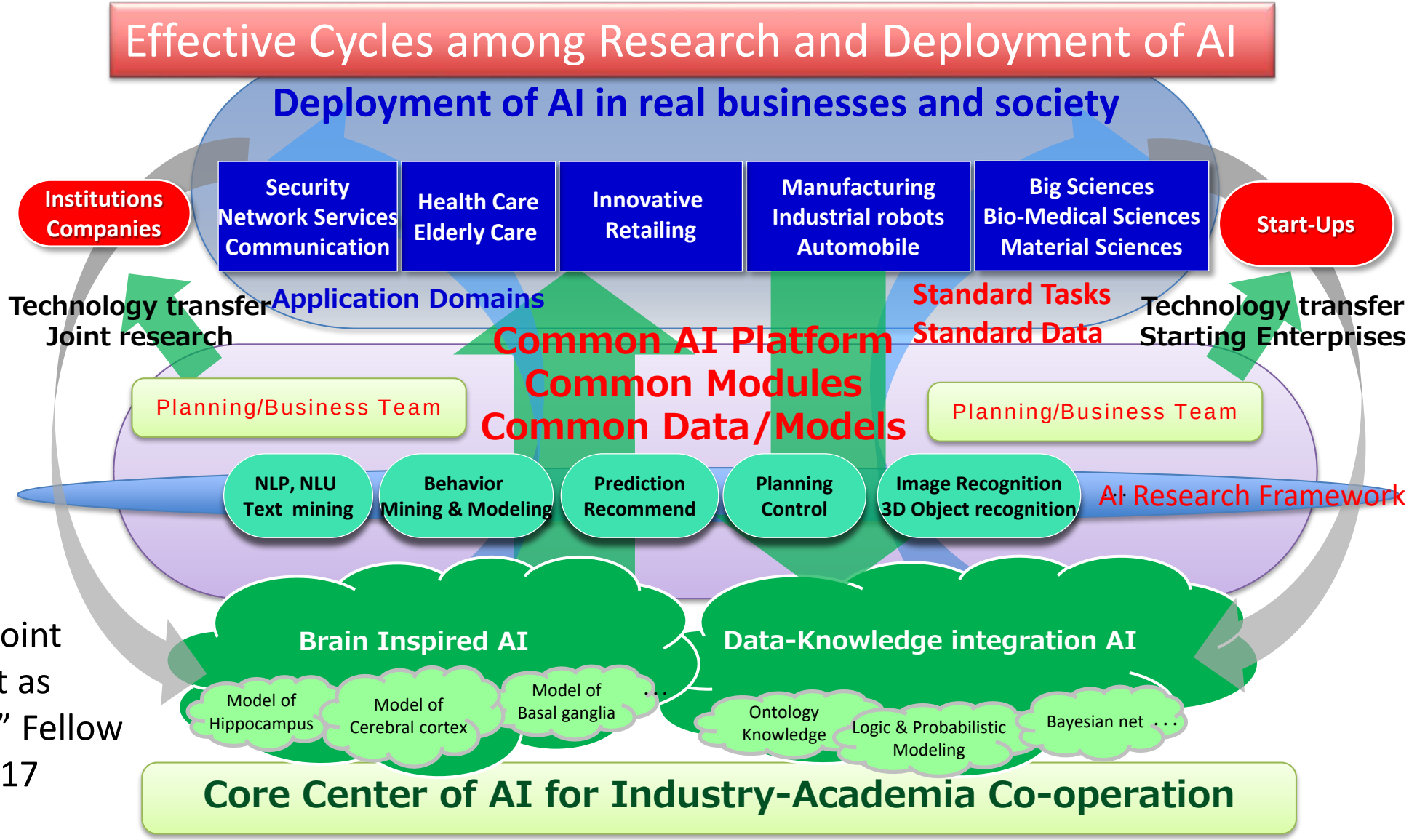
Tokyo Tech GSIC leads Japan in aggregated AI-capable FLOPS TSUBAME3+2.5+KFC, in all Supercomuters and CloudsNV



NVIDIA Pascal P100 DGEMM Performane

AI Research Center (AIRC), AIST

Now > 300+ FTEs



Matsuoka : Joint appointment as “Designated” Fellow since July 2017

National Institute for Advanced Industrial Science and Technology (AIST)

ラボ長（産総研研究職 or 東工大教員/クロスアポ） Director: Satoshi Matsuoka

- 副ラボ長（産総研研究職）
- 副ラボ長（産総研事務職）
- ラボ研究主幹（産総研研究職）
- ラボ構成員

TSUBAME
Tokyo Institute of Technology

GSIC (HPC)

Ministry of Economics Trade and Industry (METI)

Resources and Acceleration of AI / Big Data, systems research

Tsubame 3.0/2.5 Big Data /AI resources

GSIC
Global Scientific Information and Computing Center

Joing Organization@Odaiba

AIST-TokyoTech Real World Big Data /AI Open Innovation Laboratory (OIL)

AIST Artificial Intelligence Research Center（AIRC）

Application Area
Natural Langauge Processing
Robotics
Security

Joint Research on AI / Big Data and applications

Industrial Collaboration in data, applications

Basic Research in Big Data / AI algorithms and methodologies

ITCS Departments

Other Big Data / AI research organizations and proposals

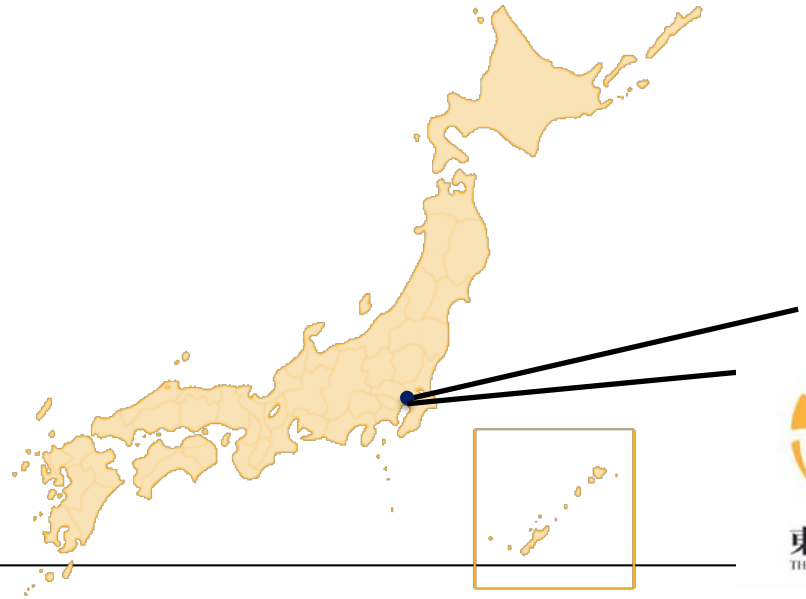
Industry

DENSO IT LABORATORY, INC.

METI AIST-AIRC ABCI

as the *worlds first large-scale OPEN AI Infrastructure*

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



- 130~200 AI-Petaflops
- < 3MW Power
- < 1.1 Avg. PUE
- Operational 2017Q4 ~2018Q1

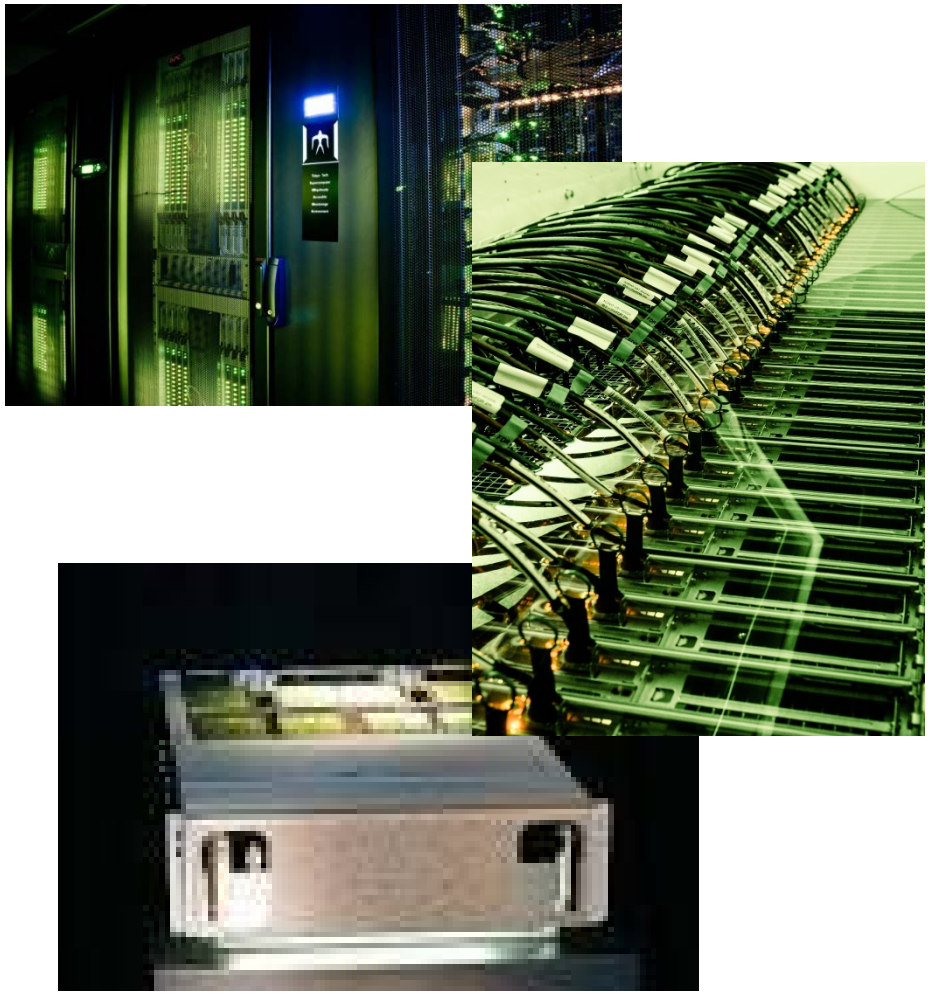
東京大学 THE UNIVERSITY OF TOKYO

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

Univ. Tokyo Kashiwa Campus

ABCI – 2017Q4~ 2018Q1

- **Extreme computing power**
  - w/ **130~200 AI-PFlops** for AI/ML especially DNN
  - **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.**



ABCI Cloud Infrastructure

- **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

ABCI AI-IDC CG Image

イメージスケッ

Reference Image

引用元: NEC導入事例

ABCI Procurement Benchmarks

- Big Data Benchmarks
  - (SPEC CPU Rate)
  - Graph 500
  - MinuteSort
  - Node Local Storage I/O
  - Parallel FS I/O

No traditional HPC Simulation Benchmarks Except SPECCPU

- AI/ML Benchmarks
  - Low precision GEMM
    - CNN Kernel, defines “AI-Flops”
  - Single Node CNN
    - AlexNet and GoogLeNet
    - ILSVRC2012 Dataset
  - Multi-Node CNN
    - Caffe+MPI
  - Large Memory CNN
    - Convnet on Chainer
  - RNN / LSTM
    - To be determined

Software Ecosystem for HPC in AI

Different SW Ecosystem between HPC and AI/BD/Cloud
How to achieve convergence—for real, for rapid tech transfer

Existing Clouds

BD/AI User Applications

Machine Learnig MLlib/ Mahout/Chainer | Graph Processing GraphX/ Giraph /ScaleGraph | SQL/Non-SQL Hive/Pig

Java · Scala · Python + IDL

MapReduce Framework Spark/Hadoop

RDB PostgresQL | CloudDB/NoSQL Hbase/Cassandra/MondoDB

Distributed Filesysem HDFS & Object Store | Coordination Service ZooKeeper

VM(KVM), Container(Docker), Cloud Services (OpenStack)

Linux OS

Ethernet TOR Swtiches High Latency/Low Capacity NW | Local Node Storage | x86 CPU

Application Layer

- Cloud Jobs often Interactive w/resource control REST APIs
- HPC Jobs are Batch-Oriented, resource control by MPI

System Software Layer

- Cloud employs High Productivity Languages but performance neglected, focus on data analytics and dynamic frequent changes
- HPC employs High Performance Languages but requires Ninja Programmers, low productivity. Kernels & compilers well tuned & result shared by many programs, less rewrite
- Cloud focused on databases and data manipulation workflow
- HPC focused on compute kernels, even for data processing. Jobs scales to thousands of jobs, thus debugging and performance tuning
- Cloud requires purpose-specific computing/data environment as well as their mutual isolation & security
- HPC requires environment for fast & lean use of resources, but on modern machines require considerable system software support

OS Layer

Hardware Layer

- Cloud HW based on Web Server "commodity" x86 servers, distributed storage on nodes assuming REST API access
- HPC HW aggressively adopts new technologies such a s GPUs, focused on ultimate performance at higher cost, shared storage to support legacy apps

Existing Supercomputers

HPC User Code

Numerical Libraries LAPACK, FFTW | Various DSLs | Workflow Systems

Fortran · C · C++ + IDL

MPI · OpenMP/ACC · CUDA/OpenCL

Parallel Debuggers and Profilers

Parallel Filesystem Lustre, GPFS, | Batch Job Schedulers PBS Pro, Slurm, UGE

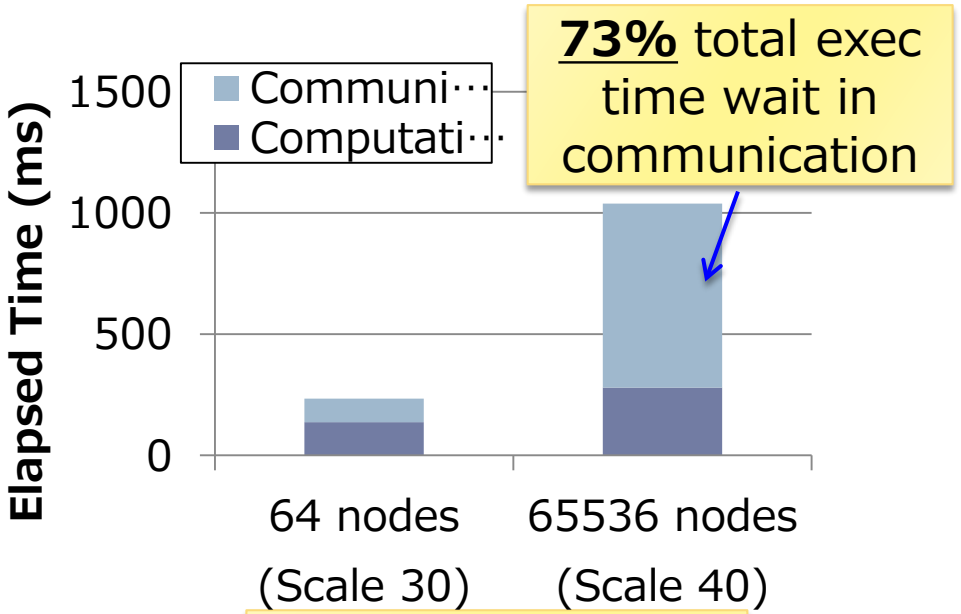
Linux OS

InfiniBand/OPA High Capacity Low Latency NW | High Performance SAN+Burst Buffers | X86 + Accelerators e.g. GPUs, FPGAs

Various convergence research efforts underway but no realistic converged SW Stack yet➔ What is the Low Hanging Fruit?

The Graph500 – 2015~2016 – 4 Consecutive world #1

K Computer #1 Tokyo Tech[EBD CREST] Univ. Kyushu [Fujisawa Graph CREST], Riken AICS, Fujitsu



73% total exec time wait in communication

88,000 nodes,
660,000 CPU Cores
1.3 Petabyte mem
20GB/s Tofu NW

K computer

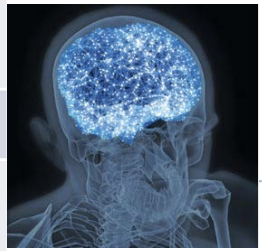
Effective x13 performance c.f. Linpack

TaihuLight
10 million CPUs
1.3 Petabyte mem

LLNL-IBM Sequoia
1.6 million CPUs
1.6 Petabyte mem

IBM

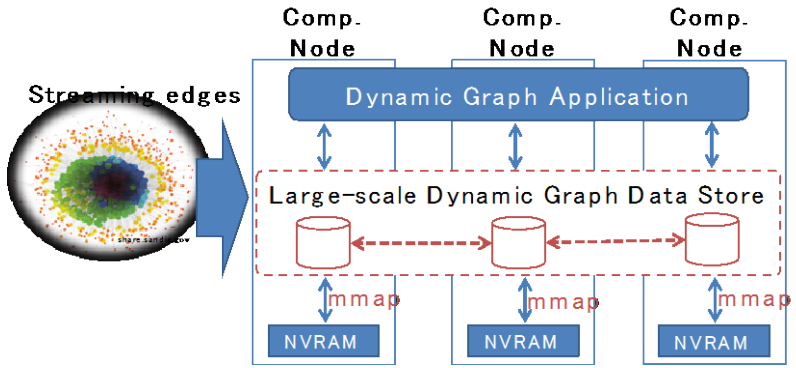
*Problem size is weak scaling
"Brain-class" graph



List	Rank	GTEPS	Implementat
November 2013	4	5524.12	Top-down or
June 2014	1	17977.05	Efficient hybrid
November 2014	2		Efficient hybrid
June, Nov 2015 June Nov 2016	**1**	**38621.4**	Hybrid + Node Compression

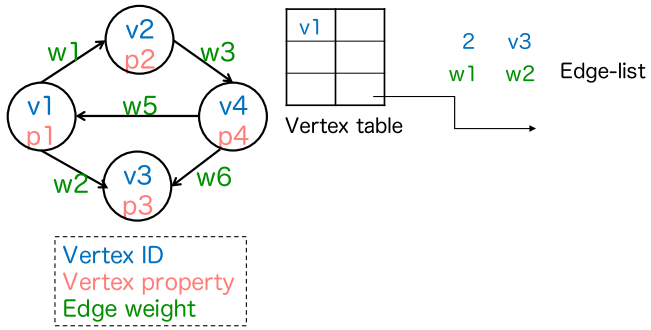
Towards a Distributed Large-Scale Dynamic Graph Data Store (SC16)

Goal: to develop the data store for large-scale dynamic graph analysis on supercomputers



Node Level Dynamic Graph Data Store

Follows an adjacency-list format and leverages an open address hashing to construct its tables



Extend for multi-processes using an async MPI communication framework

Dynamic Graph Construction (on-memory)

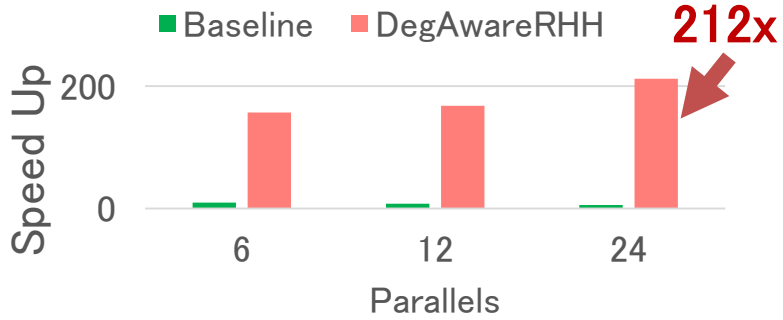
Against STINGER (single-node)

STINGER

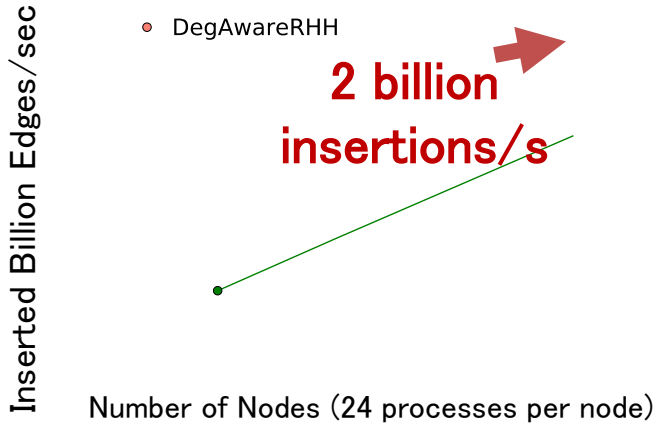
- A state-of-the-art dynamic graph processing framework developed at Georgia Tech

Baseline model

- A naïve implementation using *Boost* library (C++) and the MPI communication framework



Multi-node Experiment

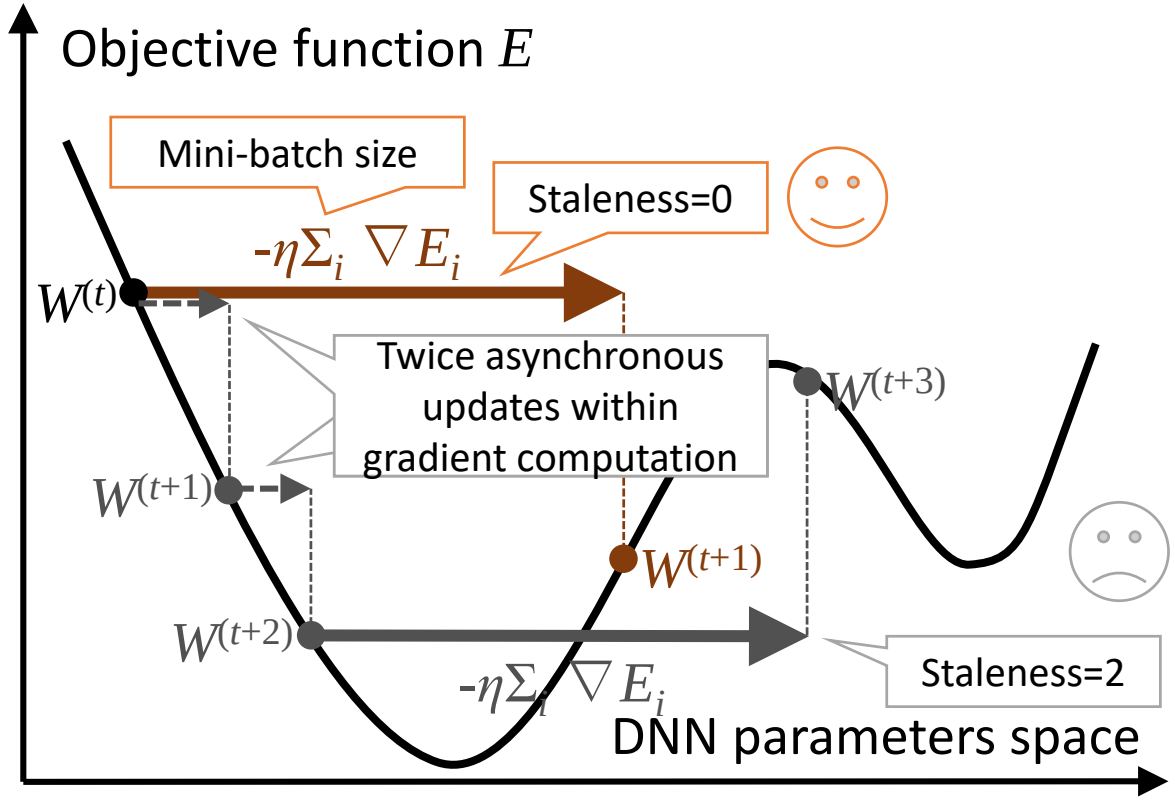


K. Iwabuchi, S. Sallinen, R. Pearce, B. V. Essen, M. Gokhale, and S. Matsuoka, Towards a distributed large-scale dynamic graph data store. In 2016 IEEE Interna- tional Parallel and Distributed Processing Symposium Workshops (IPDPSW)

Predicting Statistics of Asynchronous SGD Parameters for a Large-Scale Distributed Deep Learning System on GPU Supercomputers [BigData16]

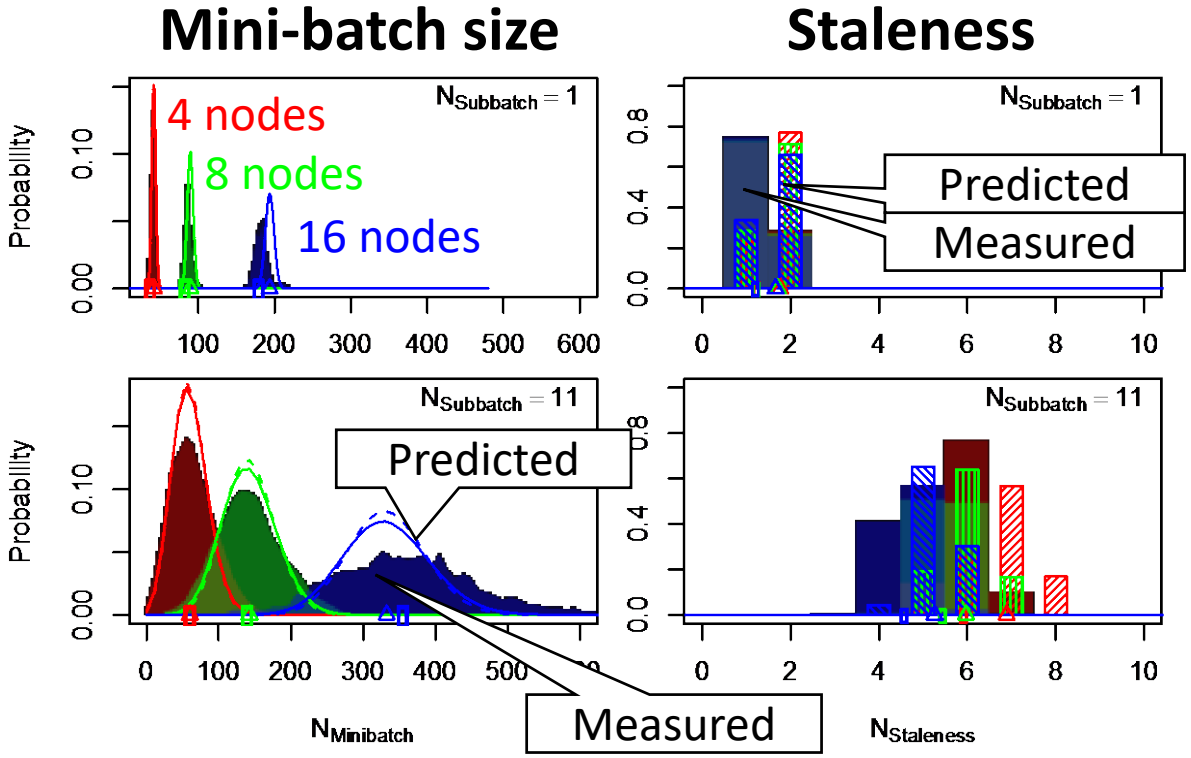
Background

- In large-scale Asynchronous Stochastic Gradient Descent (ASGD), mini-batch size and gradient staleness tend to be large and unpredictable, which increase the error of trained DNN



Proposal

- We propose a empirical performance model for an ASGD deep learning system SPRINT which considers probability distribution of mini-batch size and staleness



($N_{Subbatch}$: # of samples per one GPU iteration)

- Yosuke Oyama, Akihiro Nomura, Ikuro Sato, Hiroki Nishimura, Yukimasa Tamatsu, and Satoshi Matsuoka, "**Predicting Statistics of Asynchronous SGD Parameters for a Large-Scale Distributed Deep Learning System on GPU Supercomputers**", in proceedings of 2016 IEEE International Conference on Big Data (IEEE BigData 2016), Washington D.C., Dec. 5-8, 2016 (to appear)

Performance Prediction of Future HW for CNN

- Predicts the best performance with two future architectural extensions
  - FP16: precision reduction to double the peak floating point performance
  - EDR IB: 4xEDR InfiniBand (100Gbps) upgrade from FDR (56Gbps)

→ Not only # of nodes, but also fast interconnect is important for scalability

TSUBAME-KFC/DL ILSVRC2012 dataset deep learning
Prediction of best parameters (average minibatch size 138±25%)

	N_Node	N_Subbatch	Epoch Time	Average Minibatch Size
(Current HW)	8	8	1779	165.1
FP16	7	22	1462	170.1
EDR IB	12	11	1245	166.6
FP16 + EDR IB	8	15	1128	171.5

Fujitsu Deep Learning Processor (DLU™)

FUJITSU

京 K computer

Supercomputer K technologies

FY2018～

DLU™

(Deep Learning Unit)

FUJITSU DLU™ Deep Learning Unit

DLU™ features

- Architecture designed for Deep Learning
- High performance HBM2 memory
- Low power design

→ Goal: 10x Performance/Watt compared to others

- Massively parallel： Apply supercomputer interconnect technology

→ Ability to handle large scale neural networks

→ TOFU Network derivative for massive scaling

"Exascale" AI possible in 1H2019

Cutting Edge Research AI Infrastructures in Japan Accelerating BD/AI with HPC

(and my effort to design & build them)

Tokyo Tech

1H 2019?
"**ExaAI**"
~1 AI-ExaFlop
Undergoing Engineering Study

Mar. 2018 x5.0~7.7
ABCI (AIST-AIRC)
130-200 AI-PF

Draft RFC out
IDC under construction

Being Manufactured
Aug. 2017 x2.8~4.2
TSUBAME3.0 (Tokyo Tech.)
47.2 AI-PF (65.8 AI-PF w/Tsubame2.5)

Under Acceptance
Mar. 2017 x5.8
AIST AI Cloud
(AIST-AIRC)
8.2 AI-PF

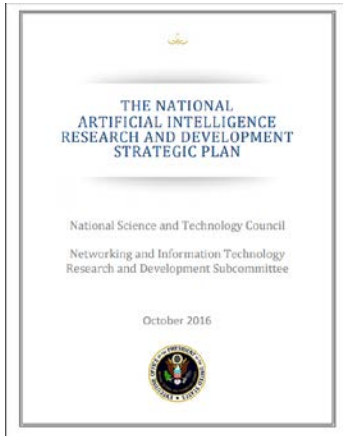
In Production
x5.8

Oct. 2015
TSUBAME-KFC/DL
(Tokyo Tech.)
1.4 AI-PF(Petaflops)

Mar. 2017
AI Supercomputer
Riken AIP
4.1 AI-PF

~x1000 in 3.5 years

R&D Investments into world leading AI/BD HW & SW & Algorithms and their co-design for cutting edge Infrastructure absolutely necessary (just as is with Japan Post-K and US ECP in HPC)


THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN

National Science and Technology Council

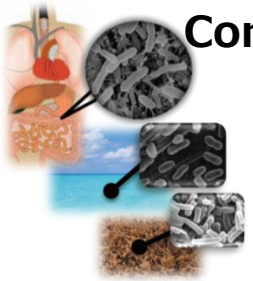
Networking and Information Technology Research and Development Subcommittee

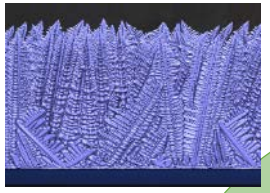
October 2016

Co-Design of BD/ML/AI with HPC using BD/ML/AI - for survival of HPC

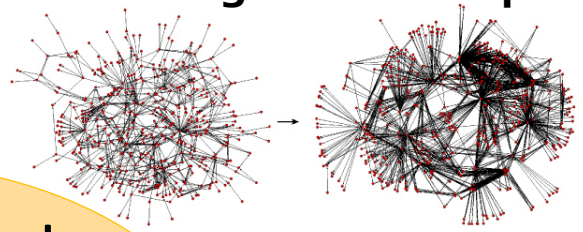
Acceleration and Scaling of BD/ML/AI via HPC and Technologies and Infrastructures

Accelerating Conventional HPC Apps





Large Scale Graphs



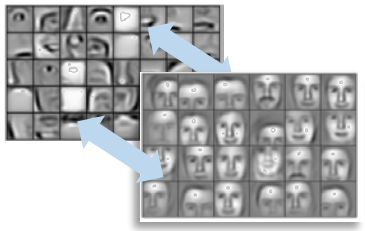
Big Data AI-Oriented Supercomputers

Mutual and Semi-Automated Co-Acceleration of HPC and BD/ML/AI

Big Data and ML/AI Apps and Methodologies

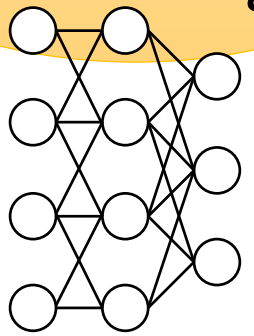
Optimizing System Software and Ops

Image and Video



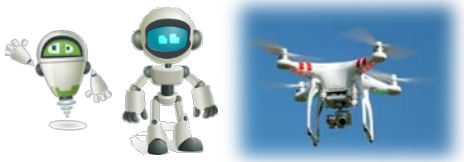
ABCI: World's first and largest open 100 Peta AI-Flops AI Supercomputer, Fall 2017, for co-design

Acceleration Scaling, and Control of HPC via BD/ML/AI and future SC designs



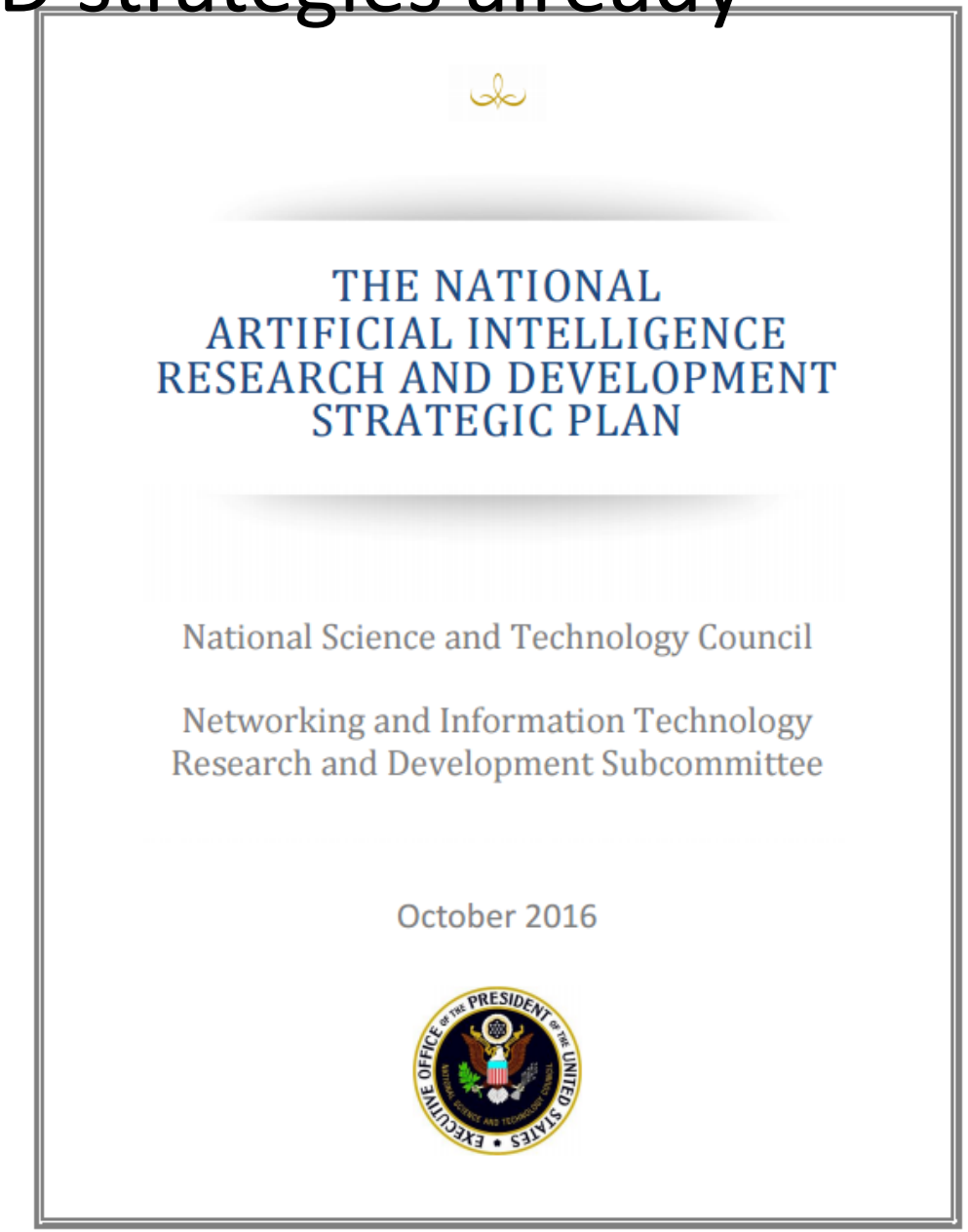
Future Big Data・AI Supercomputer Design

Robots / Drones

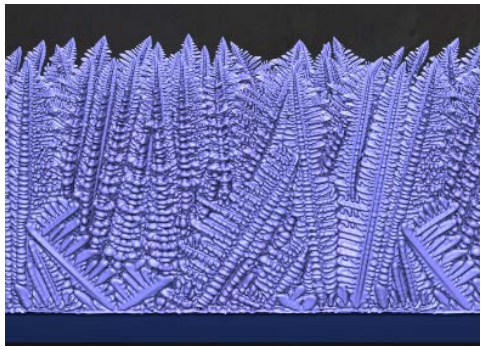


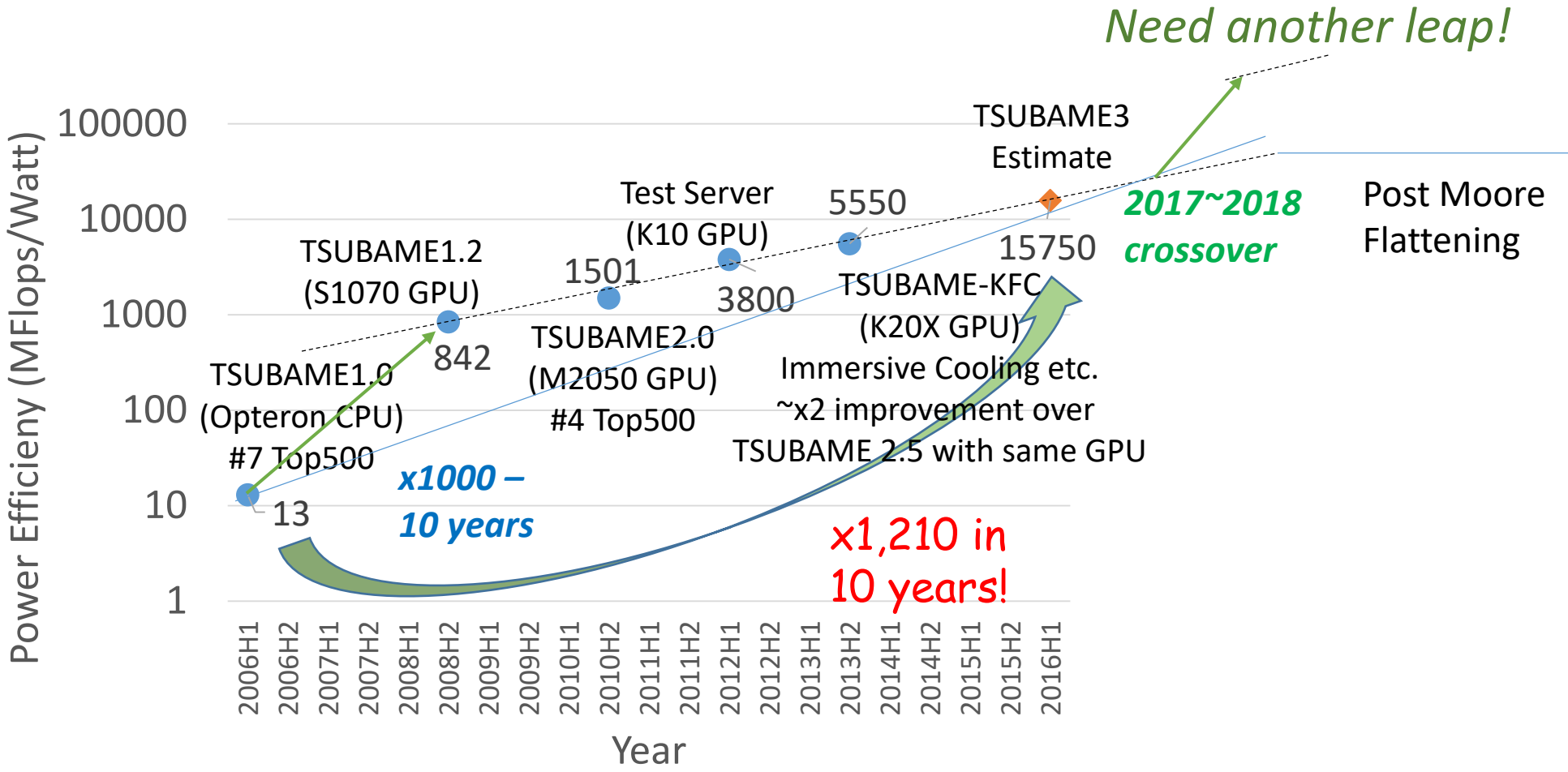
We are implementing the US AI&BD strategies already …in Japan, at AIRC w/ABCI

- Strategy 5: Develop shared public datasets and environments for AI training and testing. The depth, quality, and accuracy of training datasets and resources significantly affect AI performance. Researchers need to develop high quality datasets and environments and enable responsible access to high-quality datasets as well as to testing and training resources.
- Strategy 6: Measure and evaluate AI technologies through standards and benchmarks. Essential to advancements in AI are standards, benchmarks, testbeds, and community engagement that guide and evaluate progress in AI. Additional research is needed to develop a broad spectrum of evaluative techniques.



Many core was a good step but we already used it once, and cannot use it again for boosting





Measured for the 2011 Gordon Bell Award Dendritic Solidification App
Flop/s/W = Total #Flops / J = energy to solution given same problem

What is worse: Moore's Law will end in the 2020's

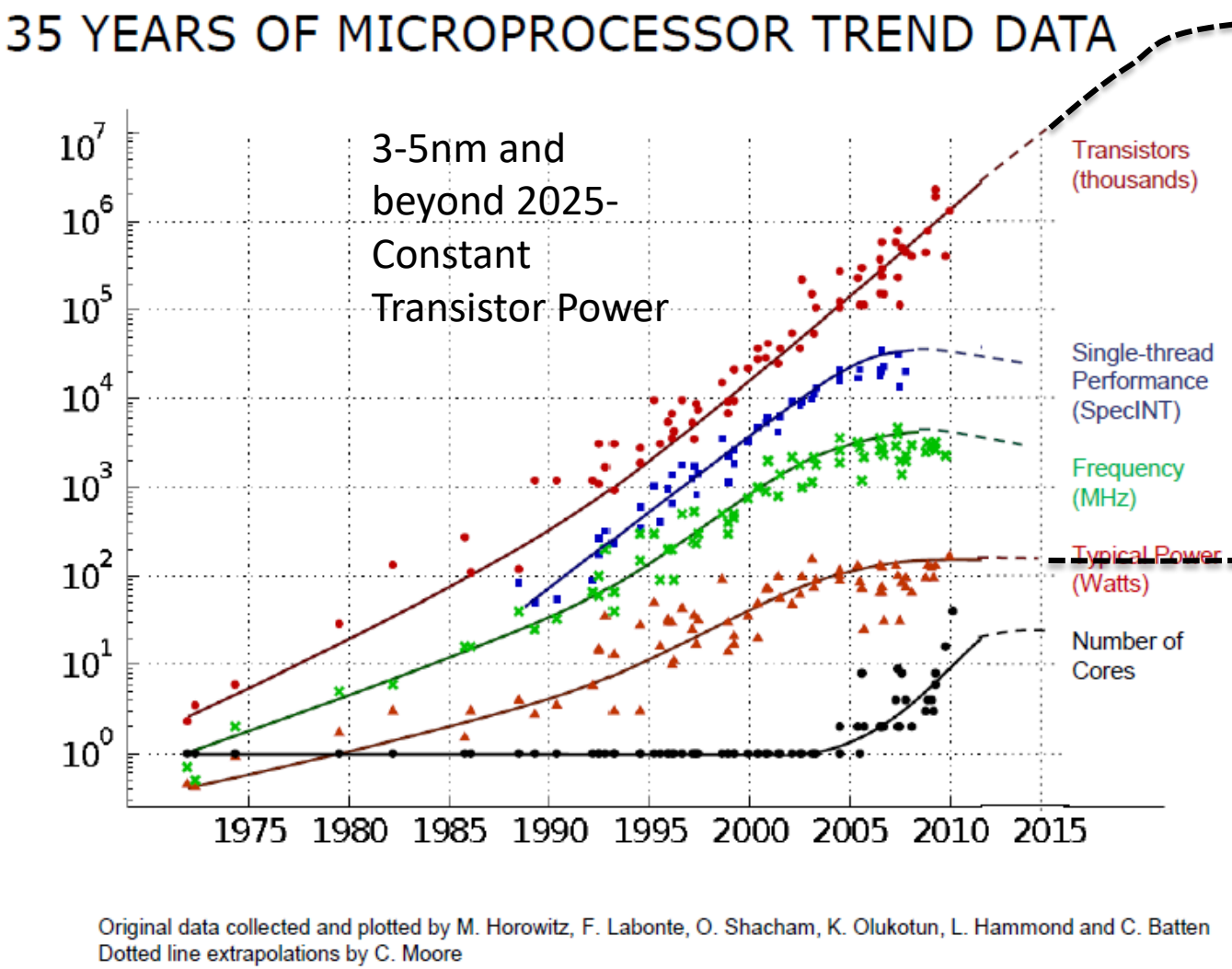
- Much of underlying IT performance growth due to Moore's law
  - "LSI: x2 transistors in 1~1.5 years"
  - Causing qualitative "leaps" in IT and societal innovations
  - The main reason we have supercomputers and Google…
- But this is slowing down & ending, by mid 2020s…!!!
  - End of Lithography shrinks
  - End of Dennard scaling
  - End of Fab Economics

Gordon Moore

The curse of constant transistor power shall soon be upon us

- How do we *sustain* "performance growth" beyond the "end of Moore"?
  - Not just one-time speed bumps
  - ***Will affect all aspects of IT, including BD/AI/ML/IoT, not just HPC***
  - ***End of IT as we know it***

20 year Eras towards of End of Moore's Law



20-year Moore-Dennard Single Core ILP-Vector Killer-Micro Era

- 1980s~2004 Dennard scaling, perf+ = single thread+ = transistor & freq+ = power+

20 year Post-Dennard Many-Core Era

- 2004~2015 feature scaling, perf+ = transistor+ = core#+, constant power
- 2015~2025 all above gets harder

20-year Next-Gen Post-Moore era

- 2025~ post-Moore, **constant feature&power = flat performance**

Need to realize the next 20-year era of supercomputing

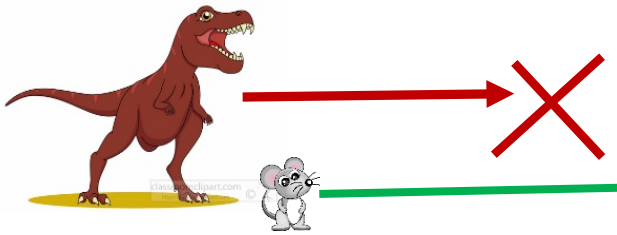
The “curse of constant transistor power”
- Ignorance of this is like ignoring global warming -

- Systems people have been telling the algorithm people that “FLOPS will be free, bandwidth is important, so devise algorithms under that assumption”
- This will certainly be true until exascale in 2020…
- But when Moore’s Law ends in 2025-2030, constant transistor power (esp. for logic) = FLOPS will no longer be free!
- So algorithms that simply increase arithmetic intensity will no longer scale beyond that point
- Like countering global warming – need disruptive change in computing – in HW-SW-Alg-Apps etc. for the next 20 year era

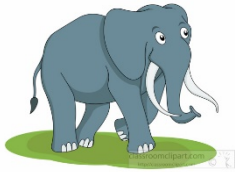
Performance growth via ***data-centric computing:***
"From FLOPS to BYTES"

- *Identify the new parameter(s) for scaling over time*
- Because data-related parameters (e.g. capacity and bandwidth) *will still likely continue to grow towards 2040s*
- Can grow transistor# for compute, but CANNOT use them AT THE SAME TIME(Dark Silicon) => **multiple computing units specialized to type of data**
- **Continued capacity growth**: 3D stacking (esp. direct silicon layering) and low power NVM (e.g. ReRAM)
- **Continued BW growth:** Data movement energy will be **capped constant** by dense 3D design and advanced optics from silicon photonics technologies
- Almost back to the old "vector" days(?), but no free lunch – latency still problem, locality still important, *need* ***general algorithmic acceleration thru data capacity and bandwidth,*** *not FLOPS*

Many Core Era



Post Moore Era

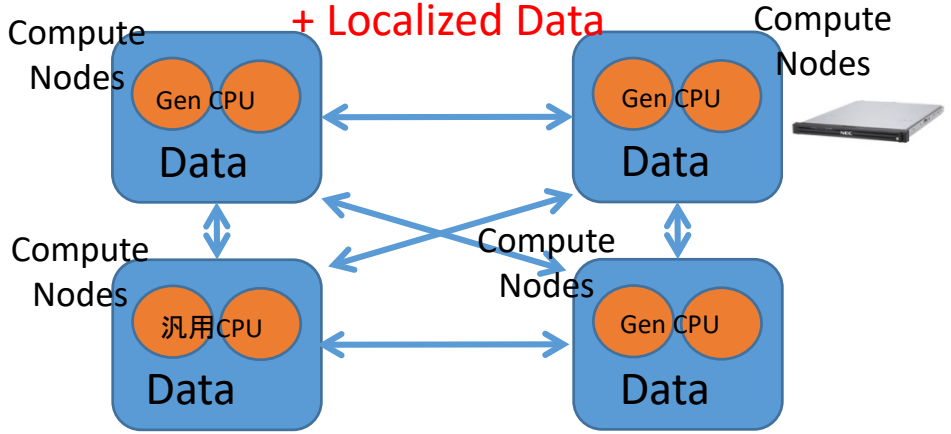


Flops-Centric Algorithms and Apps

Flops-Centric System Software

Hardware/Software System APIs
Flops-Centric Massively Parallel Architecture

Homogeneous General Purpose Nodes
+ Localized Data



Loosely Coupled with Electronic Interconnect

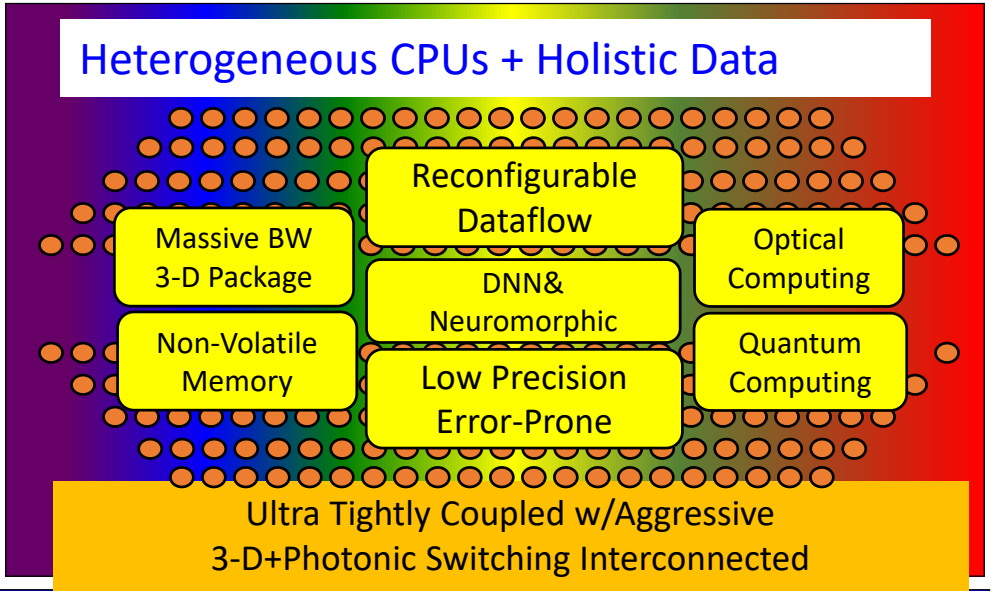
Transistor Lithography Scaling
(CMOS Logic Circuits, DRAM/SRAM)

~2025
M-P Extinction Event

Bytes-Centric Algorithms and Apps

Bytes-Centric System Software

Hardware/Software System APIs
Data-Centric Heterogeneous Architecture

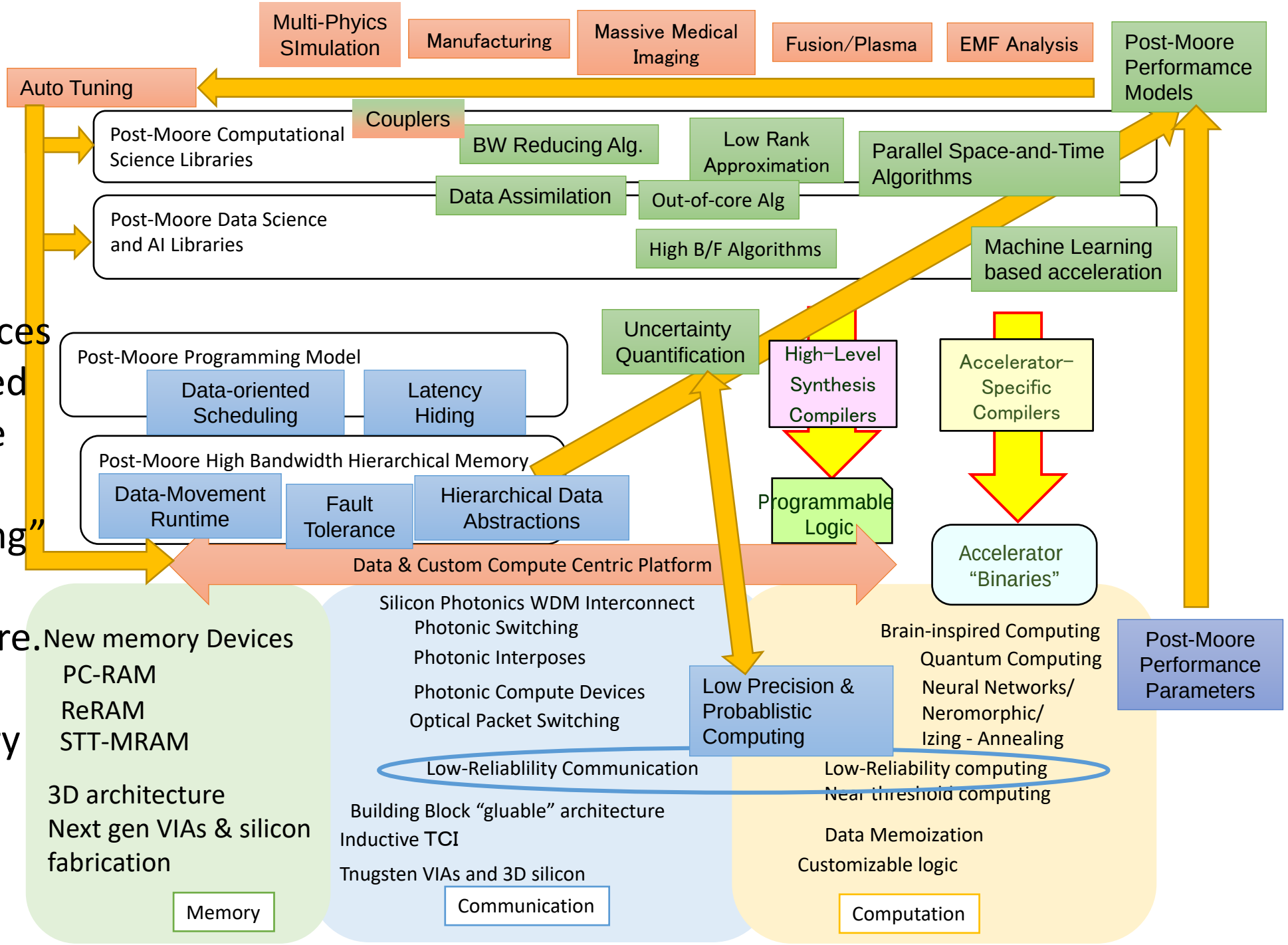


Novel Devices + CMOS (Dark Silicon)
(Nanophotonics, Non-Volatile Devices etc.)

Post-Moore is NOT a More-Moore device as a panacea

Device & arch. advances improving data-related parameters over time

"Rebooting Computing" in terms of devices, architectures, software. Algorithms, and applications necessary
=> Co-Design even more important
c.f. Exascale



Problem Specific Architectures to exploit dark silicone

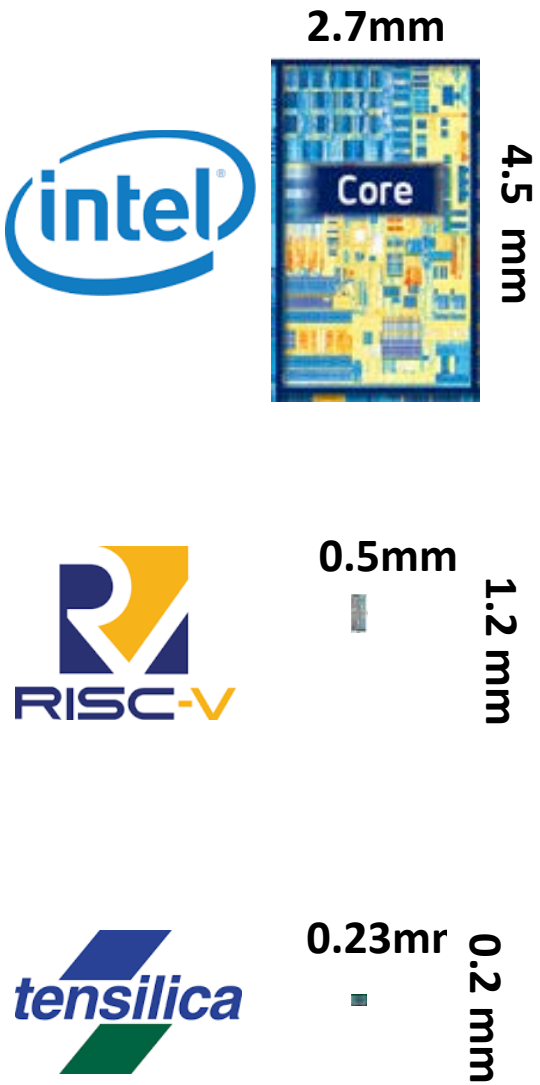
- Deep Neural Network Accelerator (Many, incl. Google)
- Spiking Neuromorphic Architecture (Manchester SpiNNaker, IBM TrueNorth, Heidelberg BrainScaleS)
- Ising Model optimization architecture (Hitachi)
- Automata Processor (Micron)
- Advanced FPGAs (Alltera, Xilinx)
- Network & I/O accelerator (Mellanox)
- …
- And of course Quantum Annealing and Computing (D-Wave)

Non-Volatile Memory and 3-D Stacking

- Many devices
- Various stacking technologies

- Results: Massive capacity, extreme bandwidth, low power
- Exploits Z-direction locality
- New breed of “in memory computing”
- Could persist as a trajectory for the next 20 years

When does data movement dominate? (Original Slide Courtecy John Shalf@LBNL)



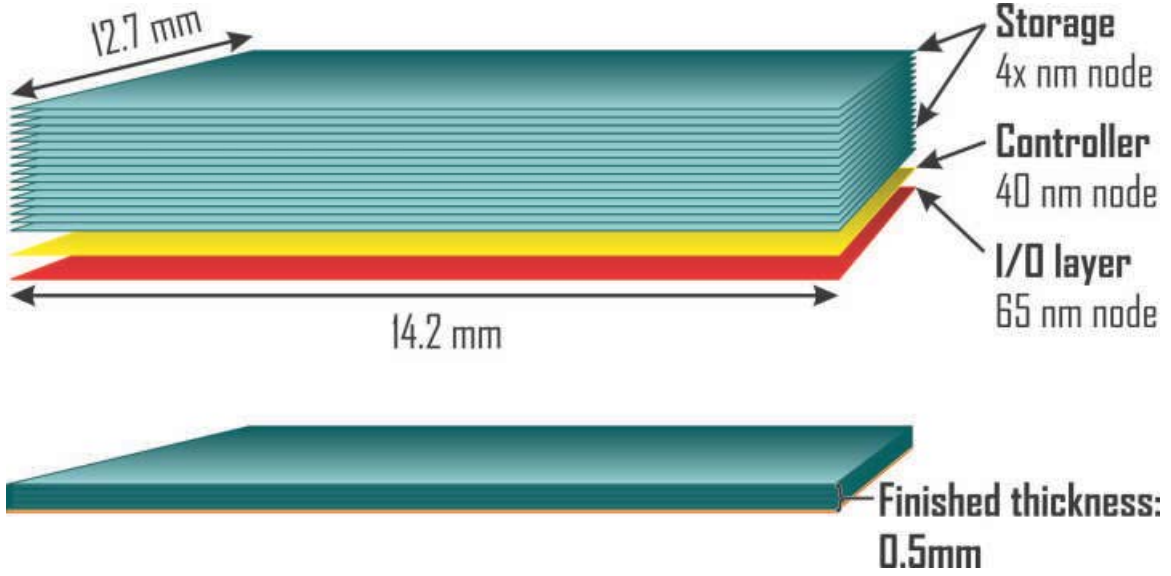
Core Energy/Area est.	Data Movement Cost
Area: 12.25 mm^2 Power: 2.5W Clock: 2.4 GHz E/op: 651 pj	*Compute Op == data movement Energy @* **108mm** *Energy Ratio for 20mm* **0.2x**
Area: 0.6 mm^2 Power: 0.3W (<0.2W) Clock: 1.3 GHz E/op: 150 (75) pj	*Compute Op == data movement Energy @* **12mm** *Energy Ratio for 20mm* **1.6x**
Area: 0.046 mm^2 Power: 0.025W Clock: 1.0 GHz E/op: 22 pj	*Compute Op == data movement Energy @* **3.6mm** *Energy Ratio for 20mm* **5.5x**

Could be reduced by orders of magnitude by 3D, as Z-direction movement is under 1mm

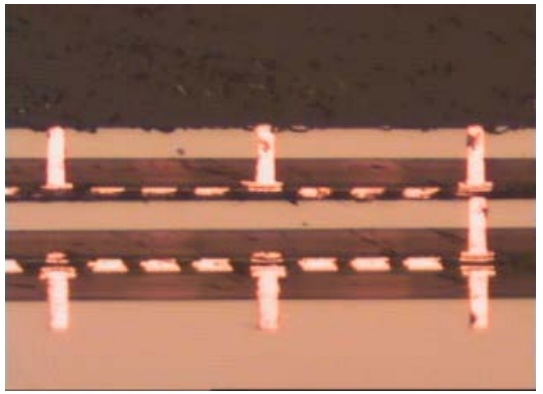
Capacity by dense NVM w/DRAM cache

Example Innovation: Tungsten TSV at 2um ultra fine pitch with die thinning by Tezzaron Semiconductor

- Suppose 4TF SFP @ 7nm, 16TB/s internal chip BW vs. 200GB/s external chip mem BW => 80 times speedup!
- High-density, high-signaling TSV challenge
  - Wide I/O 2 1024 bits 1 Ghz -> 2~3 Ghz
  - We need 128,000 bits @ 1Ghz !
  - 10 micron TSV estimation
    - 400 x 400 TSVs on 20mx20m chip -> 50 micron spacing
    - With tungsten TSVs the chip area is negligible

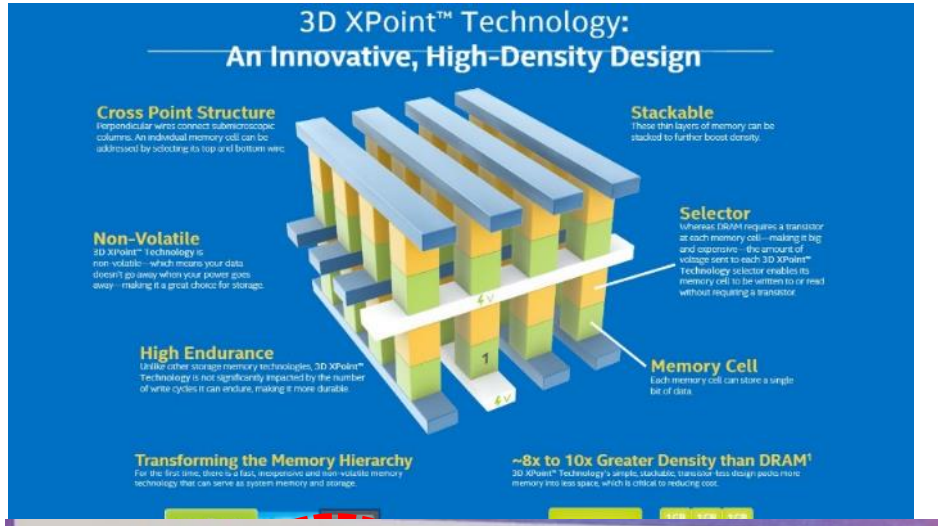


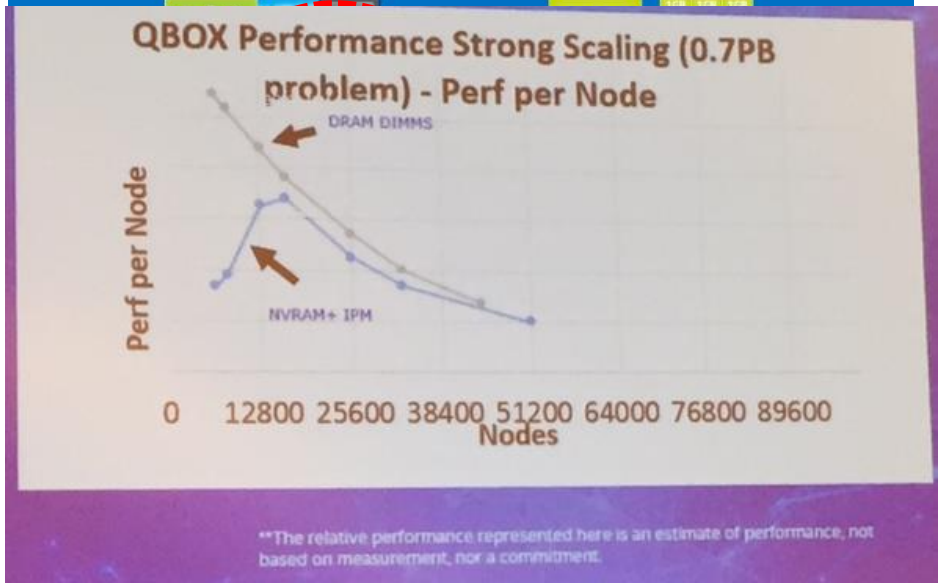
Many-layer stacking via aggressive wafer thinning and self-diagnostics



Source: Tezzaron website
http://www.tezzaron.com

Accelerating “Big Data” Non-Volatile Memory-based Exascale Architectures





From Al Gara Keynote, IEEE Custer 2015

- Accelerating
  - HPC Apps, Big Data-HPC Apps
- Using
  - Exascale machines w/NVM (Flash, ReRAM, 3D Xpoint, …)
  - BYTES, not FLOPS
- While
  - Reducing Bandwidth, Exploiting Locality
  - Dealing with higher write cost
  - Dealing with low durability
  - Maintaining Programmability
  - Exploiting other system assets such as hybrid electro-optical Exabit interconnect
  - …

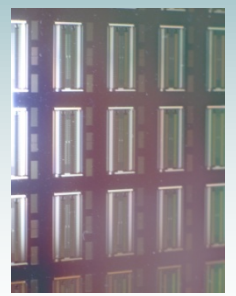
DiRAM4 Stack Overview

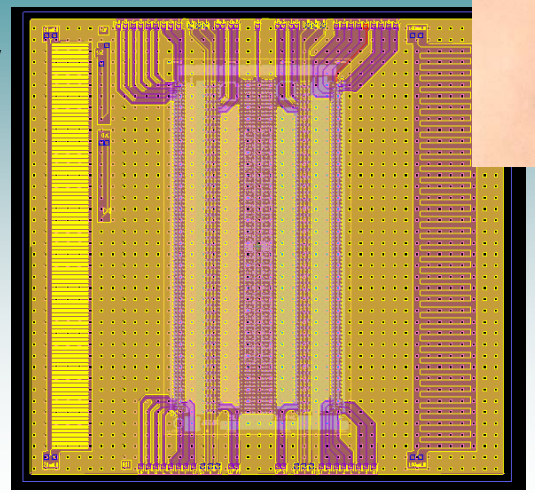
(Tezzaron slides taken from http://www.tezzaron.com/media/Tezzaron-Presentation-EPS-100814-dist-.pptx)

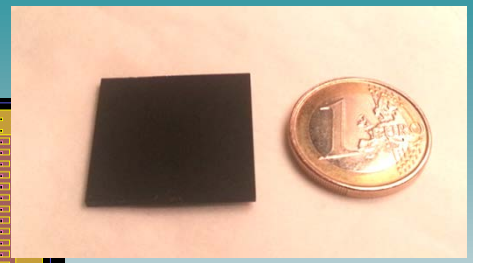
- **64 Gb** of Memory in 175 mm^2
- **256** fully independent RAMs
- **16 Banks** per RAM
- **64 bit** Sep I/O Data per RAM
- 7ns Access Time (Closed page to data)
- 12ns tRC (Page Open to Page Open in a Bank)
- 16 Tb/s Data Bandwidth
- Competitive Manufacturing Cost

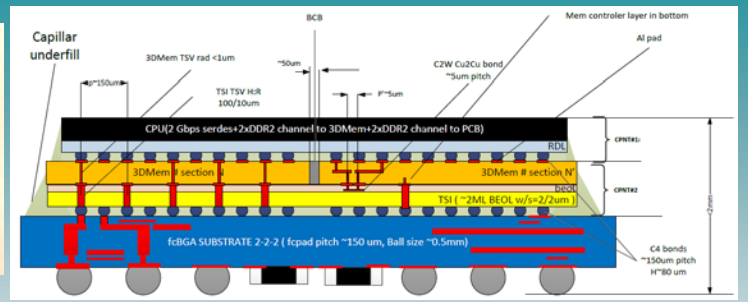
2.5/3D Circuits

IME A*STAR / Tezzaron Collaboration

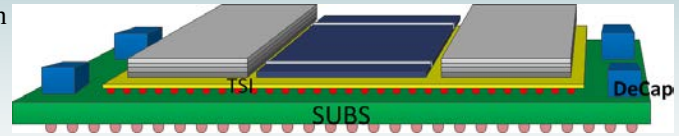


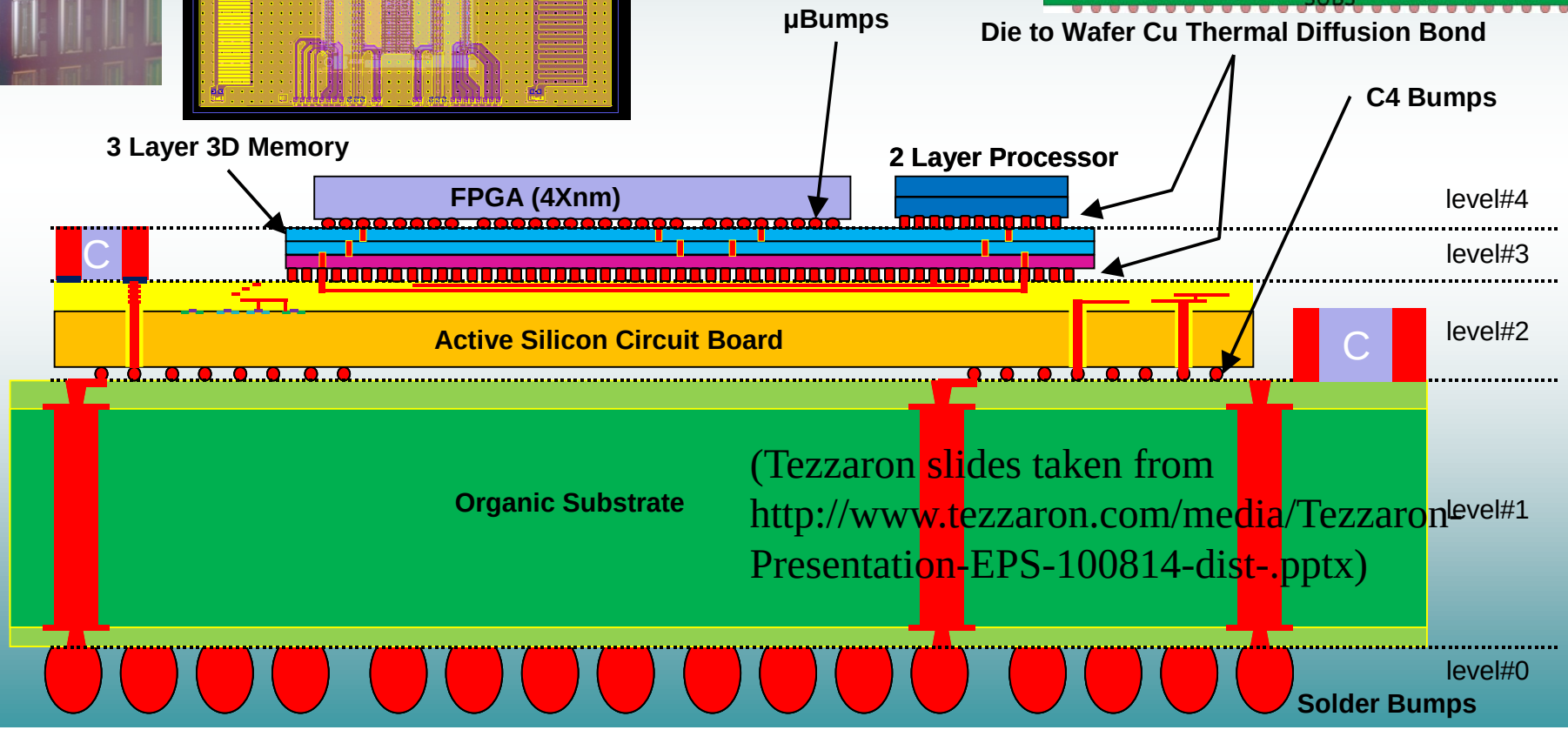






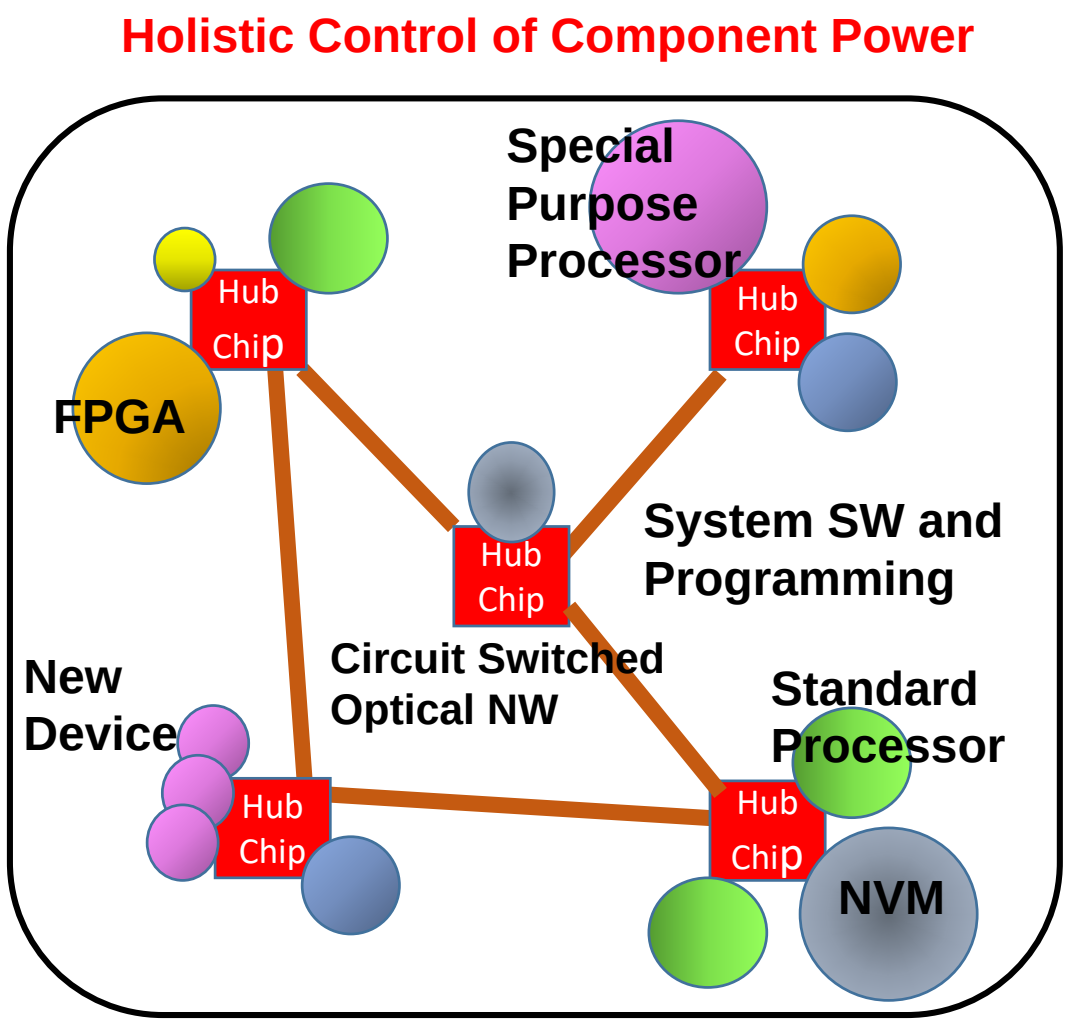
IME A*STAR / Tezzaron Collaboration





(Tezzaron slides taken from http://www.tezzaron.com/media/Tezzaron Presentation-EPS-100814-dist-.pptx)

Super Building Block Archiecture (Amano, Keio U)

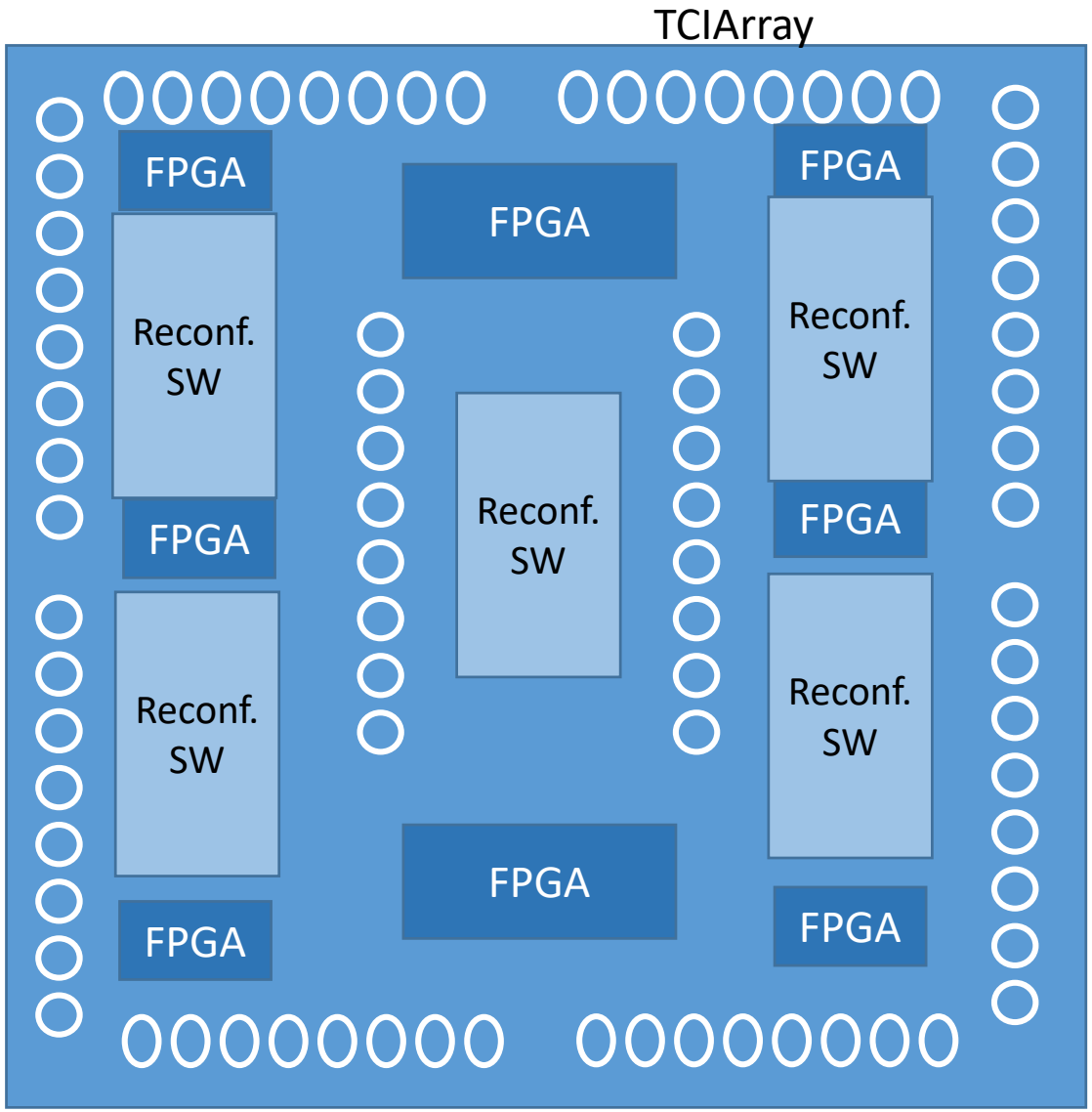


System View on "Post-Moore" Architecture

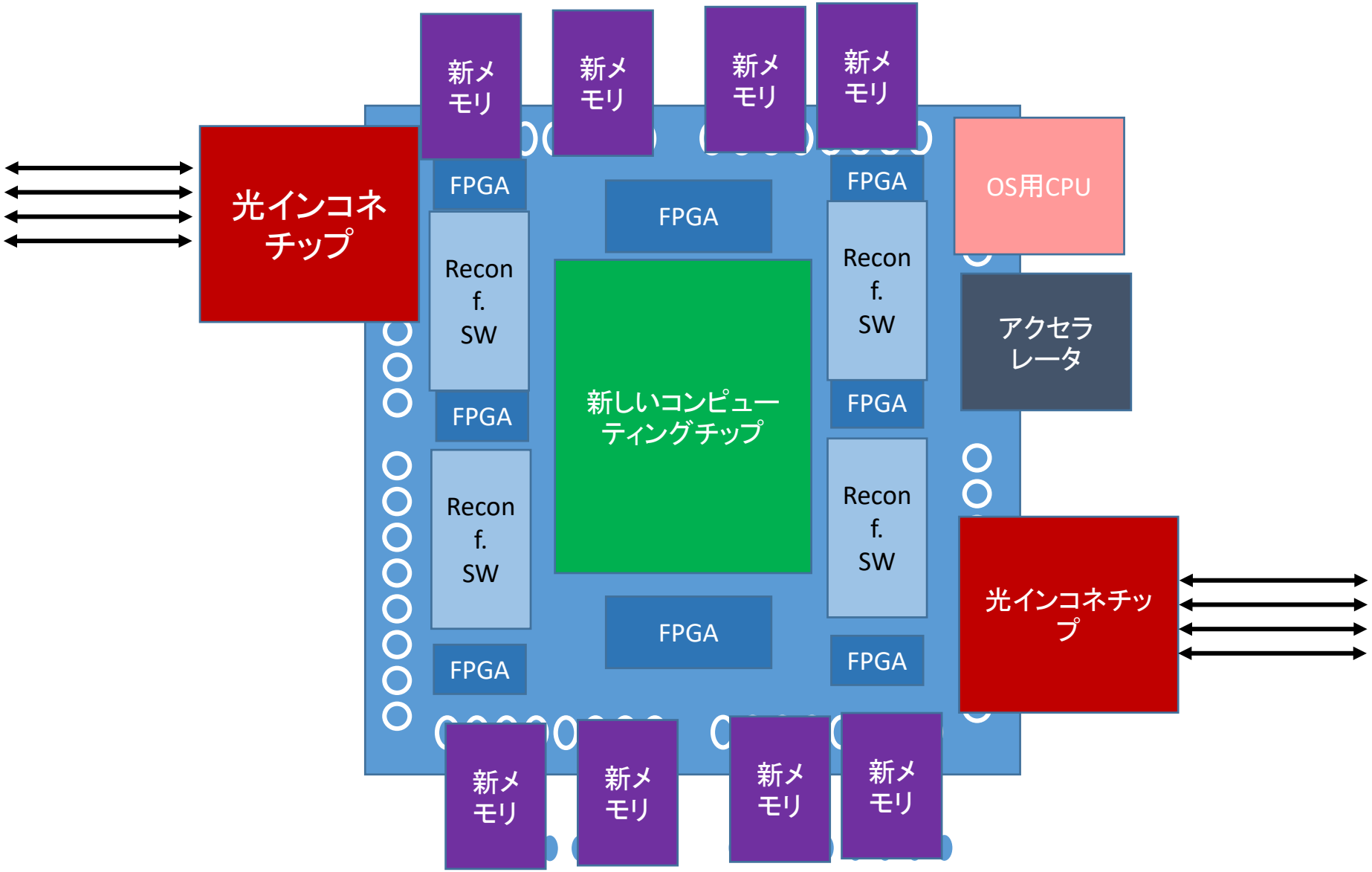
Not just a new device, but focus on how they are interconnected, and integrated as a system controlling their power

A Hub architecture that employs Inductive (3D) TCI and programmable FPGA+Switch

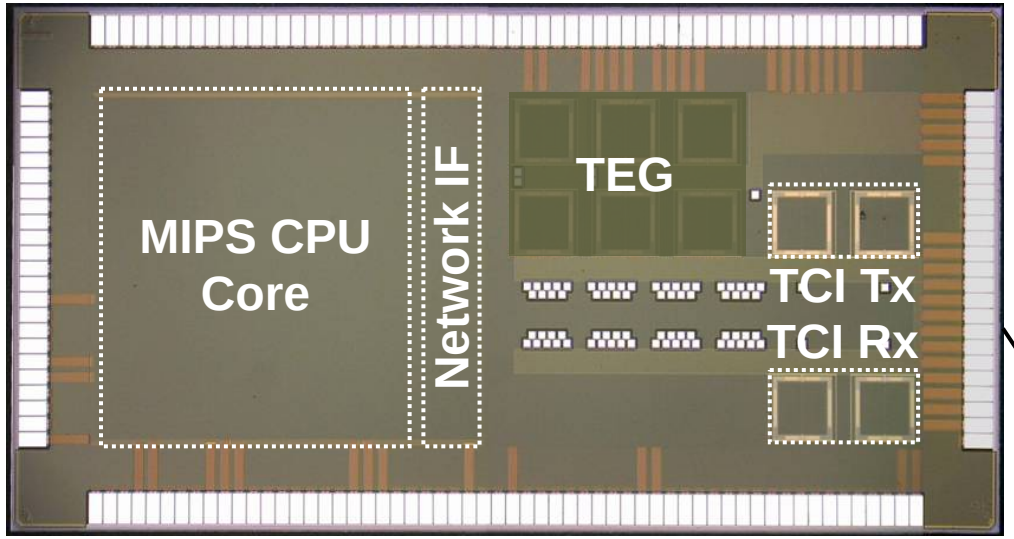
Hub Base Chip = Inductive TCI + Reconfigurable Swtich + FPGA



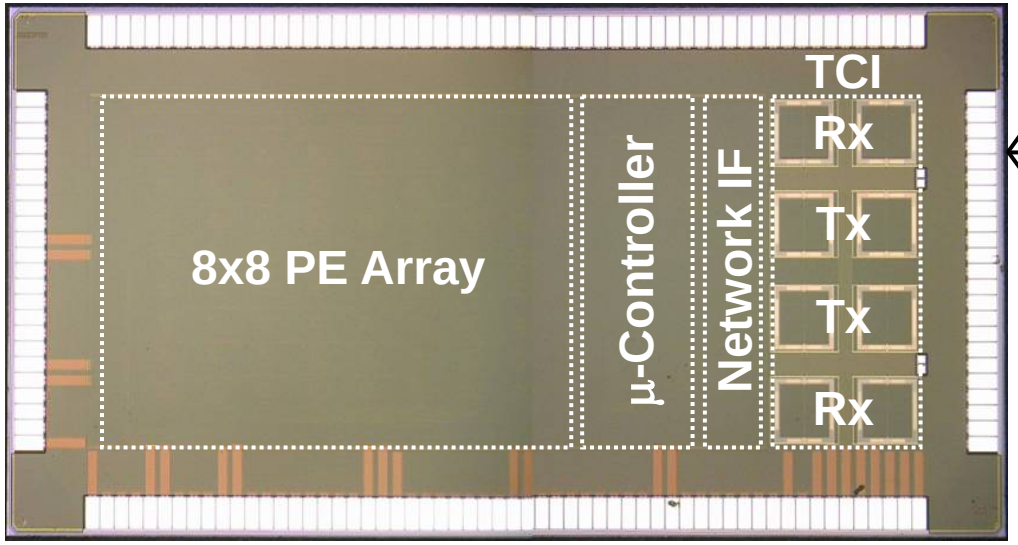
A new reconfig device based on TCI arrays and reconfigurable switches



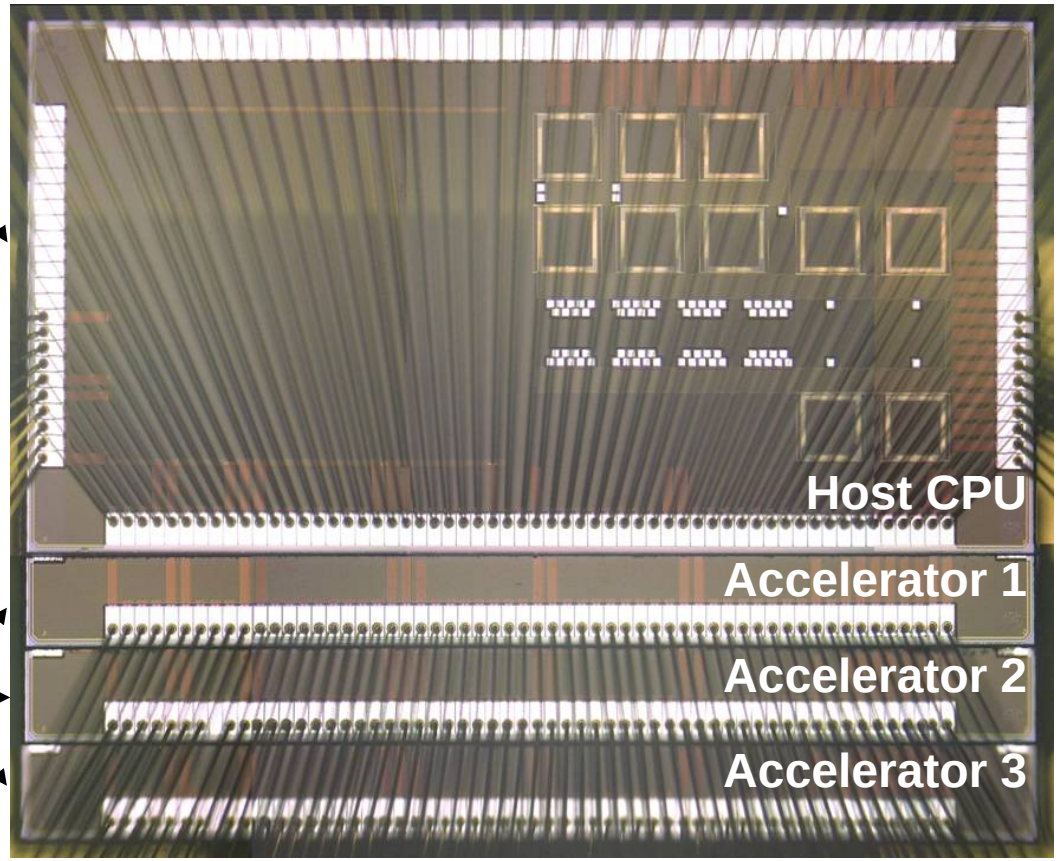
ドーターチップ接続のイメージ



Host CPU Chip



Accelerator Chip



Host CPU + Accelerator x3 Chip Stack
Fabricated in 65nm CMOS

Microphotograph of stacked test chips.

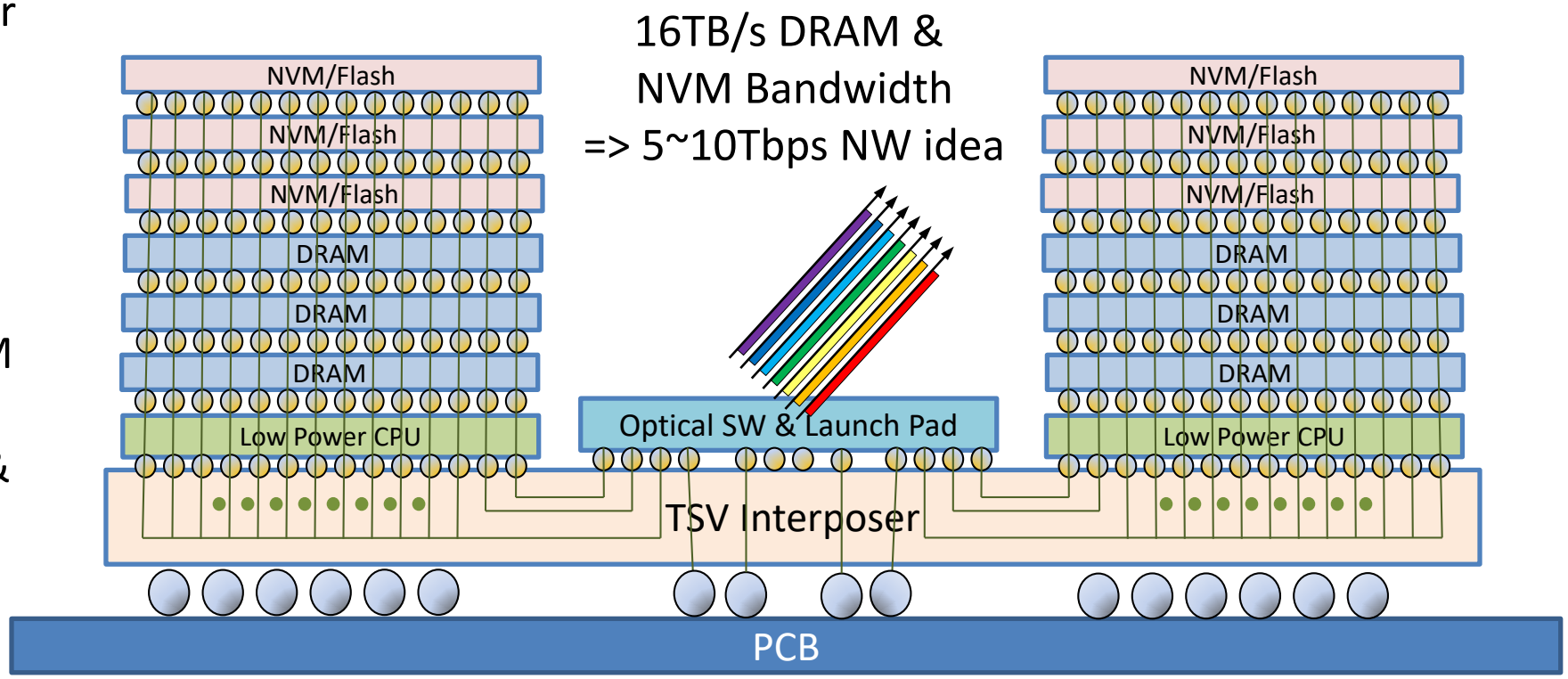
Strawman BYTES-Oriented Post-Moore Architecture

Low voltage & power CPU for direct stacking and large silicon area

Domain-specific hetero- and customizable processor configurations, including PIM

Extreme multi-layer DRAM & NVRAM stacking via high density tungsten TSV

Direct WDM optics onto Interposer

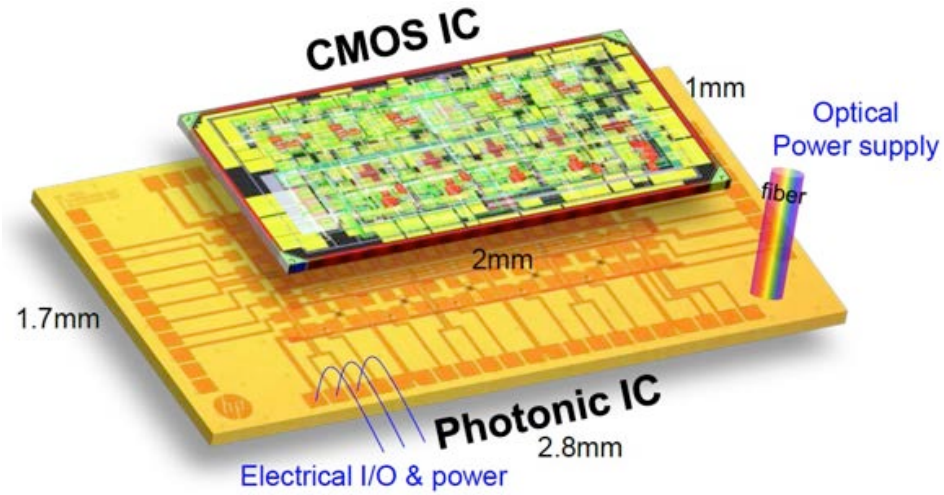


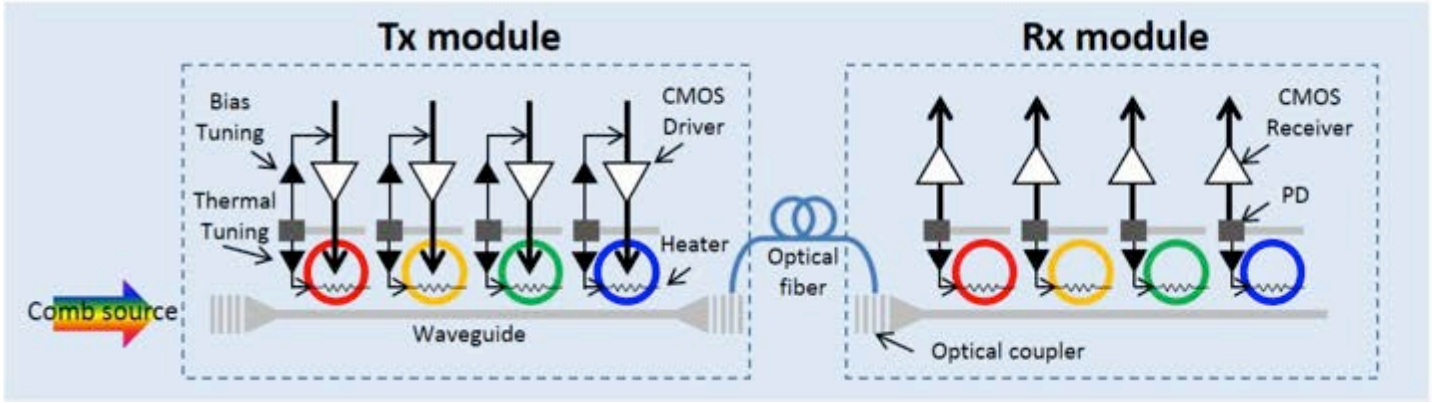
Direct Chip-Chip Interconnect with DWDM optics
Low Power Processor allows Direct 3D Stacking
Configurable Low-power CPU

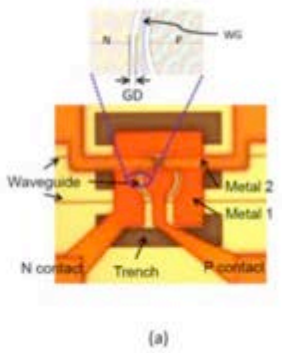
Making Silicon Photonics a reality (slide courtesy of Nick Dube @ HPE)

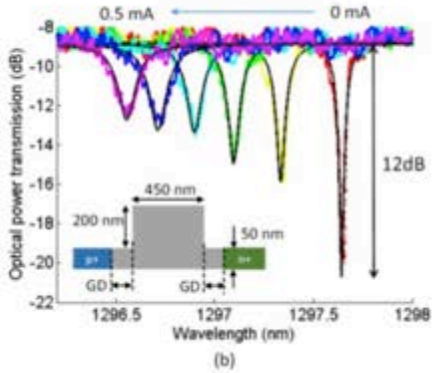
- Today's optical technology:
  - 400$ / 100Gbps optical cable @ ~4 watts
  - Or 4$ per Gbps or 40 pJ / bit per second
  - At Exascale: >10MW and >500M$ (this is problematic)

- The Silicon Photonics promise:
  - 10 cents per Gbps, projected 3-6 pJ/bit









Hybrid Integrated DWDM Silicon Photonic Transceiver with Self-Adaptive CMOS Circuits
Chen, Li, Bai, Shafik, Fiorentino, Peng, Chiang, Palermo, Beausoleil

Optical Network Technology for Future Supercomputers & IDC

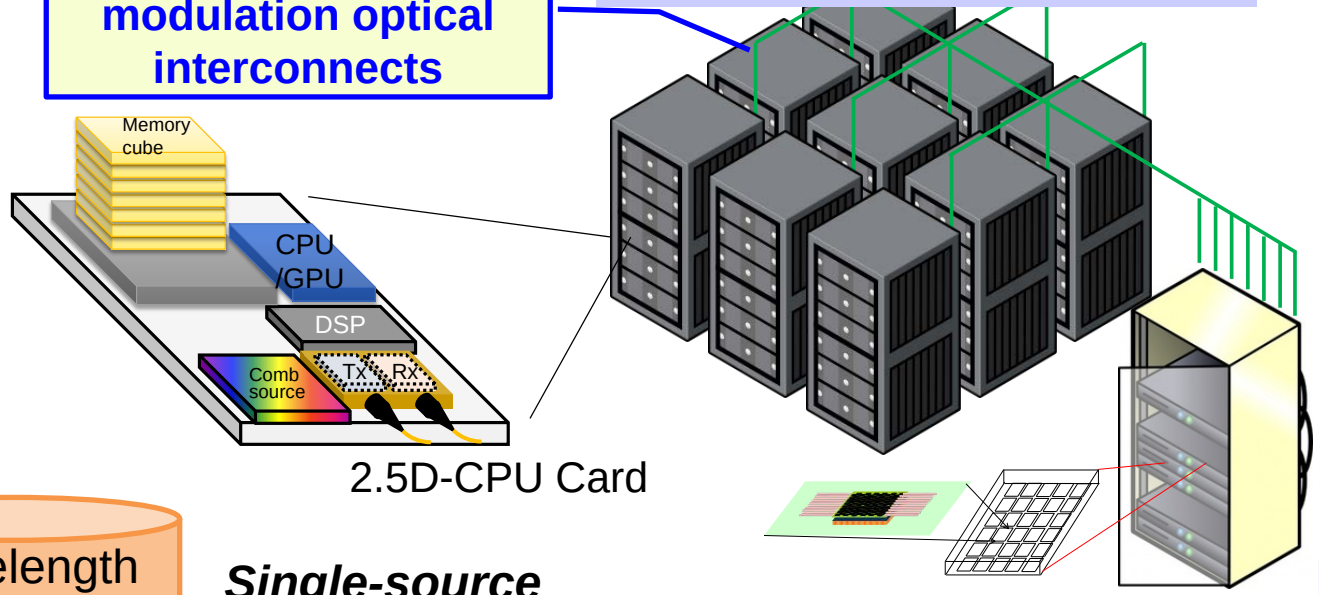
AIST IMPULSE

- Large-scale silicon photonics based cluster switches
- DWDM, multi-level modulation, highly integrated "elastic" optical interconnects
- Ultra-low energy consumption network by making use of optical switches

SC14 AIST Booth #2531

DWDM, multi-level modulation optical interconnects

Datacenter server racks



2.5D-CPU Card

Silicon photonics cluster switches

- Ultra-compact switches based on silicon photonics
- 3D integration by amorphous silicon
- A new server architecture

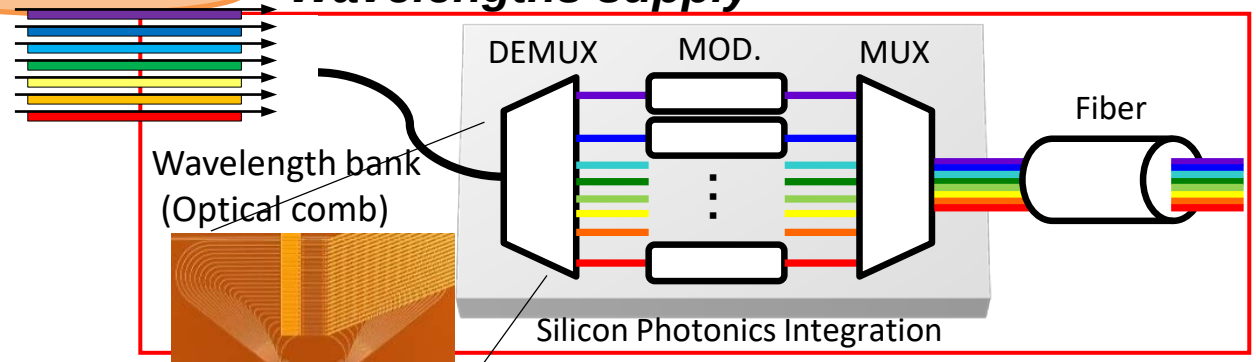
Current electrical switches:
~1pbps
↓
~500Pbps

No of λs	Order of mod.	Bit rate
1	1	20 Gbps
4	8	640 Gbps
32	8	5.12 Tbps

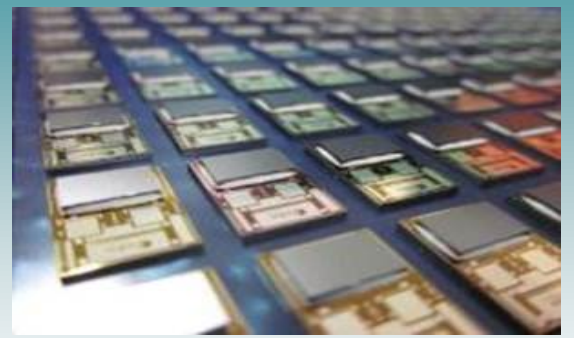
Current state-of-the-art Tx
100Gbps
↓
~ 5.12Tbps

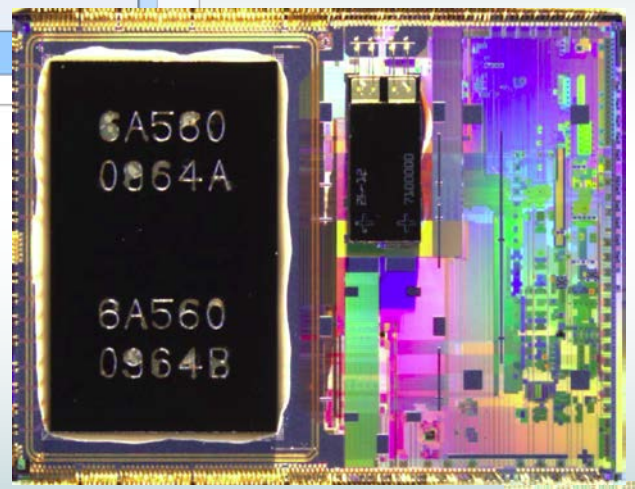
Wavelength Bank

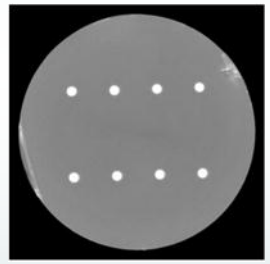
Single-source Wavelengths supply

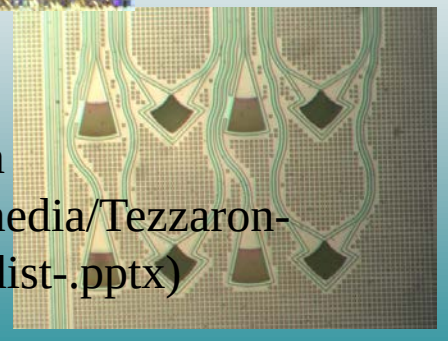


Luxtera 2.5D Photonic Data Pump









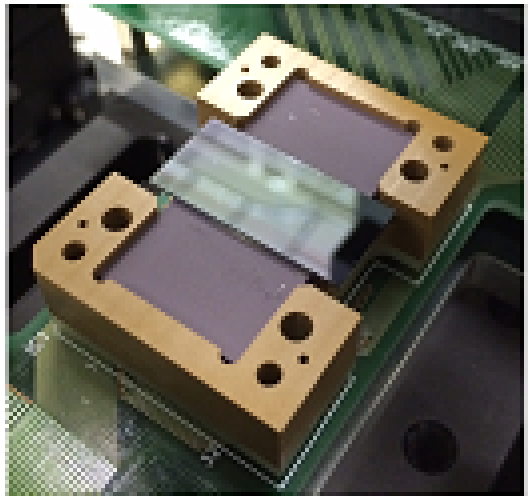
- 2.5pJ/bit power
- Bare metal protocol
  - Ultra low latency
  - Protocol agnostic
- 8 core Fiber
- 25Gb SERDES or 3.125Gb interface
- Self-calibrating self-tuning
- >1.6Tb/s payload

(Tezzaron slides taken from http://www.tezzaron.com/media/Tezzaron-Presentation-EPS-100814-dist-.pptx)

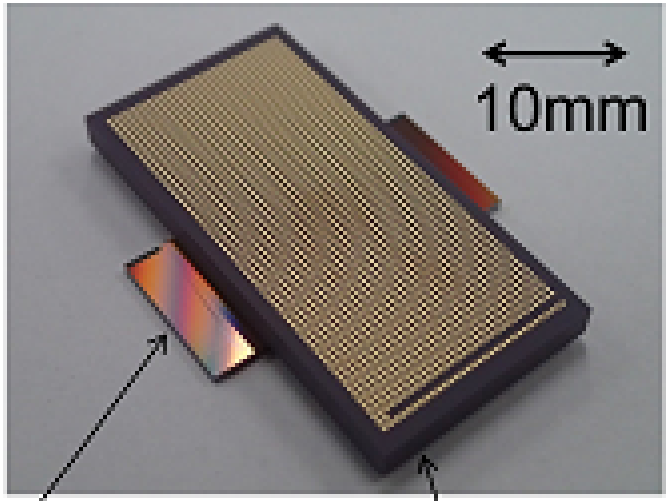
32 x 32 Optical Circuit Switch (Courtesy NTTAIST)

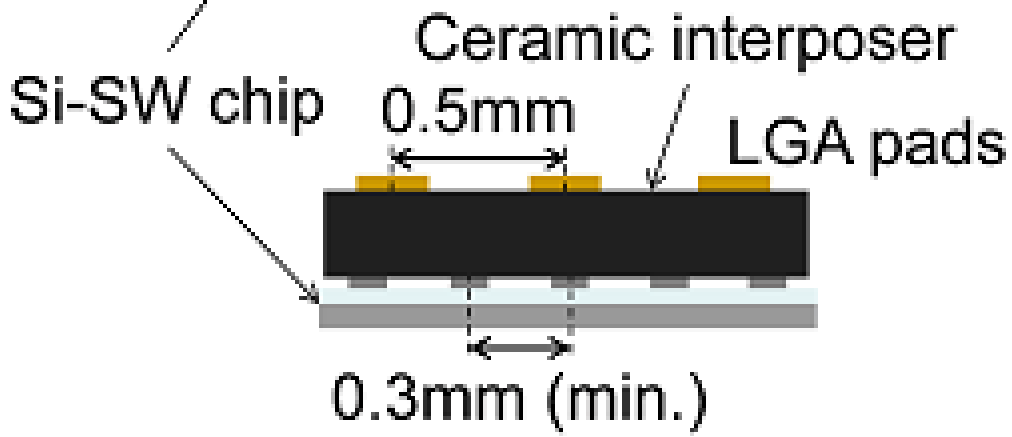
- Ceramic LGA interposer with 0.5-mm pitch
- Flip-chip bonding with Au bumps and non-conductive paste
- LGA socket to contact PCB

LGA socket



After FC bonding





Problem: heavy optical loss

Fast Optical Crossbar Swtch (EECS, UCB)

Seok et. al. "Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers" Optica, 3-1, 2016

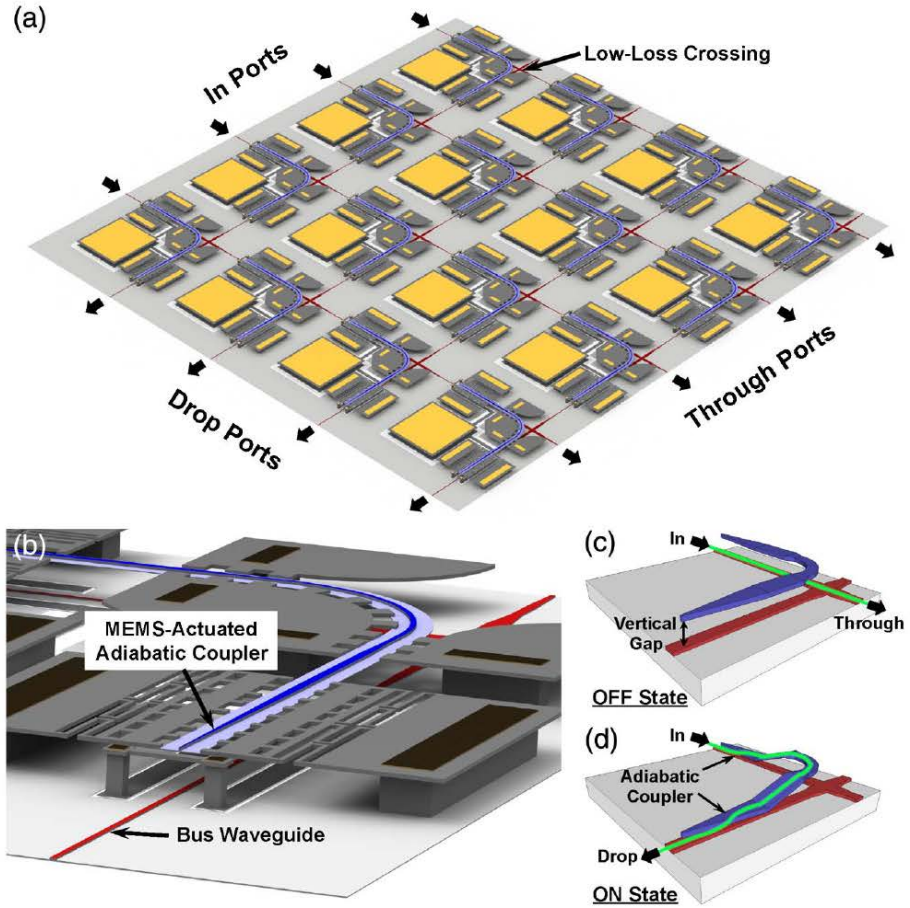


Fig. 1. Schematics of silicon photonic MEMS switches. (a) Matrix architecture of silicon photonic MEMS switch, (b) close-up view of a MEMS-actuated adiabatic coupler, (c) switch unit cell in the OFF state, and (d) switch unit cell in the ON state. The adiabatic coupler is precisely positioned at the optimum distance to the bus waveguide.

- Array of 64x64 MEMS optical crossbar switch
- 3.7db on-chip insertion loss
- 0.91microsecond switching time

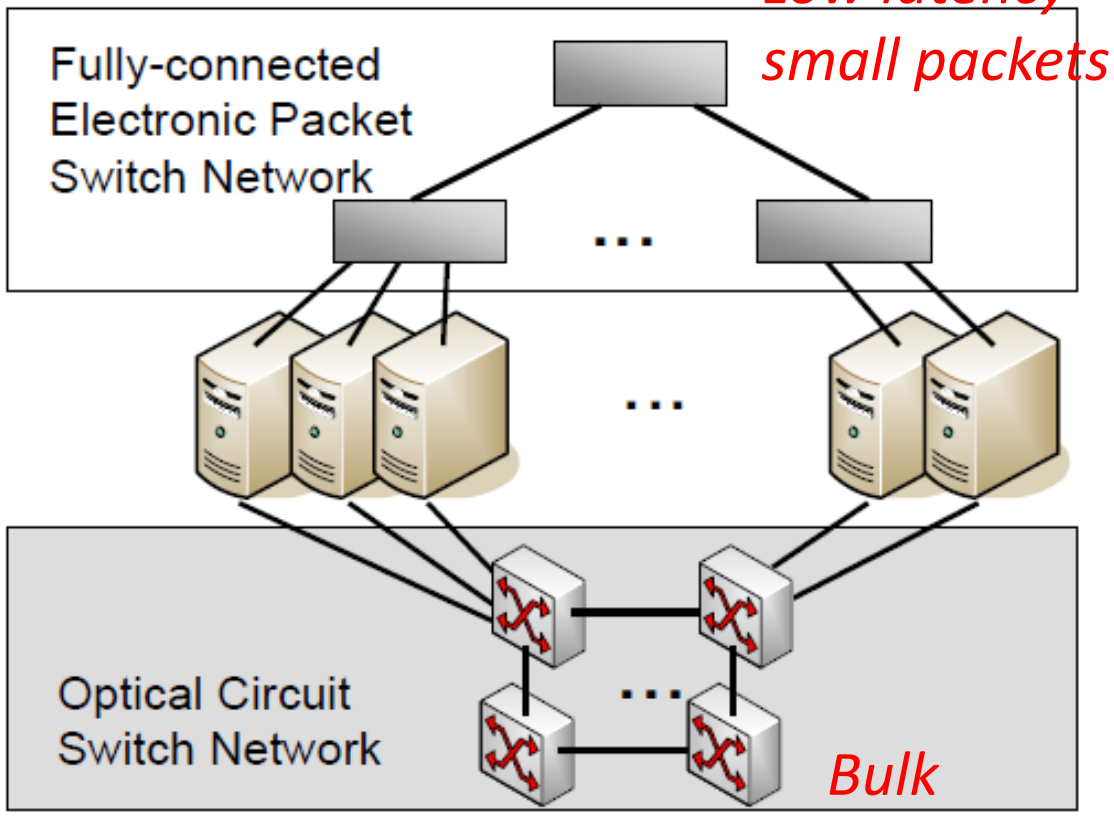
- At 100,000 ports – 9 hop network
  - 33db+ loss
  - 8.2 microsecond switching time => 1Tb 800Kbyte BW x Delay

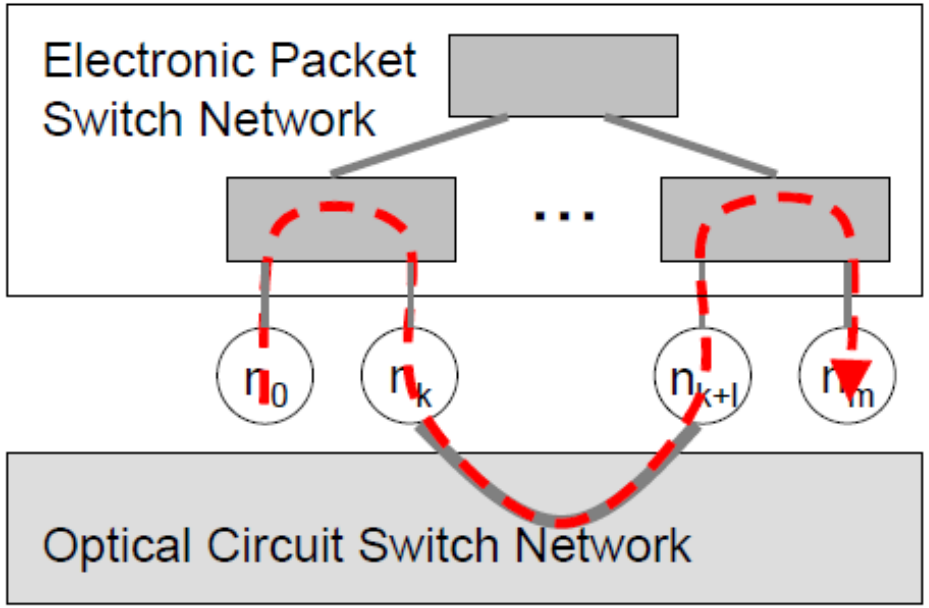
Solution: Hybrid EO Network

- Idea1: use low (latency/diameter, bandwidth, power) electrical network for low latency messages, and use optical circuits for high bandwidth and fixed topology messages
- Idea2: merge the electronic switch and optical MEMS switch, and use the latter as the control plane of the optical MEMs circuit
  - Thus the electronic switches become the optical speculative "buffer"

Hybrid Electro-Optical Network w/shortcuts [Takizawa&Matsuoka LSPP07]

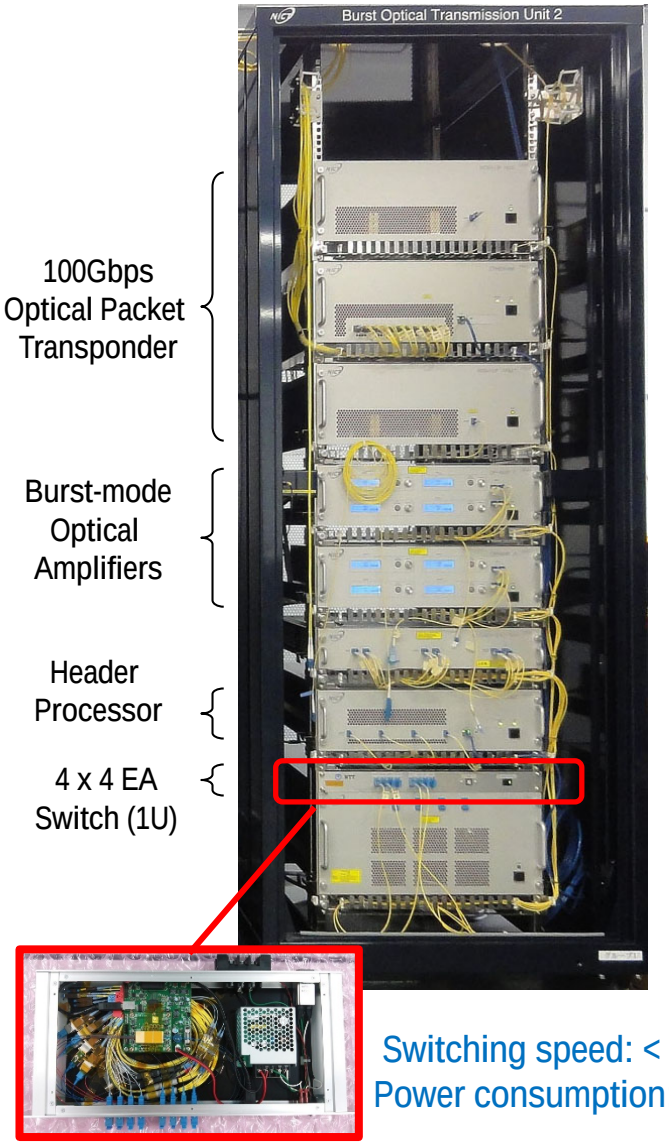
"Locality Aware MPI Communication on a Commodity Opto-Electronic Hybrid Network"





EO-OE Shortcut Forwarding of Messages Across Switches

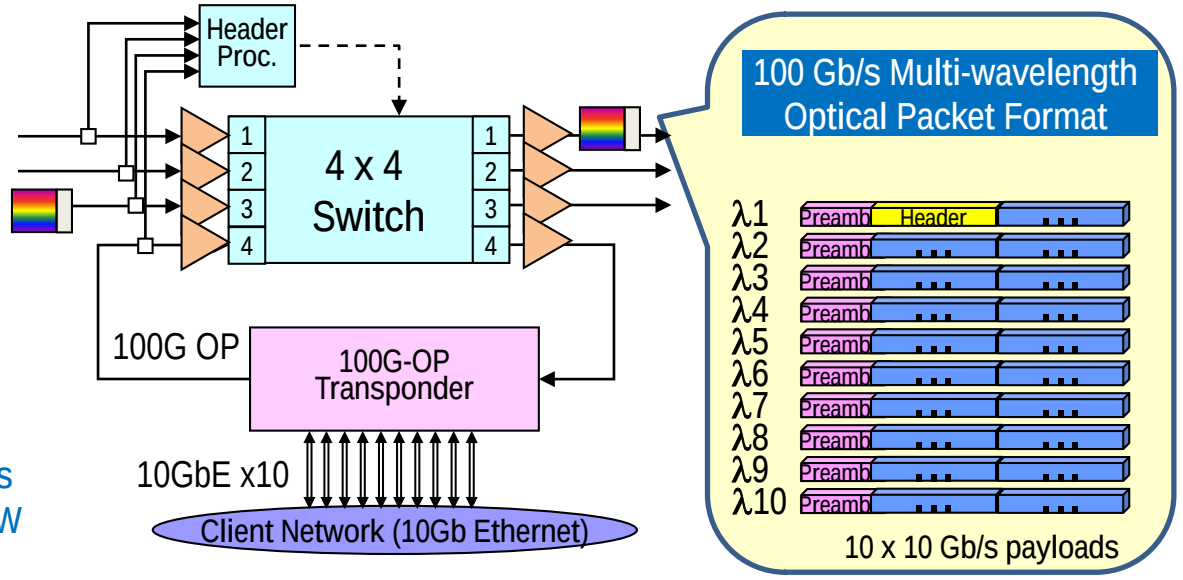
NICT Optical Packet Switch Node (Slides courtesy NICT)



Switching speed: < 8 ns
Power consumption: 3 W

Y. Muranaka, et.al, Photonics in Switching 2015.

- 4 x 4 OPS node with optical packet (OP) transponder
- 100Gb/s OPS port, 10GbE x 10 Client ports
- Stability: Tolerance for environmental disturbance (Polarization, Power fluctuation)
- Total throughput : 800 Gb/s
- Total power consumption: 141 W (w/o Transponder)
- 10-node hopping, 450 km fiber transmission



H. Furukawa, et.al, no.P.4.16, ECOC2015.

HIDEAKI FURUKAWA furukawa@nict.go.jp

Applications & Algorithms

Slides by Kengo Nakajima

Information Technology Center
The University of Tokyo

New Frontiers of Computer & Computational Science towards Post Moore Era

December 22, 2015, Takeda Auditorium, The University of Tokyo

Assumptions & Expectations towards Post-Moore Era

- **Higher Bandwidth**, Larger & Heterogeneous Latency
  - Memory: 3D Stacked Memory
  - Network: Optical Communication
  - Both of Memory & Network will be more hierarchical
- Larger Size of Memory & Cache
- Transaction/Transactional Memory

- Application-Customized Hardware, FPGA
- Large Number of Nodes/Number of Cores per Node
  - under certain constraints (e.g. power, space …)

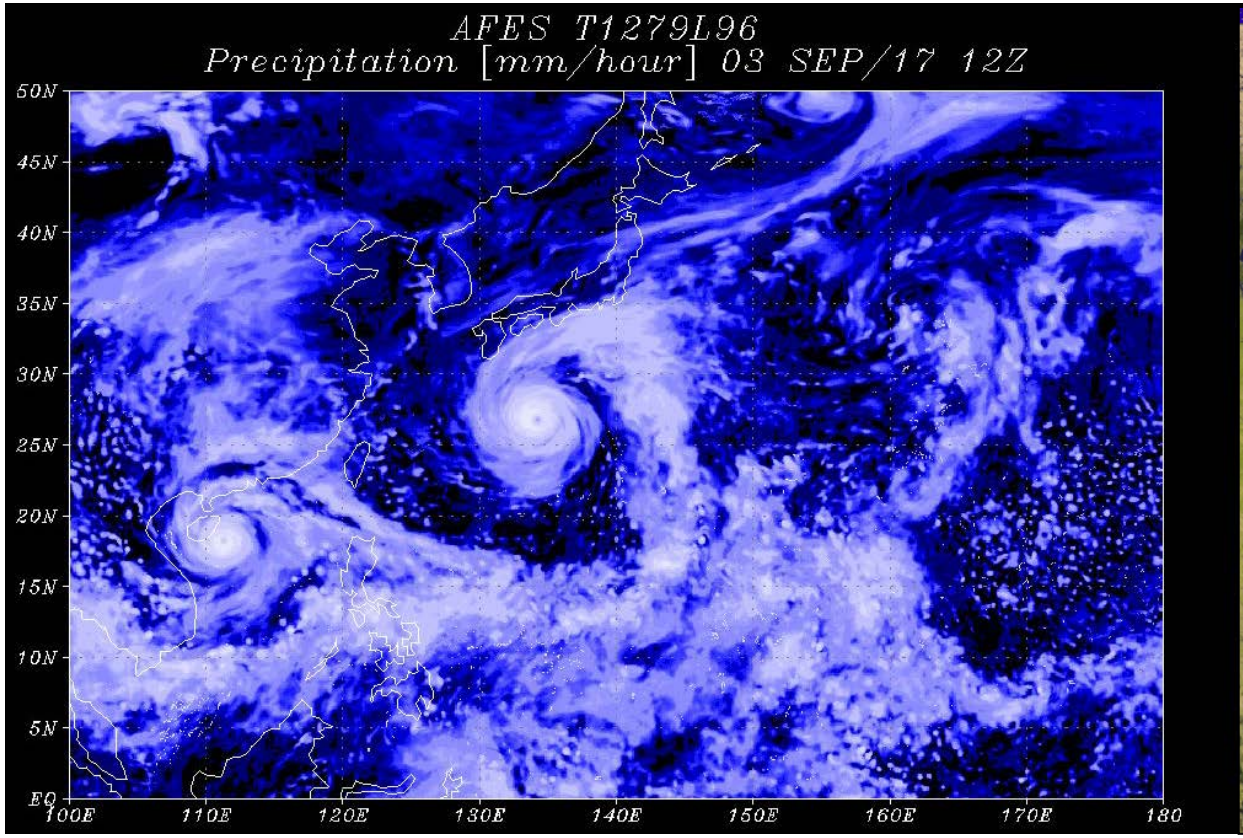
Applications & Algorithms in Post-Moore Era (1/2)

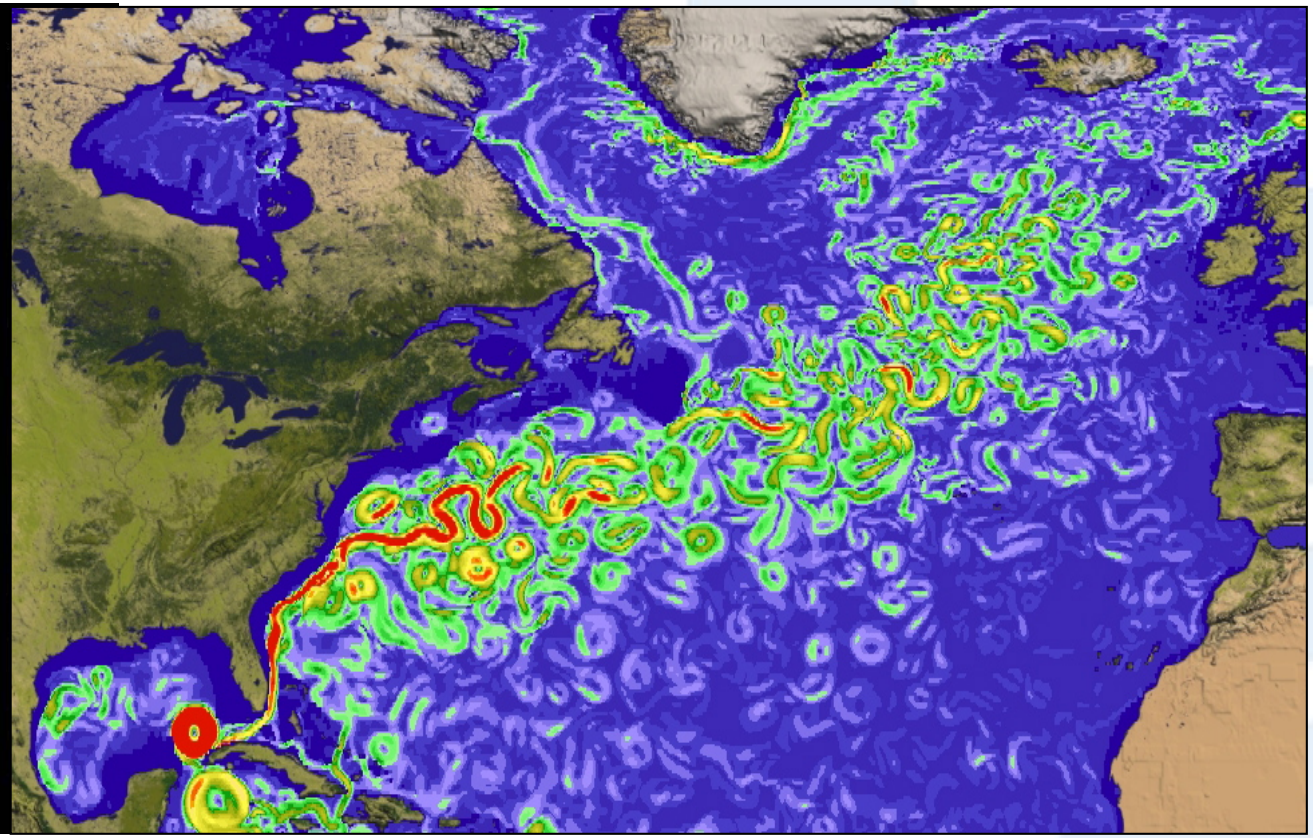
- Compute Intensity ⇒ Data Movement Intensity
  - Non-Blocking Method, Out-of-Core Algorithm
- **Implicit scheme strikes back !**
  - We believe it was never defeated
  - Improvement of performance on sparse matrix computations
  - Big change and advancement are expected in all research areas related to algorithms for sparse matrices including preconditioning
  - **Everything might be easier… but don't relax too much!**
  - **Other Compute to Data Algorithms: H-Matrices**

Applications & Algorithms in Post-Moore Era (2/2)

- **Compute Intensity -> Data Movement Intensity**
  - **Integration of CSE and Data Analysis/Machine Learning**
  - **Data Drive Approach: Machine Learning**
- **Hierarchical Methods for Hiding Latency**
  - **Hierarchical Coarse Grid Aggregation (hCGA) in MG**
  - **Parallel in Space/Time (PiST)**
- **H-Matrix Solver**
  - **Dense-H-Matrix-Sparse: O(n^3) => O(n^2) log n**
- Comm./Synch. Avoiding/Reducing Algorithms
  - Pipelined/Asynchronous CG: MPI_Iallreduce in MPI-3
  - Dynamic Loop Scheduling
- Power-aware Methods
  - Approximate Computing, Power Management, FPGA

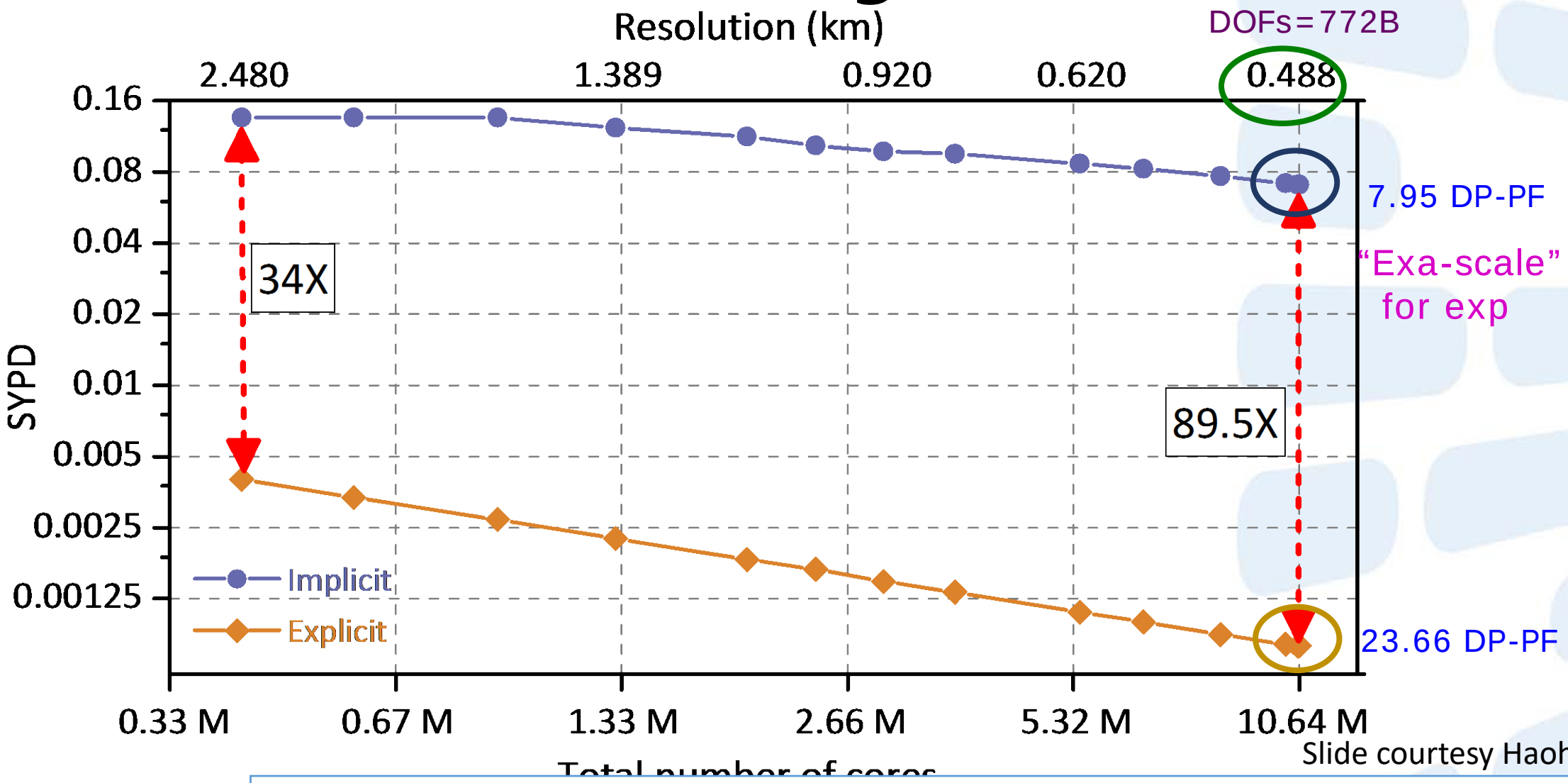
Highly-Scalable Atmospheric Simulation Framework (ACM Gordon Bell Prize 2016)





Slide courtesy Haohuan Fu

Weak-scaling results



Slide courtesy Haohuan Fu

The 488-m res run: 0.07 SYPD, 10.6M cores, dt=240s, 89.5X speedup over explicit

Memory BW Rich Matrix Algorithms: Iwashita (Hokkaido-U), Ida (U-Tokyo)

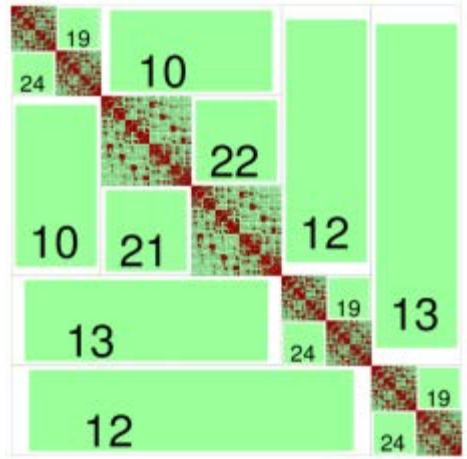
- **H-Matrix**: Low Rank Approximation of Dense Matrix with a Hierarchy of Sparse(+Dense on diagonals) Matrics

H-Matrix Based Methods

from FLOPS-cetric to BYTES (BW) -centric

Dense Matrix

H-Matrix Conversion



Full-Rank (dense)

Low-Rank (sparse)

Convert without sacrificing numerical precision

ppOpen-HPC [Nakajima et. al., U-Tokyo]

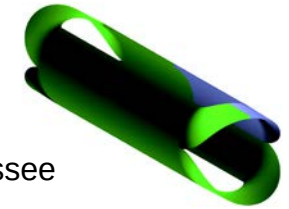
First distributed memory implementation

of H-Matrix HACApk

Earthquake Cycle Simulation（JAMSTEC, AICS）

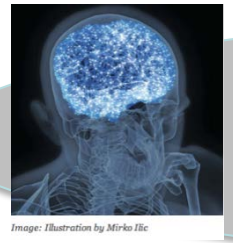
Superconducting Coils (Kyoto-U)

Workd with U-Tennessee

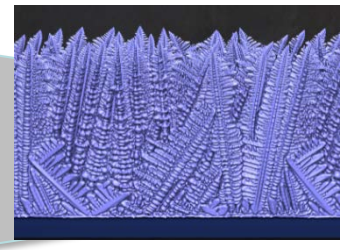


Research Issue:
How to optimize the algorithm to Post-Moore, Bandwidth Rich Architecture
How to apply this to DNN

Co-Designing Post-Moore HPC System Architecture



FLOPS-Oriented => BYTES-Oriented Numerical Applications and Algorithms



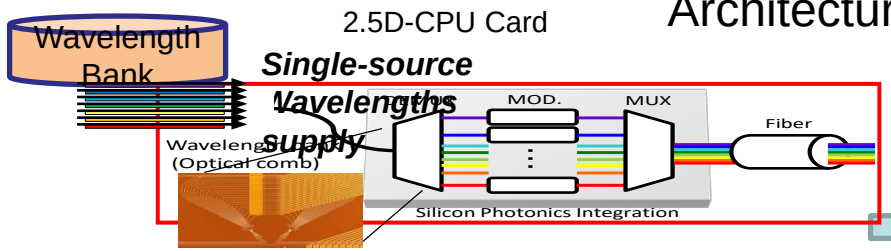
Programming Models and Abstraction?

System SW and Comm MW for Exabit Optical Inteconnect

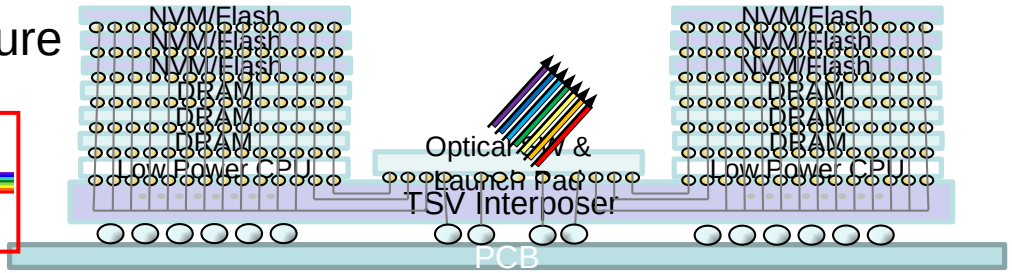
Programming & Perf Modeling of Super Building Block Architecture

System SW and & Programming of Deep and High Bandwidth Memory Hiearchy

Next Gen Exabit-class Optically Switched Interconnect



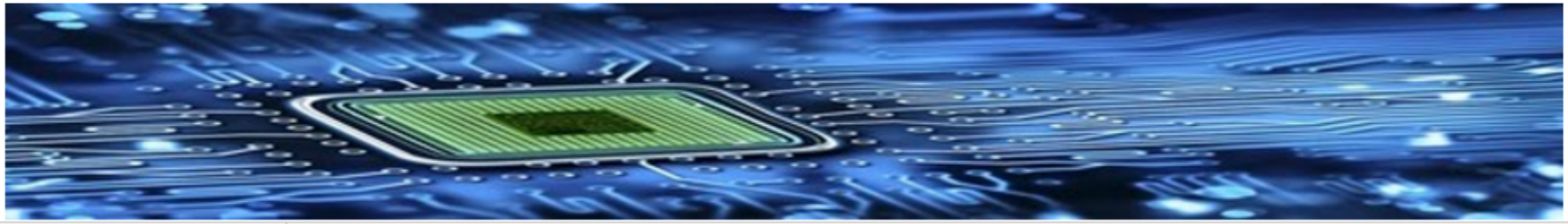
Specialized/Integrated/Reconfigurable Super Building Block Architecture



Advanced 3-D stacked Non-Volatile Memory >Tbytes, >10TByte/s

Post Moore Era Supercomputing Workshop @ SC16

- https://sites.google.com/site/2016pmes/
- Jeff Vetter (ORNL), Satoshi Matsuoka (Tokyo Tech) et. al.



Search this sit

2016 Post-Moore's Era Supercomputing (PMES) Workshop Home

News

Call For Position Papers - Submission Deadline - June 17

Invited Speakers

Photos

Program

Resources

Workshop Venue

Sitemap

188

2016 Post-Moore's Era Supercomputing (PMES) Workshop Home

Co-located with SC16 in Salt Lake City
Monday, 14 November 2016

Workshop URL: http://j.mp/pmes2016
CFP URL: http://j.mp/pmes2016cfp
Submission URL (EasyChair): http://j.mp/pmes2016submissions
Submission questions: pmes16@easychair.org

This interdisciplinary workshop is organized to explore the scientific issues, challenges, and opportunities for supercomputing beyond the scaling limits of

News

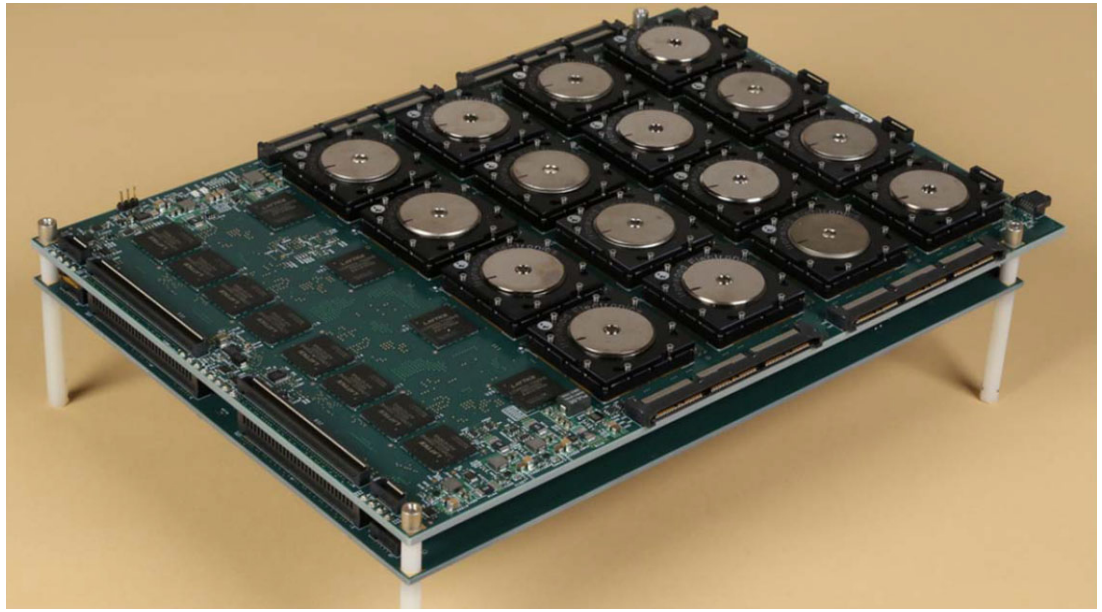
PMES Submission Site Now Open!
PMES Workshop Confirmed for SC16!
Submissions open for PMES Position Papers on April 17

Important Dates

- Submission Site Opens: 17 April 2016

Backup Slides

Neuromorphic Architectures
(Not to be confused with DNN Accelerators)



IBM TrueNorth

SpiNNaker
Biologically Inspired Massively Parallel Architectures

MANCHESTER 1824

Distributed Neuronal Computation "on chip"

Asynchronous Spiking Neural Computation Hardware for low-power real-time operation in Closed-Loop Systems

... to simulate 1 Billion Spiking Neurons in real-time

882 Core "Rack Version"

72 Core Evaluation

18 Core Stand-Alone Prototype

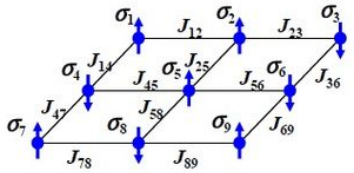
- Multi-channel spiking input and output
- Stand-alone spiking computing system
- Simulates ~20.000 neurons in real time
- Small (~20x20mm); low power (~600mW)
- Flexibly configurable, extendable, stackable

Manchester SpiNNaker
(ARM Based)

- Hitachi@ISSCC2015
  "An 1800-Times-Higher Power-Efficient 20k-spin Ising Chip for Combinational Optimization Problem with CMOS Annealing"
- Competitive to Quantum Annealing, room temperature, easy to scale
- Could be applicable to deep learning?

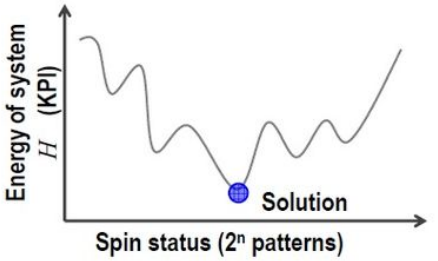
Computing with Ising model

- Ising model: expressing behavior of magnetic spins
- Using Ising model as natural phenomenon to map problems



$$H = -\sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j - \sum_j h_j\sigma_j$$

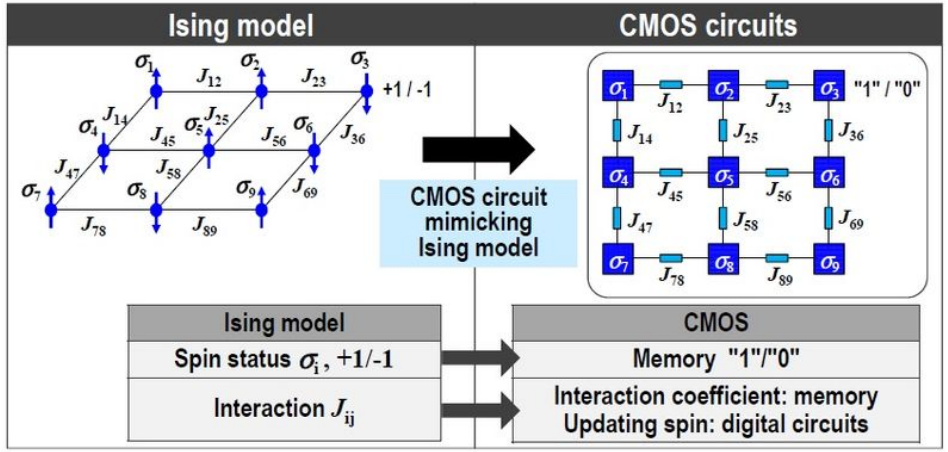
H: Energy of system
σ_i: Spin status (+1/-1)
J_{ij}: Interaction coefficient



n: number of spins

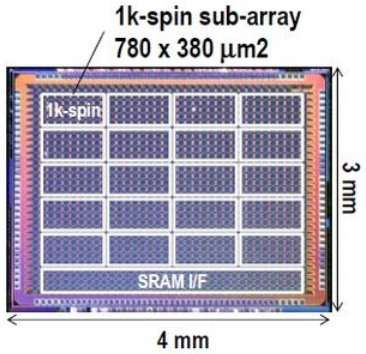
CMOS Ising computing

- Mimicking Ising model with CMOS circuits
- Easy to manufacture, easy to use, good scalability



Ising model	CMOS
Spin status σ_i, +1/-1	Memory "1"/"0"
Interaction J_{ij}	Interaction coefficient: memory Updating spin: digital circuits

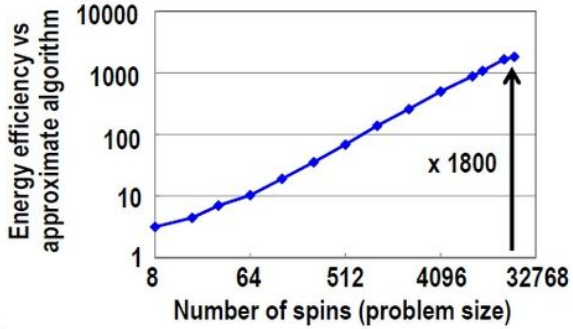
Fabrication results



Items	Value
Number of spins	20k (80 x 256)
Process	65 nm
Chip area	4x3=12 mm²
Area of spin	11.27 x 23.94 =270 µm²
Number of SRAM cells	260k bits Spin value: 1 bit Interaction factor: 2 bit x 5=10 bits External magnetic coefficient: 2 bits
Memory IF	100 MHz
Interaction speed	100 MHz
Operating current of core circuits (1.1 V) Do not include IO	Write: 2.0 mA Read: 6.0 mA Interaction: 44.6 mA

Measurement results of energy

- 1,800 times higher energy efficiency than conventional approximation algorithm on CPU



Conditions:
Randomly generated problems, energy for same preciseness solution
Ising chip: VDD=1.1 V, 100-MHz interaction, best solution among 10-times trial is selected.
Approximation algorithm: SG3(*) is operated on Core i5, 1.87 GHz, 10 W/core.

(*): Sera Kahruman et al., "On Greedy Construction Heuristics for the Max-Cut Problem," International Journal on Computational Science and Engineering, Volume 3, Number 3/2007, pp. 211-218, 2007.

Towards Understanding the Performance of FPGAs using OpenCL Benchmarks [HiPEAC Reconfigurable Computing WS 2015 Extended version to appear SC16]

Hamid Zohouri (Tokyo Tech), Naoya Maruyama (Riken AICS), Satoshi Matsuoka (Tokyo Tech), Motohiko Matsuda (RIKEN AICS)

In collaboration with:
Aaron Smith (Microsoft Research),

Supported by Altera

TSUBAME
Tokyo Institute of Technology

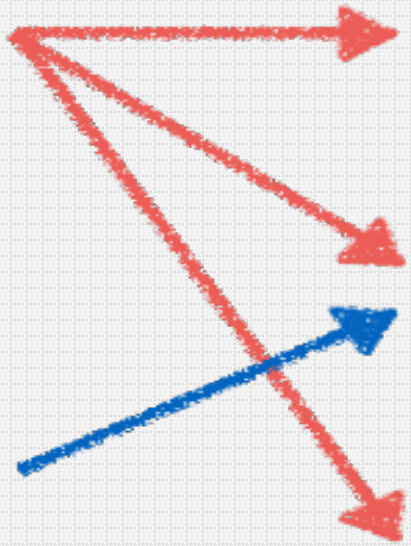
Microsoft Research

ALTERA

Parallelism in Altera OpenCL

Explicit:
Thread/SIMD parallelism

Implicit:
Pipeline parallelism



- **Inter pipelines**
  - Configurable number of duplicated pipelines
- **Intra pipeline**
- **SIMD**
  - Instantiate SIMD units base on user direction (attribute `num_simd_work_items`)

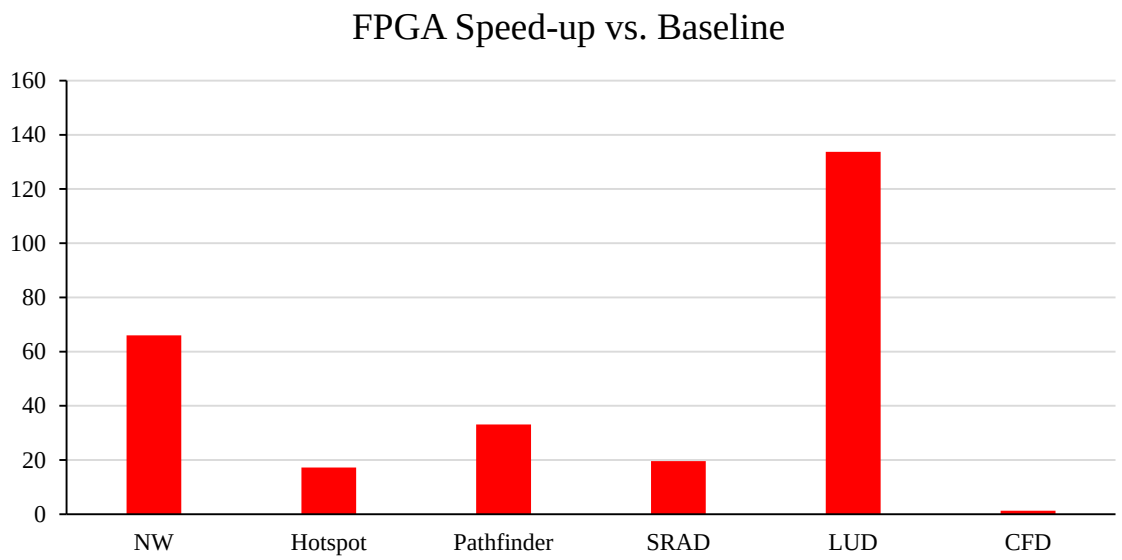
Optimization Effects

Type	Optimization	F_{max} (MHz)	Run Time (ms)	Power Dissipation (Watt)	Power Usage (J)
MT	None	277.2	16574	12.01	199.1
Pipeline	None	243.4	117523	10.59	1245.2
MT	Basic	194.7	2445	16.94	41.4
Pipeline	Basic	249.1	116457	9.93	1156.7
Pipeline	Advanced	148.0	251	15.44	3.8

Sliding window is 66x faster than baseline

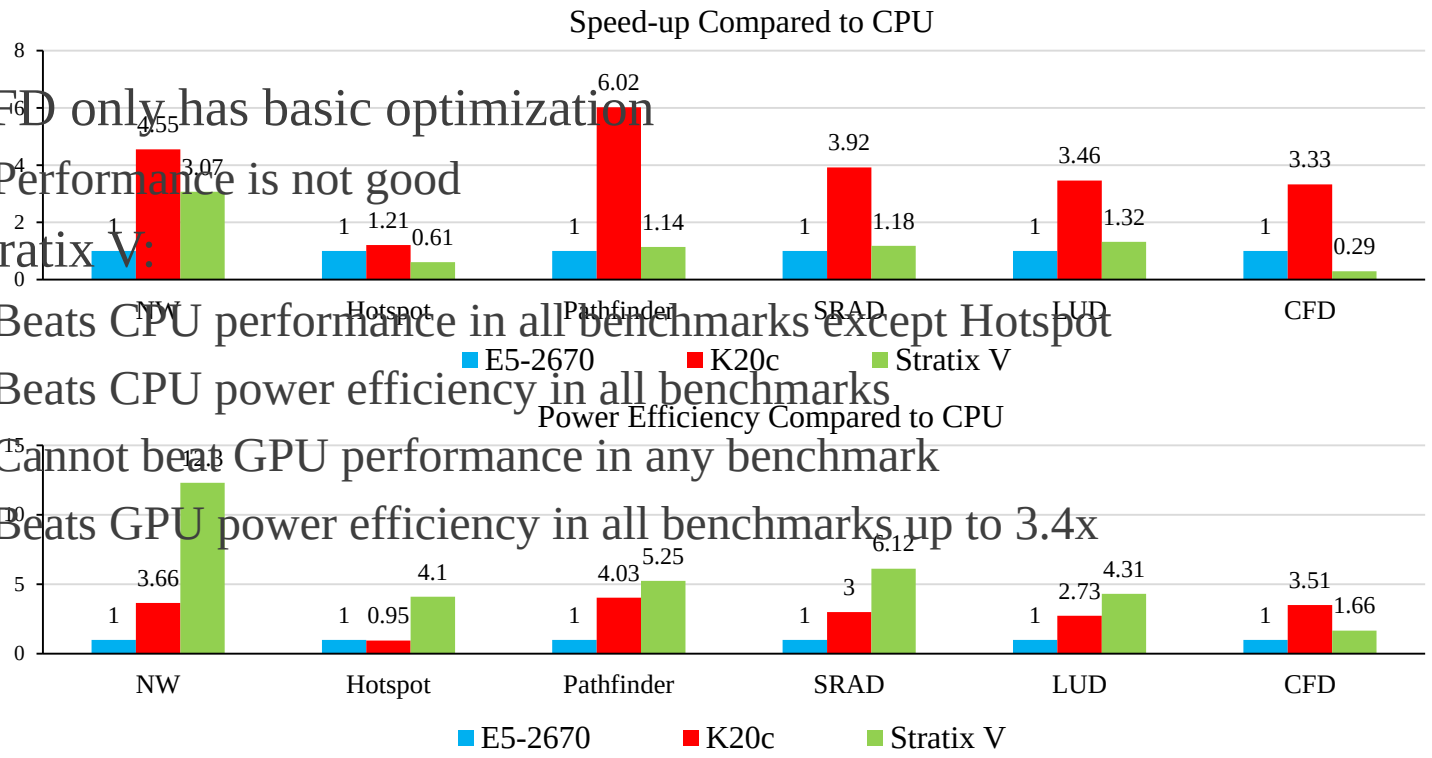
Stratix V Speed-up vs. Baseline

- Up to 133x speed-up
- CFD speed-up is minimal due to lack of area for optimization



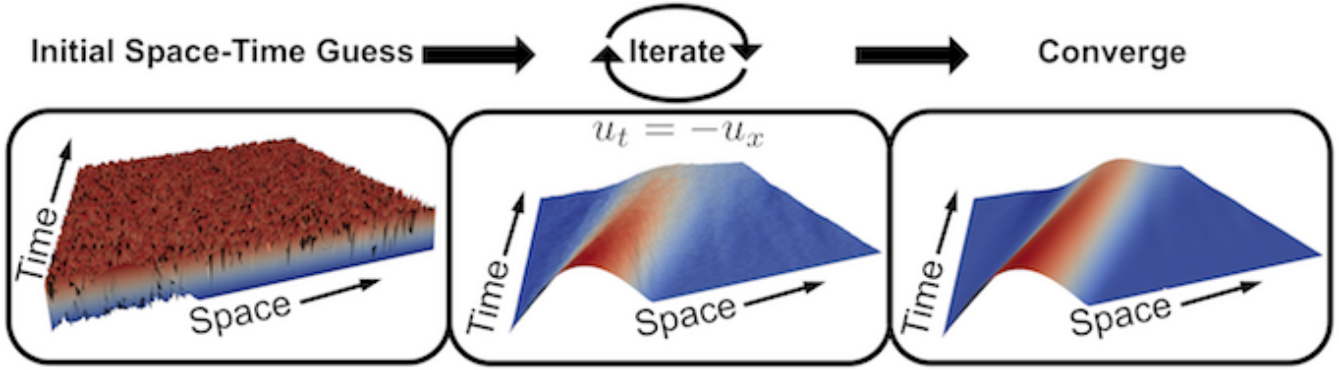
Stratix V vs. E5-2670 & K20c

- CFD only has basic optimization
  - Performance is not good
- Stratix V:
  - Beats CPU performance in all benchmarks except Hotspot
  - Beats CPU power efficiency in all benchmarks
  - Cannot beat GPU performance in any benchmark
  - Beats GPU power efficiency in all benchmarks up to 3.4x



Parallel-in-Space/Time (PiST)

- MG is scalable, but improvement of performance is limited by parallelization only in space direction
- Time-Dependent Problems: Concurrency in Time Dir.
- Multigrid in (Space+Time) Direction
  - ✓ Traditional time-dependent method: Point-Wise Gauss Seidel
  - ✓ XBraid : Lawrence Livermore National Laboratory
    - ☐ Application to nonlinear problems (Transient Navier-Stokes Eqn's)
- MS with 3 sessions in SIAM PP16 (April 2016)
- PiST approach is suitable for the Post-Moore Systems with a complex and deeply hierarchical network that causes large latency.



Specification of OPS

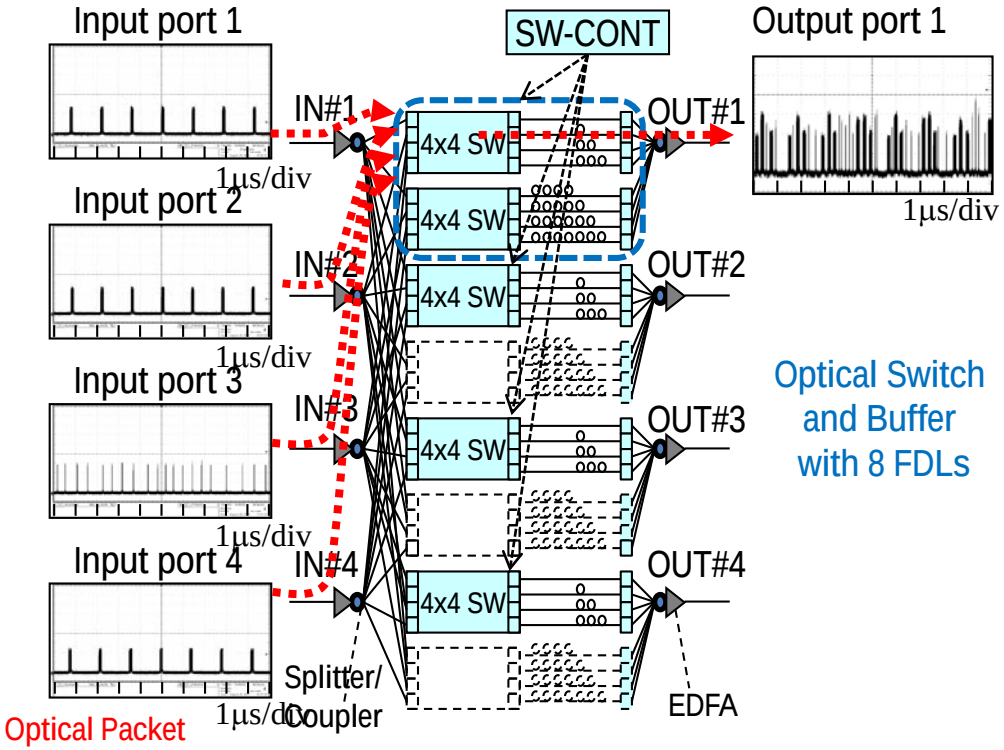
	NICT Optical Packet Switch Node	Experiment and Estimation (Possibility)
I/O ports	4 x 4	2 x 2
Max data rate/port	100 Gb/s (10λ x 10G, OOK)	12.8 Tb/s (64λ x 25G, DP-16QAM, Offline)
Total throughput	800 Gb/s	51.2 Tb/s
Header processing method	OOK header with 16 bit OP Address, 1024 Address (= Entry of Look-up Table)	
Header processing speed per port	250 ns (O/E -> FPGA (Route Selection) -> SW)	
Optical switch	4 x 4 EA switch	1 x 8 PLZT switch
Optical buffer size (Maximum delay)		31 packets (3100 ns)
Power consumption	141 W (w/o OP transponder)	2278 W (Estimation) (w/o OP transponder)

cf) S. Shinada, no. We.3.5.4 ECOC2014.
cf) H. Furukawa, et.al., IEICE Technical report, no.PN2015-20, pp.55-61, 2015.

HIDEAKI FURUKAWA furukawa@nict.go.jp

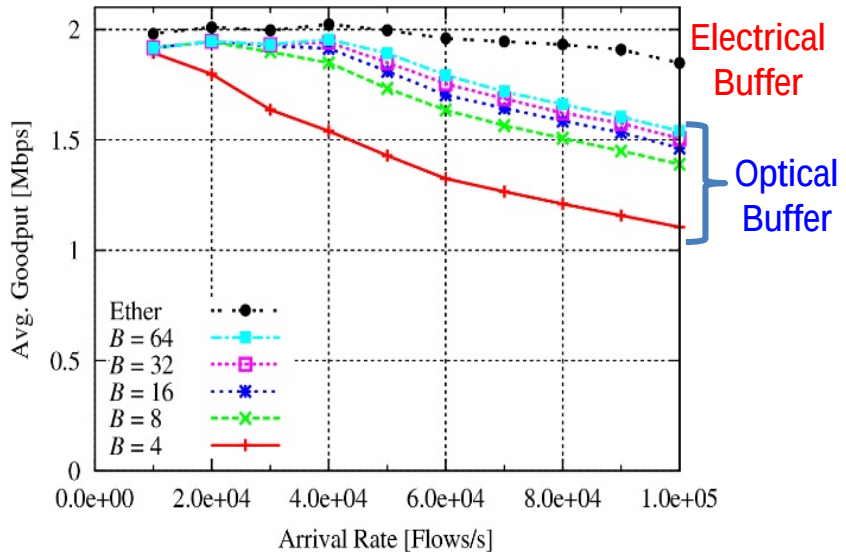
Performance of Optical FDL Buffer

- Experiment: 64, 500, 1500 Byte 100Gbps Optical Packets, 8 FDLs
- Total packet loss rate at 4 input ports was 1.2E-5 in about 35 % load condition.



H. Furukawa, et.al., W2A.19, OFC2014.

- Simulation: 64 Byte ~ 1500 Byte 100Gbps Optical Packets, 4 ~ 64 FDLs
- More than 75% performance at 8-FDL Opt. buffer compared with Ele. buffer



T. Hirayama, et.al, JOCN, vol.7, no.8, pp.776-784, 2015.

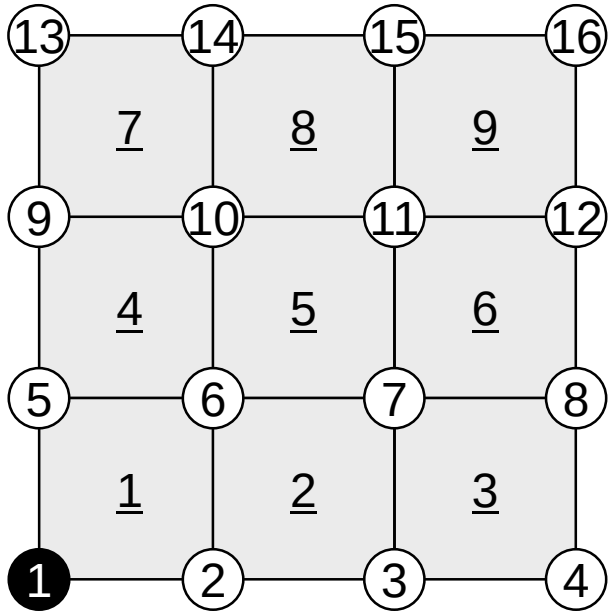
HIDEAKI FURUKAWA furukawa@nict.go.jp

GeoFEM Benchmark: ICCG for FEM Performance of a Node: Flat MPI

	SR11K/J2 Power5+	T2K AMD	FX10	K	Earth Sim 1
Core #/Node	16	16	16	8	8
Peak Performance (GFLOPS)	147.2	147.2	236.5	128.0	64.0
STREAM Triad (GB/s)	101.0	20.0	64.7	43.3	256.0
B/F	0.686	0.136	0.274	0.338	4.00
GeoFEM (GFLOPS)	19.0	4.69	16.0	11.0	25.6
% to Peak	12.9	3.18	6.77	8.59	40.0
LLC/core (MB)	18.0	2.00	0.75	0.75	-

Sparse Solver：Memory-Bound

Improvement of performance on sparse matrix computations due to higher memory bandwidth



Sparse Matrices:
- **FEM**
- **Indirect Memory Access**
- **Memory-Bound**

$$
\begin{bmatrix}
D & X & & & X & X & & & & & & & & & & \\
X & D & X & & X & X & X & & & & & & & & & \\
 & X & D & X & & X & X & X & & & & & & & & \\
 & & X & D & & & X & X & & & & & & & & \\
X & X & & & D & X & & & X & X & & & & & & \\
X & X & X & & X & D & X & & X & X & X & & & & & \\
 & X & X & X & & X & D & X & & X & X & X & & & & \\
 & & & & & & & & & & & & & & X & D
\end{bmatrix}
\begin{Bmatrix}
\Phi_1 \\ \Phi_2 \\ \Phi_3 \\ \Phi_4 \\ \Phi_5 \\ \Phi_6 \\ \Phi_7 \\ \Phi_8 \\ \Phi_9 \\ \Phi_{10} \\ \Phi_{11} \\ \Phi_{12} \\ \Phi_{13} \\ \Phi_{14} \\ \Phi_{15} \\ \Phi_{16}
\end{Bmatrix}
=
\begin{Bmatrix}
F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \\ F_6 \\ F_7 \\ F_8 \\ F_9 \\ F_{10} \\ F_{11} \\ F_{12} \\ F_{13} \\ F_{14} \\ F_{15} \\ F_{16}
\end{Bmatrix}
$$

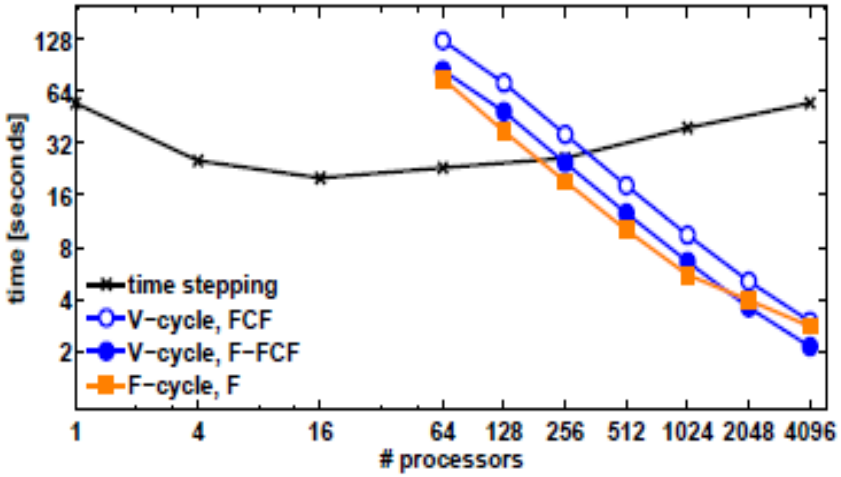
```
{Y}= [A]{X}

do i= 1, N
  Y(i)= D(i)*X(i)
  do k= INDEX(i-1)+1, INDEX(i)
    Y(i)= Y(i) + AMAT(k)*X(ITEM(k))
  enddo
enddo
```

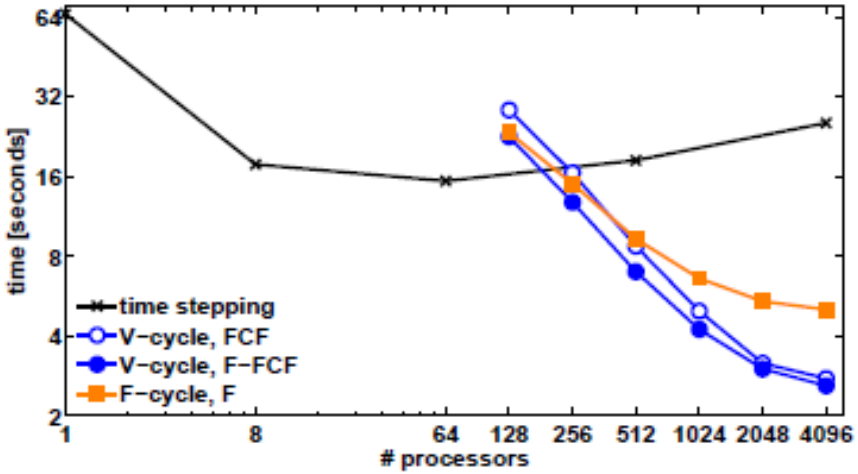
Assumptions & Expectations towards Post-K/Post-Moore Era

- Post-K (-2020)
  - Memory Wall
  - Hierarchical Memory (e.g. KNL: MCDRAM-DDR)
- **Post-Moore (-2025? -2029?)**
  - **Larger Size of Memory & Cache**
  - **Higher Bandwidth, Larger & Heterogeneous Latency**
    - **3D Stacked Memory, Optical Network**
    - **Both of Memory & Network will be more hierarchical**
  - **Application-Customized Hardware, FPGA**
- **Common Issues**
  - **Hierarchy, Latency (Memory, Network etc.)**
  - **Large Number of Nodes/Number of Cores per Node**
    - **under certain constraints (e.g. power, space …)**

Comparison between PiST and "Time Stepping" for Transient Poisson Equations

Effective if processor# is VERY large





2D: $129^2 \times 16385$
16 processors in space direction for PiST

3D: $33^3 \times 4097$
8 processors in space direction for PiST

R. D. Falgout, S. Friedhoff, T. V. Kolev, S. P. MacLachlan, and J. B. Schroder. Parallel time integration with multigrid. *SIAM Journal on Scientific Computing*, *36*(6), C635-C661. 2014

Applications & Algorithms in Post-Moore Era

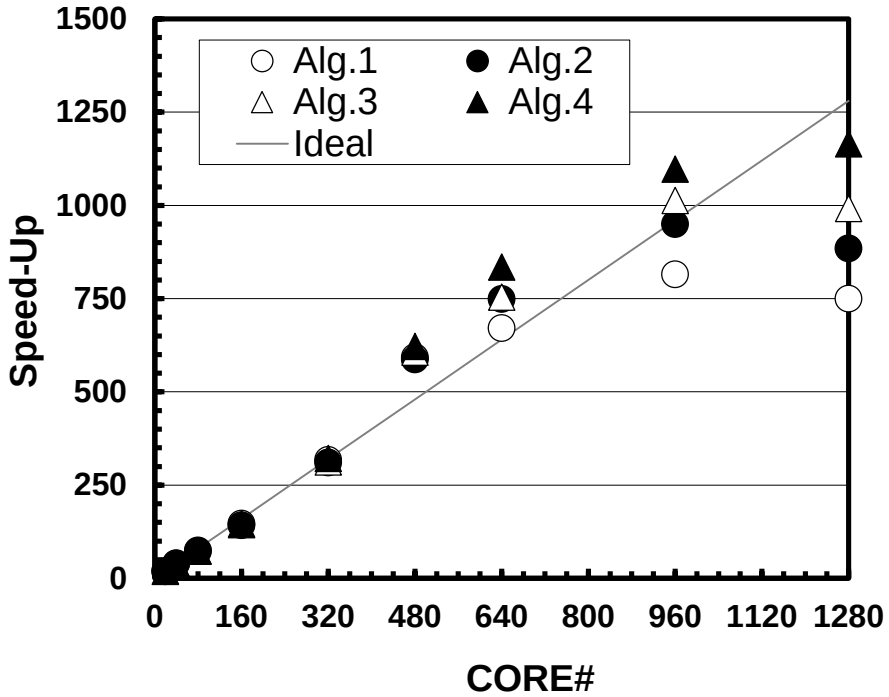
- 計算量重視（Compute Intensity）⇒データ移動重視（Data Movement Intensity）
- Implicit scheme strikes back !: but not straightforward
- Hierarchical Methods for Hiding Latency
  - Hierarchical Coarse Grid Aggregation (hCGA) in Multigrid
  - Parallel in Space/Time (PiST)
- **Communication/Synchronization Avoiding/Reducing Algorithms**
  - **Network latency is already a big bottleneck for parallel sparse linear solvers (SpMV, Dot Products)**
- Utilization of Manycores
- Power-aware Methods
  - Approximate Computing, Power Management, FPGA

3D FEM
Solid Mechanics
96x80x64 nodes
Strong Scaling

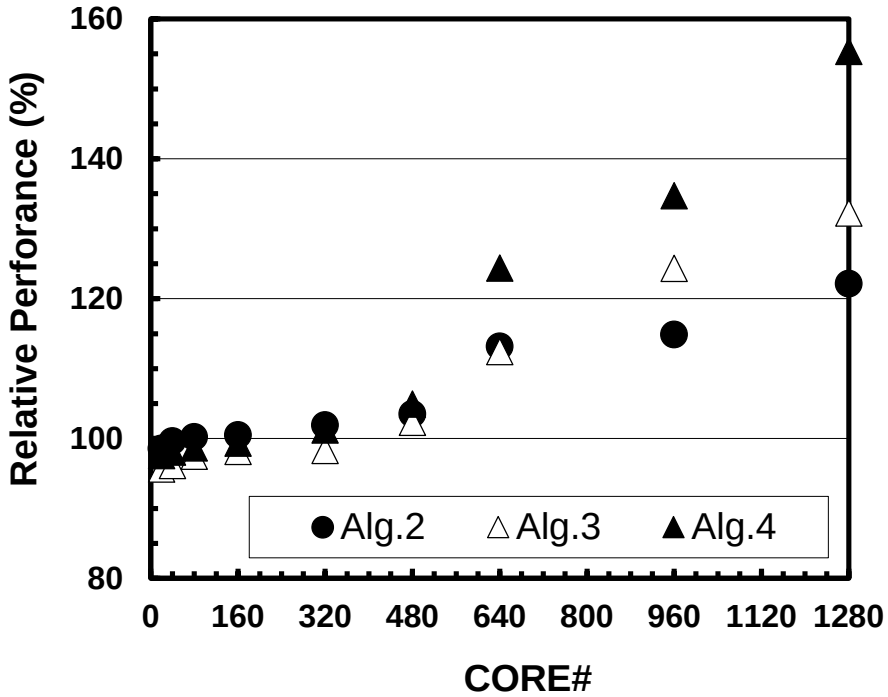
Alg.1 Original PCG
Alg.2 Chronopoulos/Gear
Alg.3 Pipelined CG
Alg.4 Gropp's CG

P. Ghysels et al., Hiding global synchronization latency in the preconditioned Conjugate Gradient algorithm, Parallel Computing 40, 2014

Speed-Up (20-1,280 cores)



Relative Performance to Alg.1 (Original)



Alg.4: Gropp's Asynch. CG

Dot products – Preconditioning/SpMV

Algorithm 7. Gropp's asynchronous CG

1: $r_0 := b - Ax_0$; $u_0 := M^{-1}r_0$; $p_0 := u_0$; $s_0 := Ap_0$; $\gamma_0 := (r_0, u_0)$
2: **for** $i = 0, \ldots$
3: $\quad \delta := (p_i, s_i)$
4: $\quad q_i := M^{-1}s_i$
5: $\quad \alpha_i := \gamma_i / \delta$
6: $\quad x_{i+1} := x_i + \alpha_i p_i$
7: $\quad r_{i+1} := r_i - \alpha_i s_i$
8: $\quad u_{i+1} := u_i - \alpha_i q_i$
9: $\quad \gamma_{i+1} := (r_{i+1}, u_{i+1})$
10: $\quad w_{i+1} := Au_{i+1}$
11: $\quad \beta_{i+1} := \gamma_{i+1} / \gamma_i$
12: $\quad p_{i+1} := u_{i+1} + \beta_{i+1} p_i$
13: $\quad s_{i+1} := w_{i+1} + \beta_{i+1} s_i$
14: **end for**