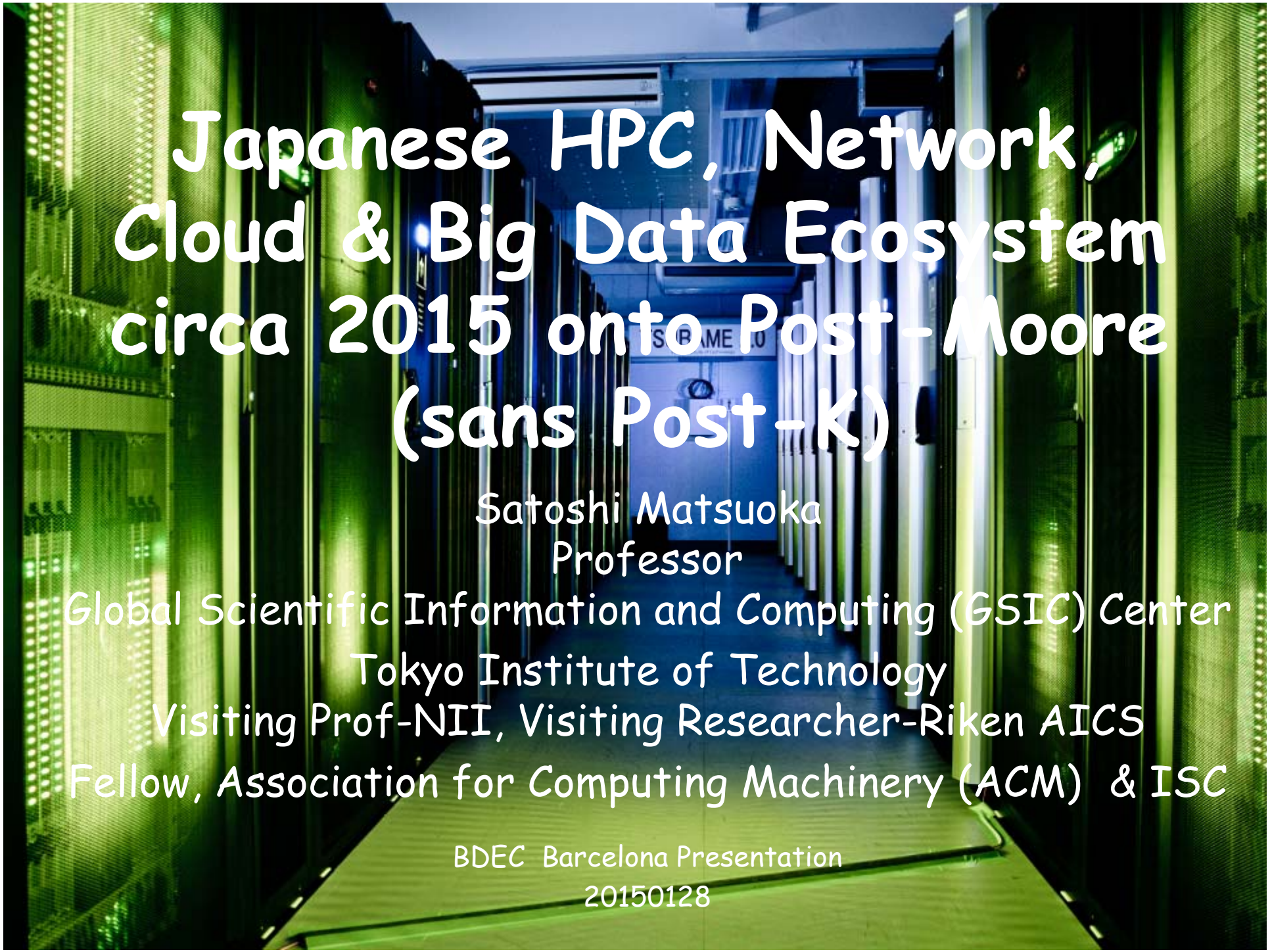


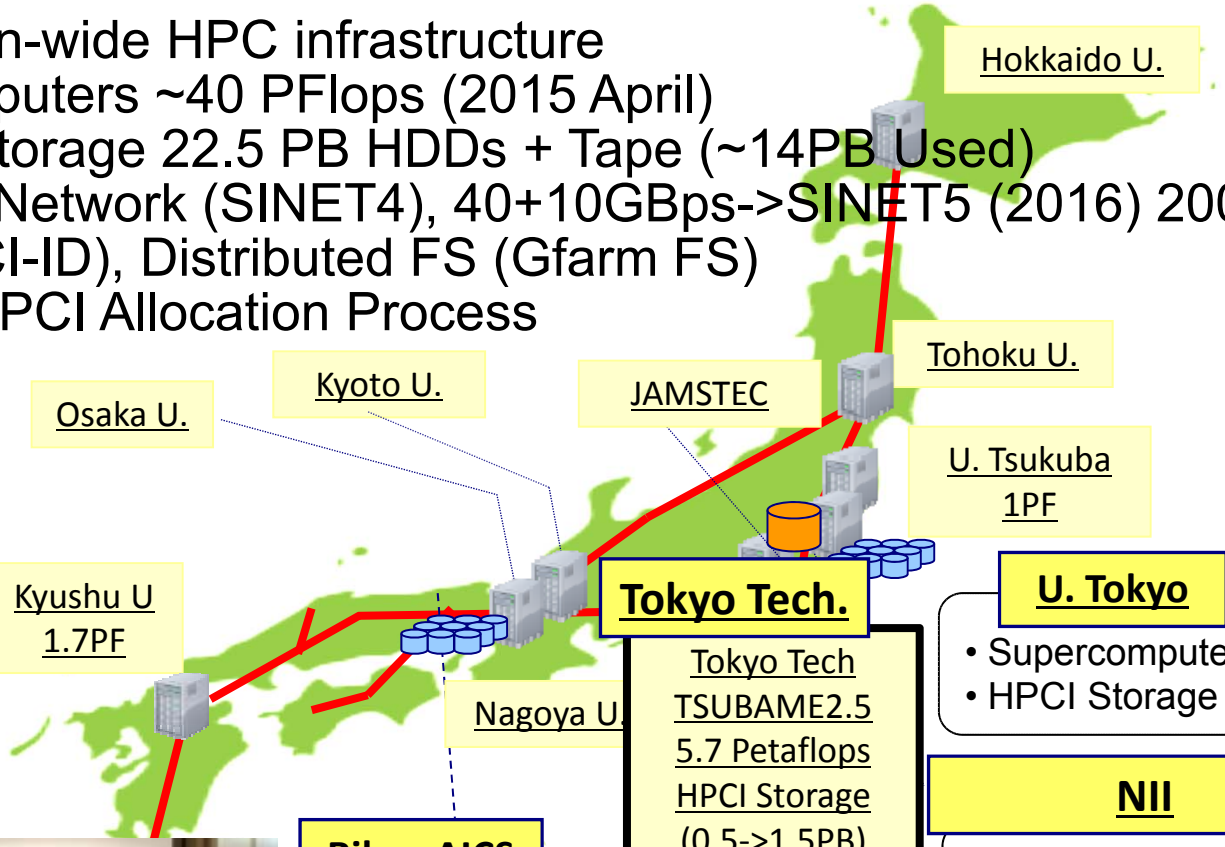
Japanese HPC, Network, Cloud & Big Data Ecosystem circa 2015 onto Post-Moore (sans Post-K)

Satoshi Matsuoka
Professor
Global Scientific Information and Computing (GSIC) Center
Tokyo Institute of Technology
Visiting Prof-NII, Visiting Researcher-Riken AICS
Fellow, Association for Computing Machinery (ACM) & ISC

BDEC Barcelona Presentation
20150128

Japan's High Performance Computing Infrastructure (HPCI)

HPCI: a nation-wide HPC infrastructure
- Supercomputers ~40 PFlops (2015 April)
- National Storage 22.5 PB HDDs + Tape (~14PB Used)
- Research Network (SINET4), 40+10GBps->SINET5 (2016) 200~400Gbps
- SSO (HPCI-ID), Distributed FS (Gfarm FS)
- National HPCI Allocation Process



Tokyo Tech.

Tokyo Tech
TSUBAME2.5
5.7 Petaflops
HPCI Storage
(0.5->1.5PB)

U. Tokyo

- Supercomputers 1.1PF
- HPCI Storage (12PB)

NII

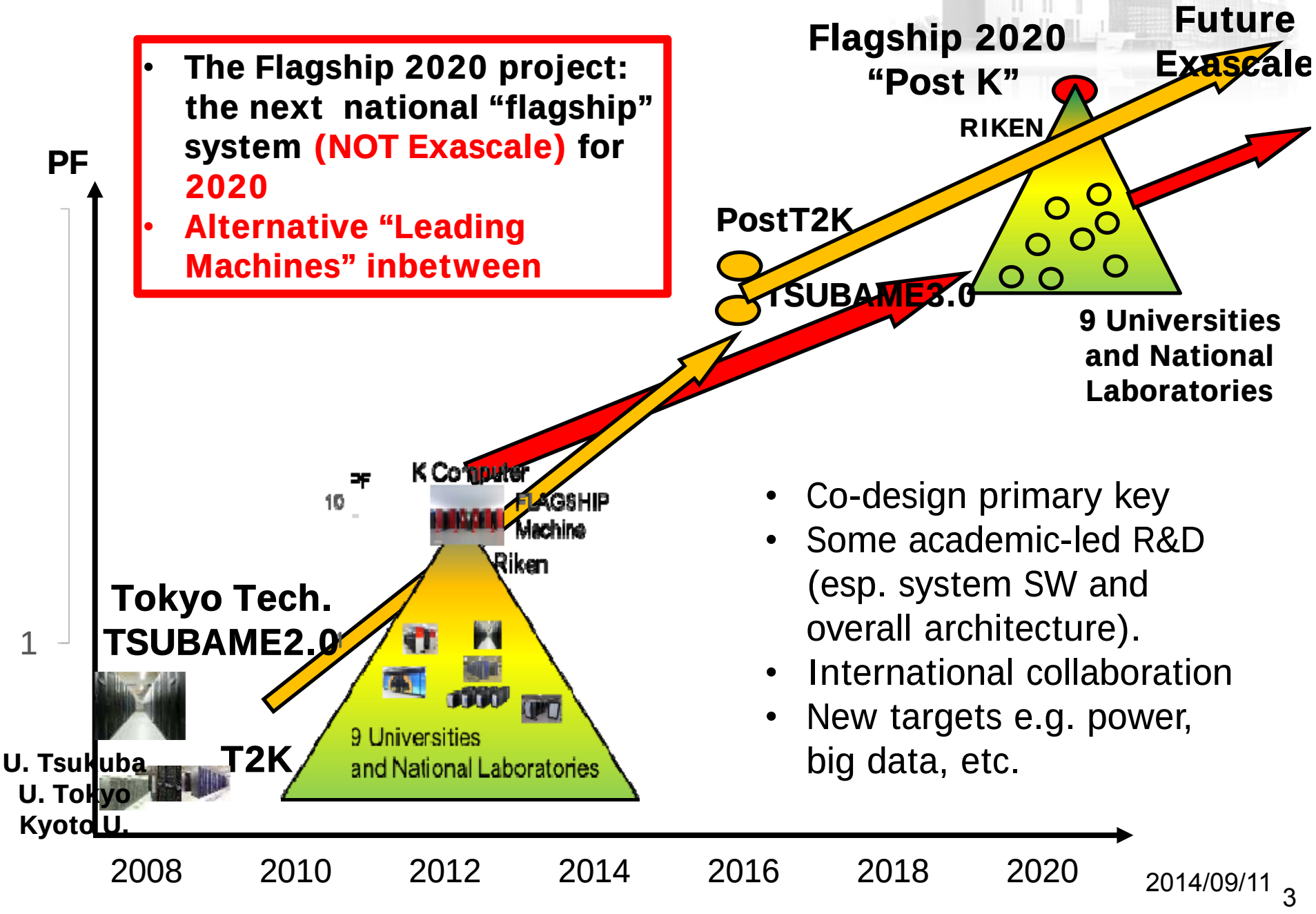
- Management of SINET &Single sign-on

Riken AICS

- "K" computer 11Petaflops
- HPCI Storage (10PB)

Towards the Next Flagship Machine & Beyond

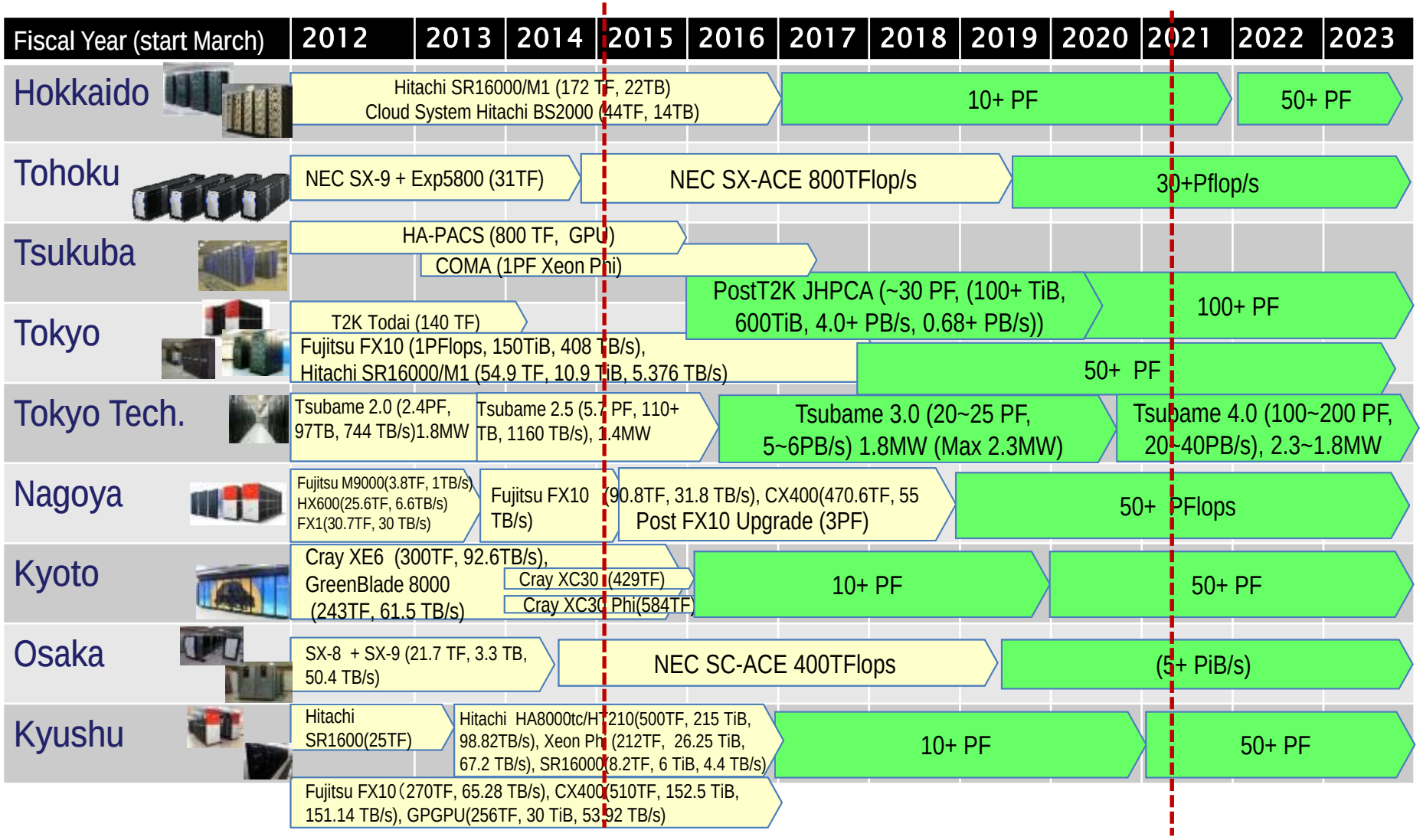
- **The Flagship 2020 project: the next national "flagship" system (NOT Exascale) for 2020**
- **Alternative "Leading Machines" inbetween**



- Co-design primary key
- Some academic-led R&D (esp. system SW and overall architecture).
- International collaboration
- New targets e.g. power, big data, etc.

Japanese “Leading Machine” Candidates
Roadmap of the 9 HPCI University Centers

+ Post K -> Total ExaFlop?



~17PF April 2015, Japan-wide ~40PF(incl. K),

HPCI Nationwide HPC Storage Cloud

- 21.8 PB (separate from local)
  ~70% full
- High resiliency and availability
  - Redundant Servers･RAID6
  - Active Repair
- Multi-Tier Distributed Storage
  - Multi-vendor utility
  - ZABBIX, Ganglia
- Fault Detection & Information Sharing

HPCI East HUB
Univ. Tokyo
• 11.5PB + 20PB Tape

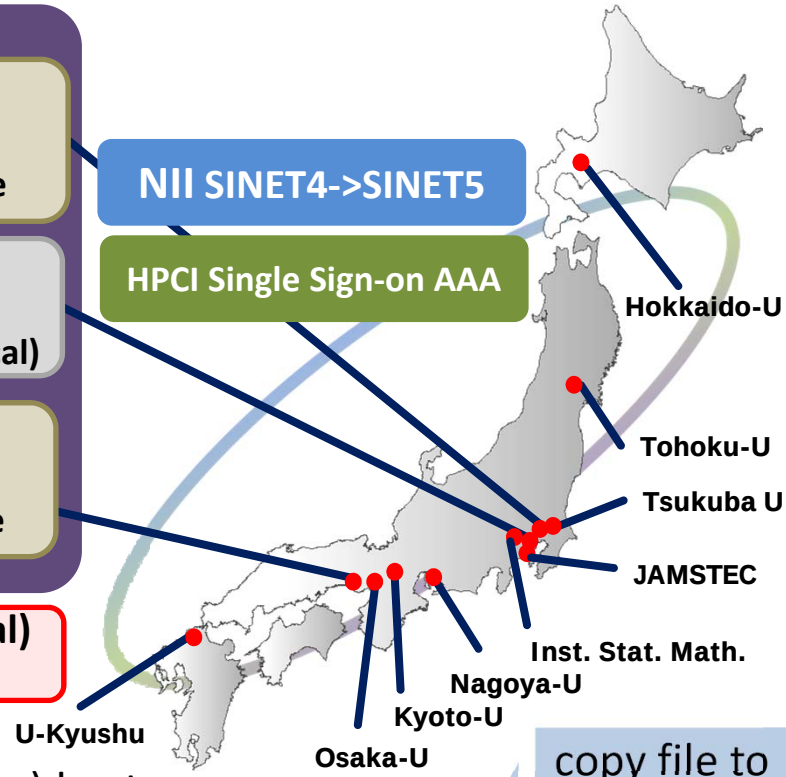
Tokyo Tech
• 0.3PB -> 1.2PB
• (TSUBAME 11PB Local)

HPCI West HUB
Riken AICS
• 10PB + 60PB Tape

K Computer (30PB Local)
=> PostK (2020)

NII SINET4->SINET5

HPCI Single Sign-on AAA

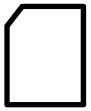


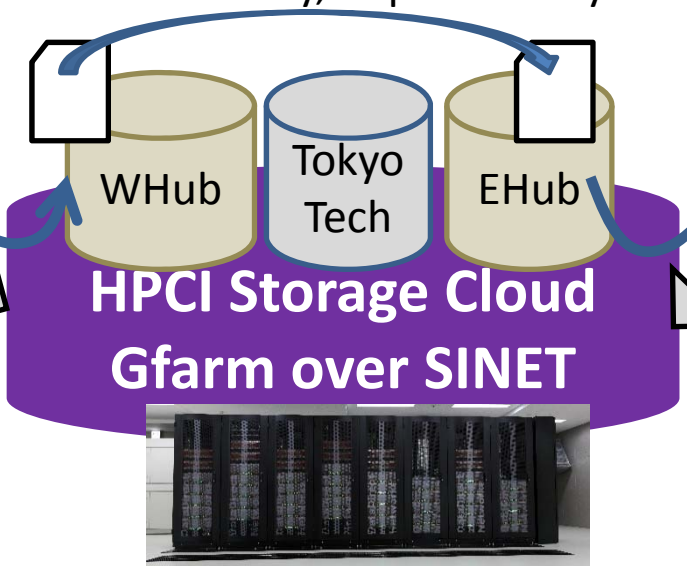
HPCI High Performance Computing Infrastructure

replication to (neighbor) host
- access efficiency, dependability

copy file to
"Gfarm"
via login node

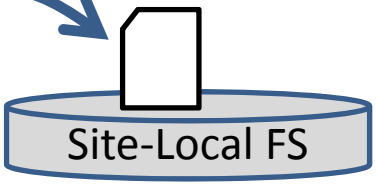
mount

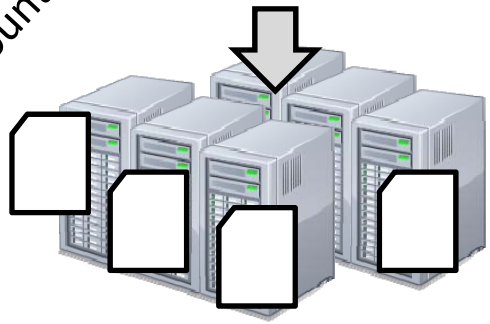




mount

copy file to
Site-local FS
via login node





SINET5: Nationwide Academic Network

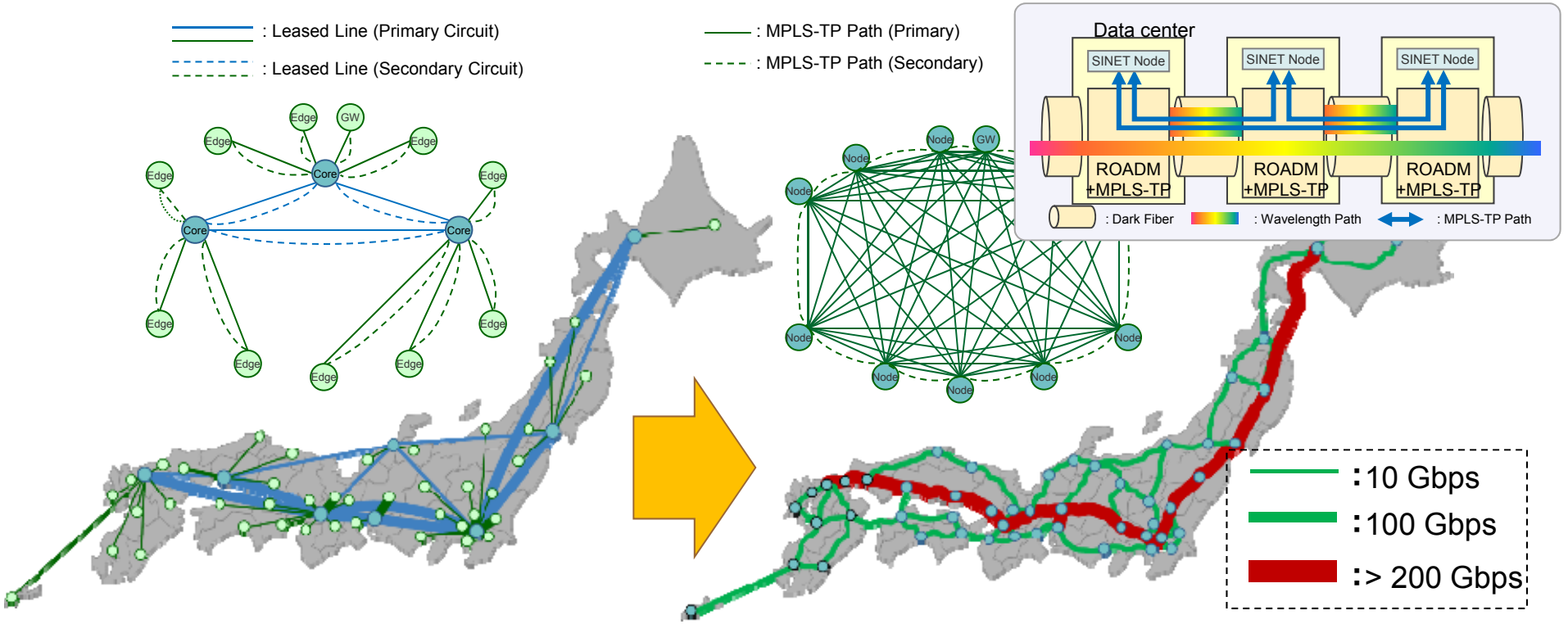
- 2016 SINET5 connects all the SINET nodes in a fully-meshed topology and minimizes the latency between every pair of the nodes using nationwide dark fiber
- MPLS-TP devices connect a pair of the nodes by primary and secondary MPLS-TP paths.

SINET4 present

- Connects nodes in a star-like topology
- Secondary circuits of leased lines need dedicated resources

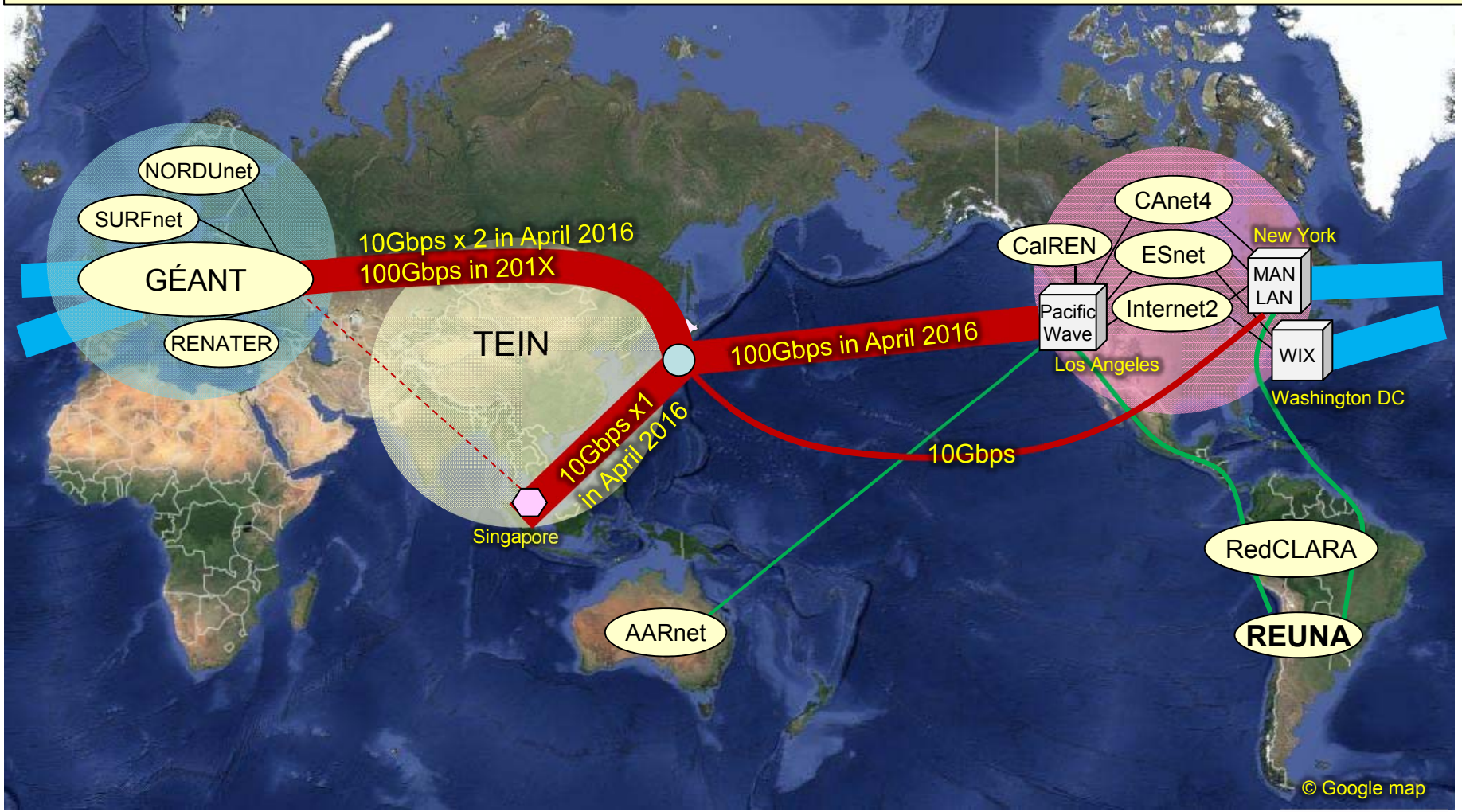
SINET5 2016

- Connects all the nodes in a fully-meshed topology with redundant paths
- Secondary paths do not consume resources



International Lines of SINET5

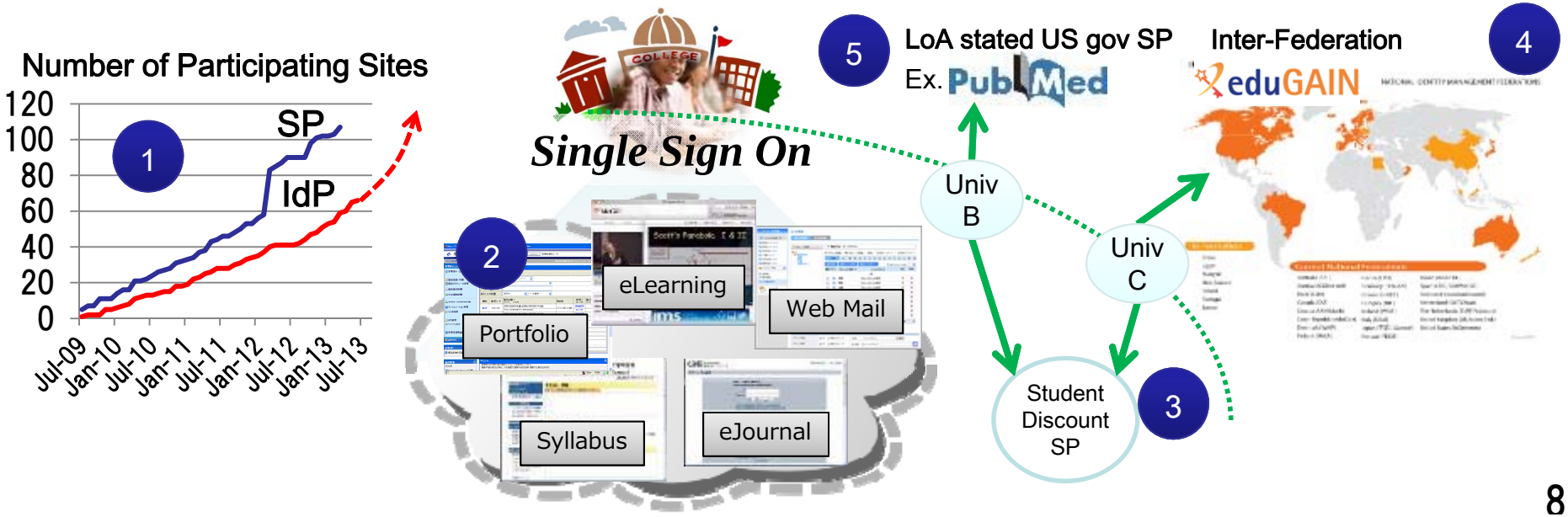
- 100-Gbps line to U.S. West Coast and will keep a 10-Gbps line to U.S. East Coast.
- Two direct 10-Gbps lines to Europe in April 2016, possibility of a 100-Gbps in the near future.
- SINET will keep a 10-Gbps line to Singapore in April 2016.



Academic Access Federation GakuNin

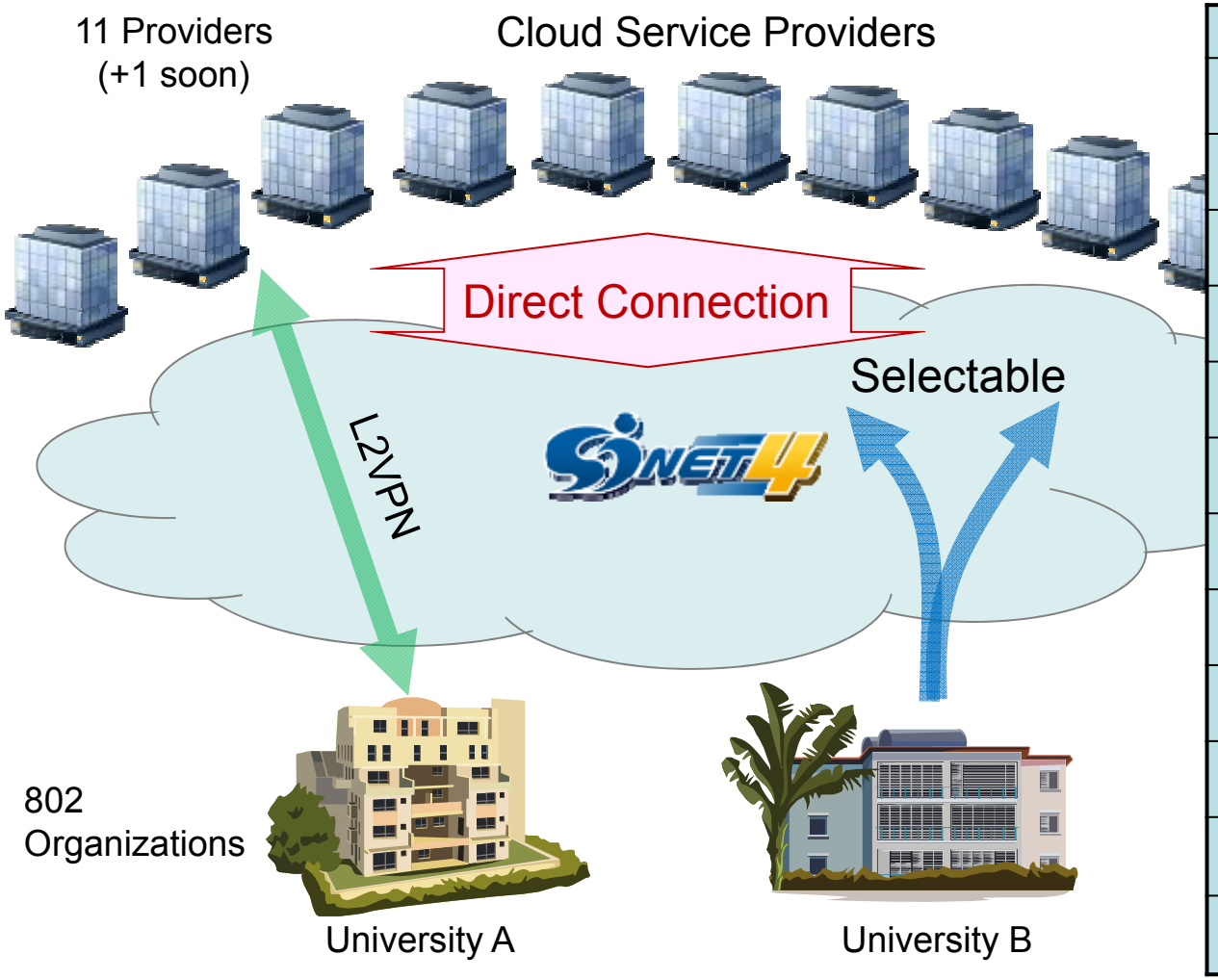
The Japanese academic access federation, GakuNin, is deploying federated identify in Japan using the SAML 2.0 standard, primarily with Shibboleth software.

1. Number of participants are **rapidly increasing** and becoming as a **de facto standard** of the current **HE** (higher education) infrastructure.
2. Some of the commercial service providers are very interested in as a tool for proofing the **student status** on the Internet.
3. GakuNin is also a member of **eduGAIN** which facilitates the **Inter-Federation**.
4. GakuNin is a **level 1 TFP** (Trust Framework Provider) certified by OIX and now preparing for higher LoA (Level of Assurance).
5. Future HPCI authorization to incorporate Gakunin



Infrastructure for Cloud Services

- ◆ Eleven service providers directly connect to SINET and offer cloud services.
- ◆ SINET users expect high-performance, secure, and inexpensive cloud services.



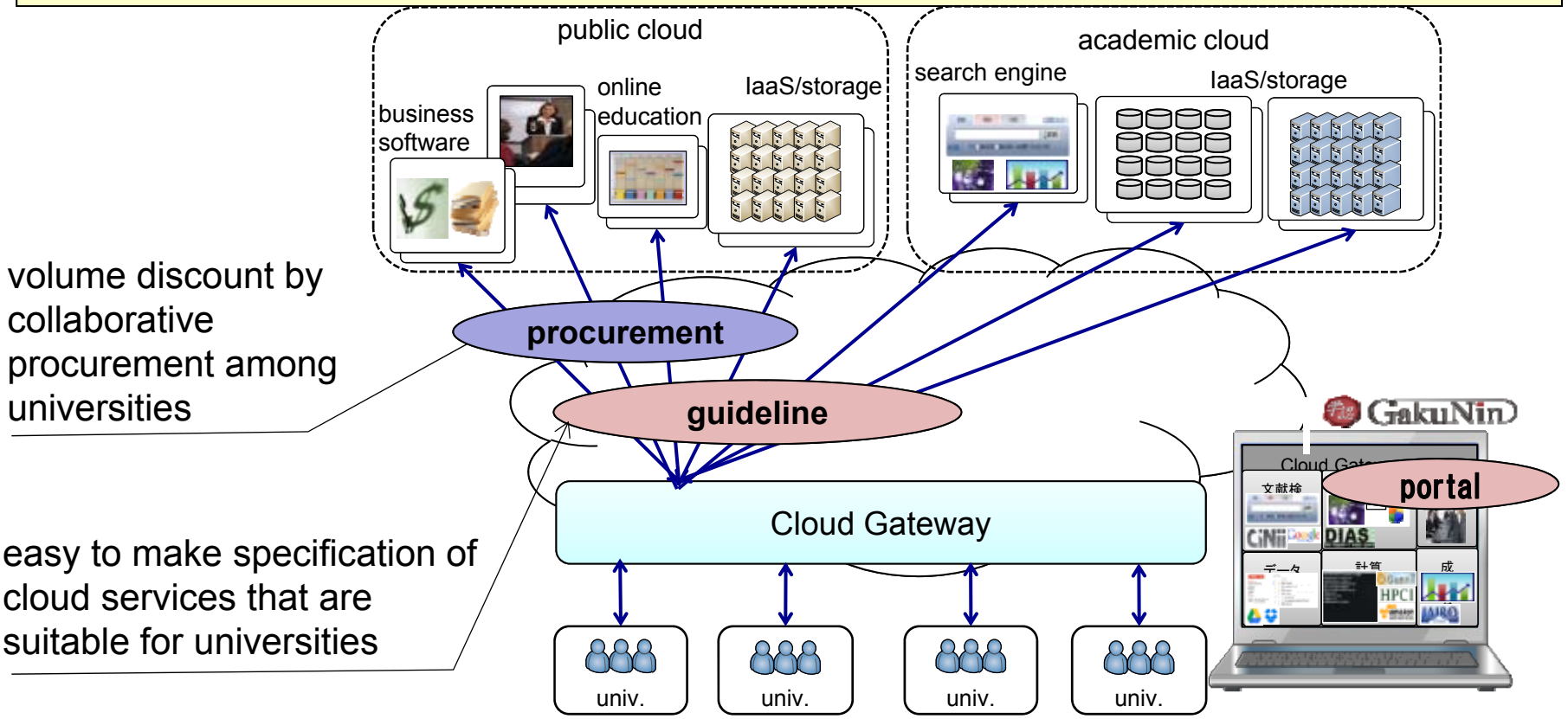
Sapporo: 2, Tokyo: 6, Osaka: 3, Fukuoka: 2

DC	ISP	Status
Sapporo	SAKURA Internet	OK
Sapporo	NTT EAST	OK
Tokyo	amazon web services	OK
Tokyo	CTC	OK
Tokyo	NTT Communications	OK
Tokyo	DATAHOTEL	OK
Tokyo	FUJITSU	OK
Tokyo	UQ WiMAX	OK
Osaka	IIJ	OK
Osaka	GMO CLOUD WEST	Soon
Osaka Fukuoka	NTT Smart Connect	OK
Osaka Fukuoka	NTT DATA	OK

Cloud Service Marketplace

◆ The cloud service marketplace will help the users to easily develop cloud service specifications and enable joint procurements for the same cloud services, which will lead to dramatic cost reduction in academia as a whole.

- the checklist (or guideline) to select suitable cloud services
- the evaluation results of cloud services in accordance with the checklist



CREST: Development of System Software Technologies for post-Peta Scale High Performance Computing 2010H2-2018

- Objectives
  - Co-design of system software with applications and post-peta scale computer architectures
  - Development of deliverable software pieces
- Research Supervisor
  - Akinori Yonezawa, Deputy Director of RIKEN AICS

- Run by JST (Japan Science and Technology Agency)
- Budget and Formation (2010 to 2018)
  - About 60M $ (47M$ in normal rate) in total
  - Round 1: From 2010 for 5.5 year
  - Round 2: From 2011 for 5.5 year
  - Round 3: From 2012 for 5.5 year

http://www.postpeta.jst.go.jp/en/

- *NEW: Joint DFG (Germany) & ANR (France) SPPEXA2 call 2016*

ISP2S2: JST CREST International Symposium on Post Petescale System Software

http://wallaby.aics.riken.jp/isp2s2/

- **December 2-4, 2014**
  - **RIKEN AICS, Kobe University**
  - **Joint Symposium of 14 Projects of "Development of System Software Technologies for Post-Peta Scale High Performance Computing" (Supervisor: Prof. A. Yonezawa, RIKEN AICS)**
  - **14 Invited international speakers from US, Europe, …**

Overview of PPC CREST (slide 1 of 3)

CREST: Development of System Software Technologies for post-Peta Scale High Performance Computing

2013	2014	2015	2016	2017
Round 1: 5 teams run				
Round 2: 5 teams run				
Round 3 : 4 teams run				

Takeshi Nanri, Kyushu University

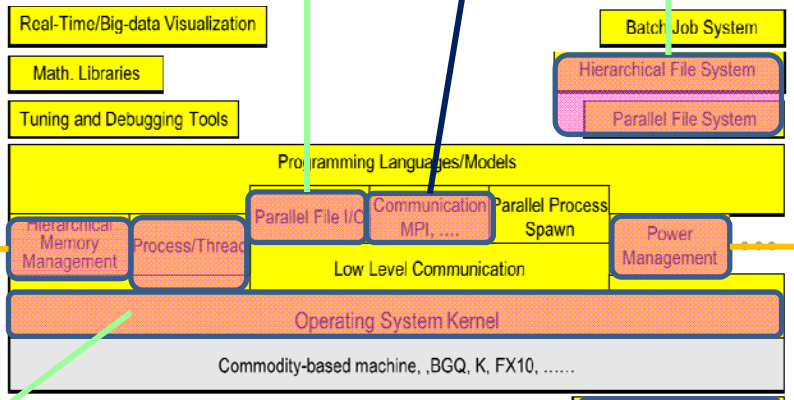
Development of Scalable Communication Library with Technologies for Memory Saving and Runtime Optimization

Toshio Endo, Tokyo Tech.

Software Technology that Deals with Deeper Memory Hierarchy in Post-petascale Era

Osamu Tatebe, U. of Tsukuba

System Software for Post Petascale Data Intensive Science



Taisuke Boku, U. of Tsukuba

Research and Development on Unified Environment of Accelerated Computing and Interconnection for Post-Petascale Era

Atsushi Hori, RIKEN AICS

Parallel System Software for Multi-core and Many-core

Masaaki Kondo, U. of Electro-Comm.

Power Management Framework for Post-Petascale Supercomputers

Overview of PPC CREST (slide 2 of 3)

CREST: Development of System Software Technologies for post-Peta Scale High Performance Computing

2013	2014	2015	2016	2017
Round 1: 5 teams run				
Round 2: 5 teams run				
Round 3 : 4 teams run				

Naoya Maruyama, Riken AICS

Highly Productive, High Performance Application Frameworks for Post Petascale Computing

Hiroyuki Takizawa, Tohoku University

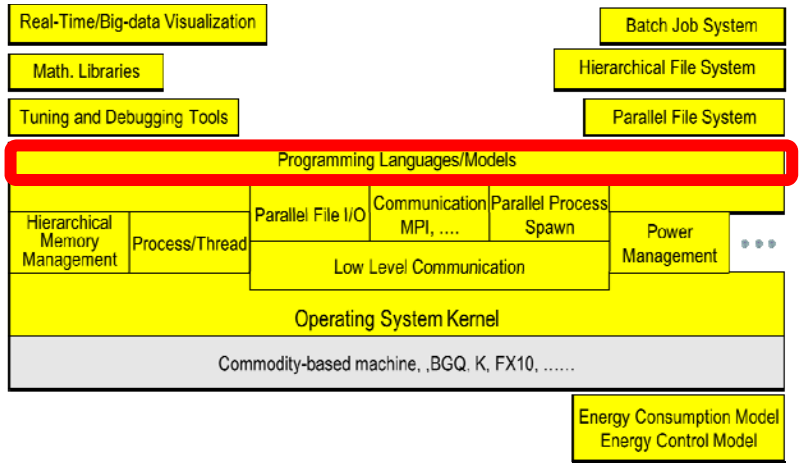
An evolutionary approach to construction of a software development environment for massively-parallel heterogeneous systems

Shigeru Chiba, Tokyo Tech.

Software development for post petascale super computing --- Modularity for Super Computing

Itsuki Noda, AIST

Framework for Administration of Social Simulations on Massively Parallel Computers



Overview of PPC CREST (slide 3 of 3)

CREST: Development of System Software Technologies for post-Peta Scale High Performance Computing

2013	2014	2015	2016	2017
Round 1: 5 teams run				
Round 2: 5 teams run				
Round 3 : 4 teams run				

Tetsuya Sakurai, University of Tsukuba

Development of an Eigen-Supercomputing Engine using a Post-Petascale Hierarchical Model

Kengo Nakajima, University of Tokyo

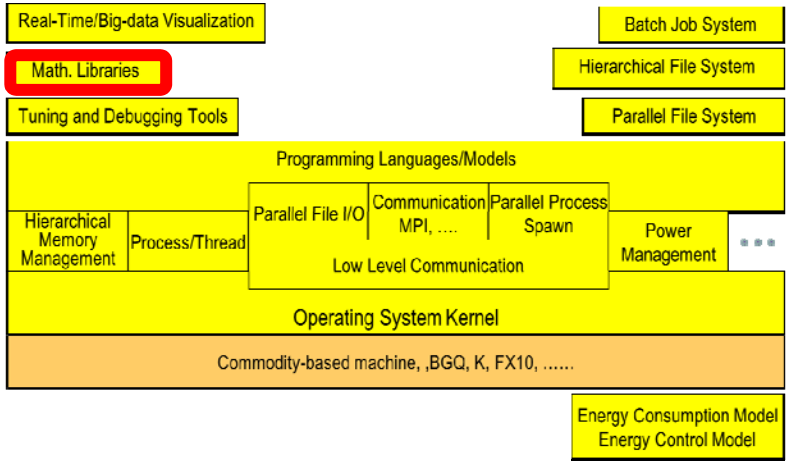
ppOpen-HPC

Ryuji Shioya, Toyo University

Development of a Numerical Library based on Hierarchical Domain Decomposition for Post Petascale Simulation

Katsuki Fujisawa, Chuo University

Advanced Computing and Optimization Infrastructure for Extremely Large-Scale Graphs on Post Peta-Scale Supercomputers



Associated Post Petascale Projects

- Univ. of Tsukuba
  - HA PACS(Highly Accelerated Parallel Advanced system for Computational Sciences) project (2011 to 2013, total $5 mil)
    - Objective: to investigate acceleration technologies for post-petascale computing and its software, algorithms and computational science applications, and demonstrate by building a prototype system
    - Design and deploy a GPGPU-based Cluster system
- Tokyo Institute of Technology – PI Satoshi Matsuoka
  - JSPS Grant-in-Aid for Scientific Research(S) "Billion-Way Parallel System Fault Tolerance"
    2011-15, Total $2 mil
    - Collaborators Franck Cappello (ANL), Bronis de Spinski (LLNL
  - MEXT – Tokyo Tech "Ultra Green Supercomputing"
    2011-15 Total $3 mil
    - **TSUBAME-KFC (TSUBAME3.0 Prototype)**
  - JST CREST "Extreme Big Data" 2013-2017 Total $3mil

Two Big Data CREST Programs (2013-2020) ~$60 mil

Advanced Core Technologies for Big Data Integration

Research Supervisor: Masaru Kitsuregawa
Director General, National Institute of Informatics

Advanced Application Technologies to Boost Big Data Utilization for Multiple-Field Scientific Discovery and Social Problem Solving

Research Supervisor: Yuzuru Tanaka
Professor, Graduate School of Information Science and Technology, Hokkaido University

CREST Big Data Projects circa 2014

(blue = big data application area)

Advanced Core Technologies for Big Data Integration

- Establishment of Knowledge-Intensive Structural Natural Language Processing and Construction of Knowledge Infrastructure
- Privacy-preserving data collection and analytics with guarantee of information control and its application to personalized medicine and genetic epidemiology
- **EBD: Extreme Big Data – Convergence of Big Data and HPC for Yottabyte Processing**
- Discovering Deep Knowledge from Complex Data and Its Value Creation
- Data Particlization for Next Generation Data Mining
- Foundations of Innovative Algorithms for Big Data
- Recognition, Summarization and Retrieval of Large-Scale Multimedia Data
- The Security Infrastructure Technology for Integrated Utilization of Big Data

Advanced Application Technologies to Boost Big Data Utilization for Multiple-Field Scientific Discovery and Social Problem Solving

- Development of a knowledge-generating platform driven by big data in drug discovery through production processes.
- Innovating "Big Data Assimilation" technology for revolutionizing very-short-range severe weather prediction
- Establishing the most advanced disaster reduction management system by fusion of real-time disaster simulation and big data assimilation
- Exploring etiologies, sub-classification, and risk prediction of diseases based on big-data analysis of clinical and whole omics data in medicine
- Detecting premonitory signs and real-time forecasting of pandemic using big biological data
- Statistical Computational Cosmology with Big Astronomical Imaging Data

TSUBAME3.0 : Convergent Architecture 2016

- Under Design : Deployment 2016Q2
- High computational power: **~20 Petaflops, ~5 Petabyte/s Mem BW**
- Ultra high density: **~0.6 Petaflops DFP/rack (x10 TSUBAME2.0)**
- Ultra power efficient: **10 Gigaflops/W (x10 TSUBAME2.0, TSUBAME-KFC)**
  - Latest power control, efficient liquid cooling, **energy recovery**
- Ultra high-bandwidth network: **over 1 Petabit/s bisection**
  - **Bigger capacity than the entire global Intenet (several 100Tbps)**
- Deep memory hierarchy and ultra high-bandwidth I/O with NVM
  - **Petabytes of NVM, several Terabytes/s BW, several 100 million IOPS**
  - **Next generation "scientific big data" support**
- Advanced power aware resource mgmy, high resiliency SW/HW co-design, Cloud VM & container-based dynamic deployment…
- In less than 40 racks, less than 1MW operation power

TSUBAME-KFC
(Kepler Fluid Cooling)

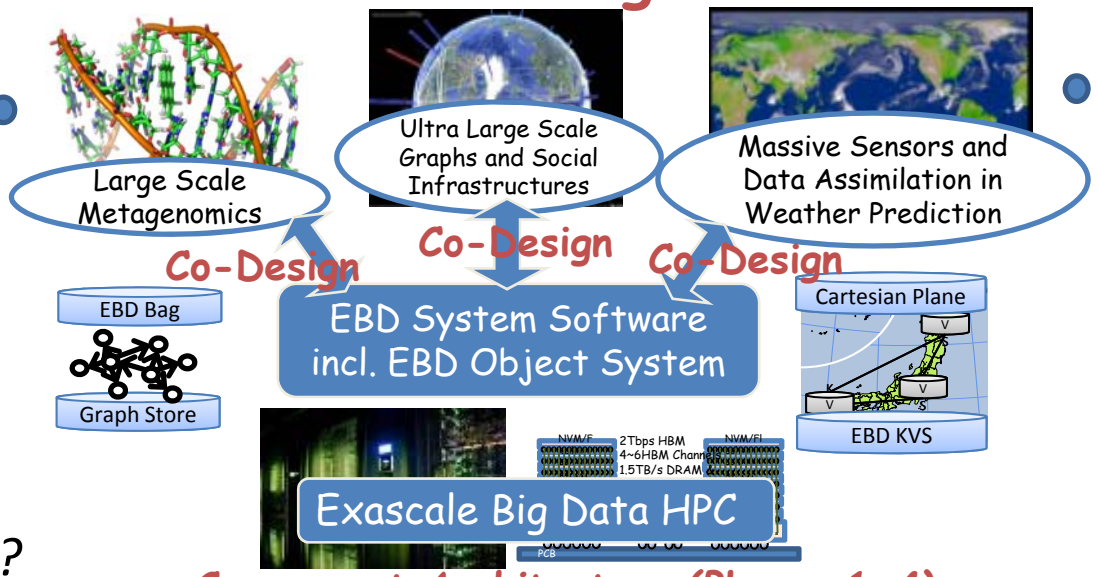
A TSUBAME3.0 prototype system
with advanced next gen cooling
40 compute nodes are oil-submerged
1200 liters of oil (Exxon PAO ~1 ton)
#1 Nov. 2013 Green 500!!

Single Node	5.26 TFLOPS DFP
System (40 nodes)	210.61 TFLOPS DFP 630TFlops SFP
Storage (3SSDs/node)	1.2TBytes SSDs/Node Total 50TBytes ~50GB/s BW

Tokyo Tech.
JST-CREST “Extreme Big Data” Project (2013-2018)

Future Non-Silo Extreme Big Data Scientific Apps

Given a top-class supercomputer, how fast can we accelerate next generation big data over Clouds?



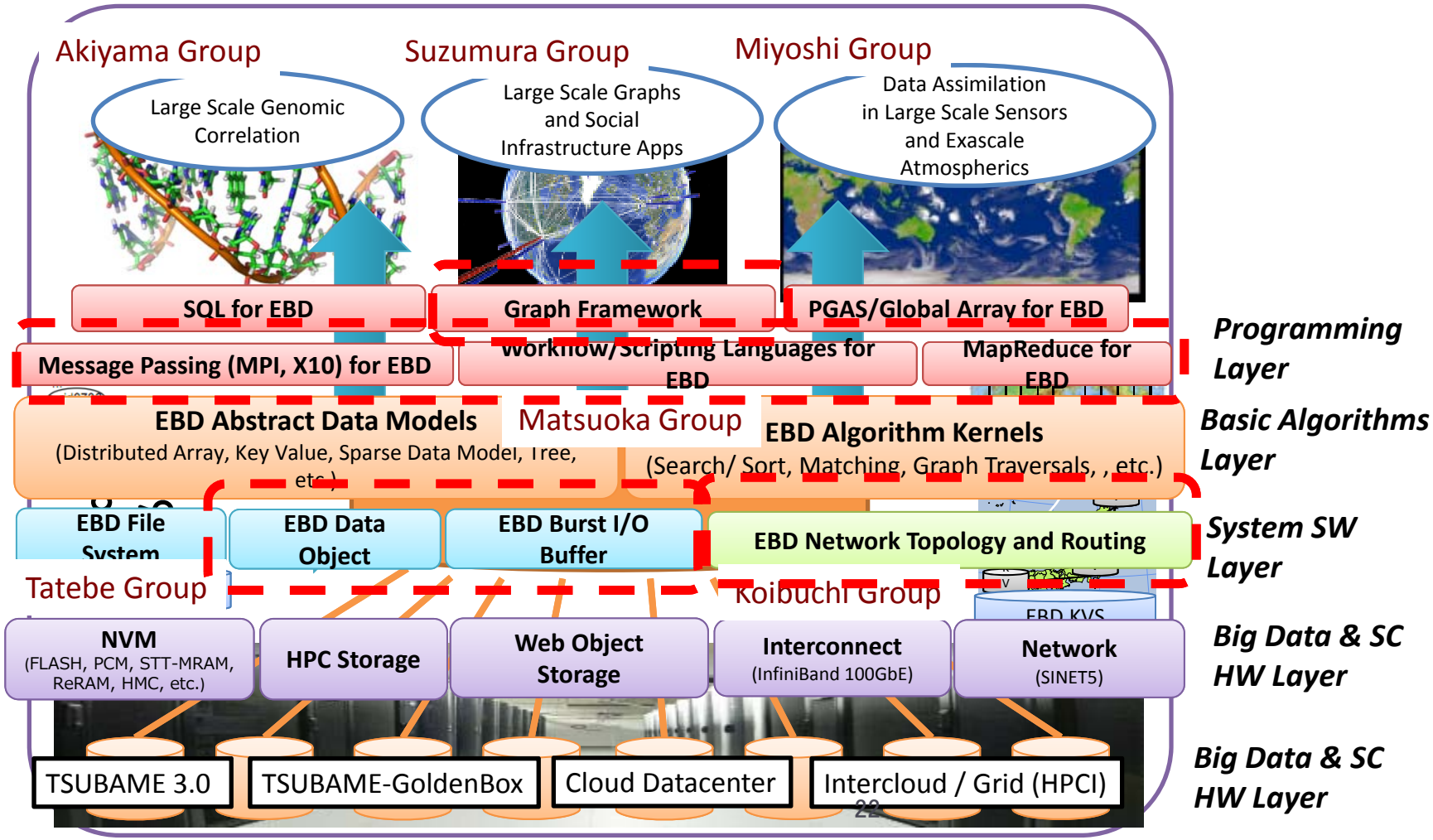
What are the issues? Architectural, algorithmic, and system software evolution?

Convergent Architecture (Phases 1~4)
Large Capacity NVM, High-Bisection NW

Cloud IDC
Very low BW & Efficiency
Highly available, resilient

Supercomputers
Compute&Batch-Oriented
More fragile

100,000 Times Fold EBD "Convergent" System Architecture
Defining the software stack via Co-Design



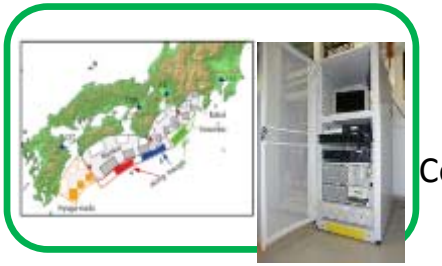
Proposed Big Data and HPC Convergent Infrastructure
=> "Nationoal Big Data Science Virtual Institute" (Tokyo Tech GSIC)

(Objective)

- **Convergence of High Bandwidth and Large Capacity HPC facilities with "Big Data" currently processed managed by domain laboratories**
- **HPCI HPC Center => HPC and Big Data Science Center**
- **People convergence: domain scientists + data scientists + CS/Infrastructure => Big data virtual institute**

Present

Domain labs segregated data facilities
No mutual collaborations
Inefficient, not scalable with
Not enough data scientists



Convergence of top-tier HPC and Big Data Infrastructure

2013 TSUBAME2.5
Upgrade
5.7Petaflops

2016 TSUBAME3.0
Green&Big Data
HPCI Leading Machine
Ultra-fast memory network, I/O

Data Management
Big Data Storage
Deep Learning
SW Infrastructure

National Labs
With Data

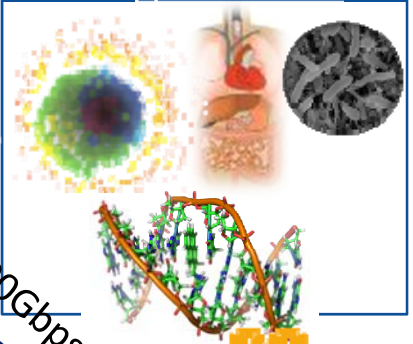
Mid-tier
"Cloud" Storage

Archival
Long-Term
Storage

Goal 100 Petabytes

SINET 5 400Gbps

Big Data Science Applications



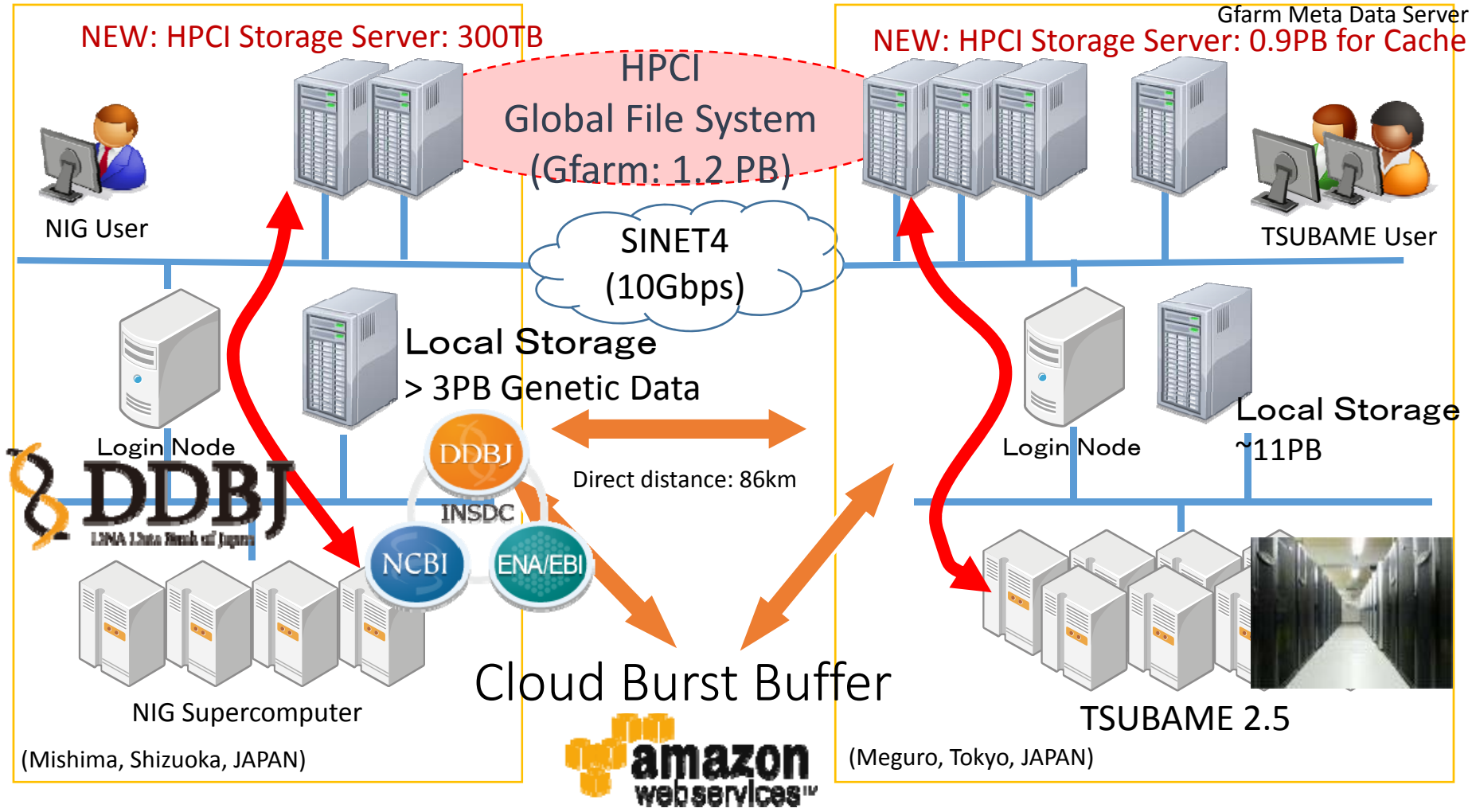
amazon web services

Virtual Multi-Institutional Data Science => People Convergence

HPCI Data Publication Prototype
GSIC and DDBJ @ National Institute of Genetics
& Amazon Storage Service (Cloud Burst Buffer)

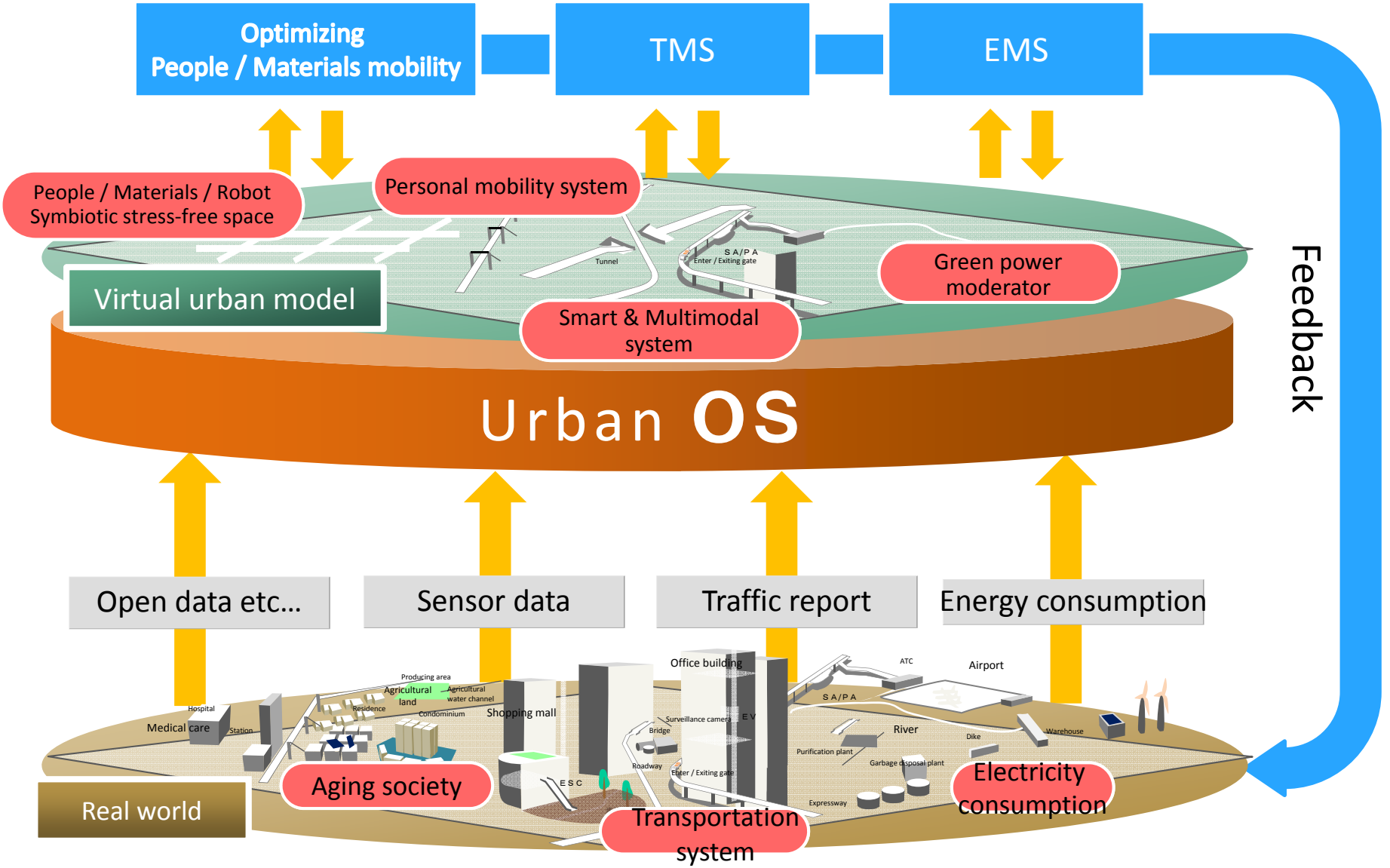
DDBJ Center, National Institute of Genetics
Data Generation and Provanance

GSIC Center, Tokyo Institute of Technology
HPCI Storage Cache for Data Publication



Univ. Kyushu Urban OS Project (2014-)
HPC+BD+IoT+Applied Math

The Urban OS to meet various needs and activate the society

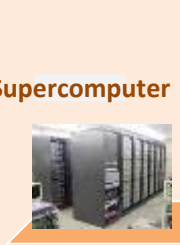


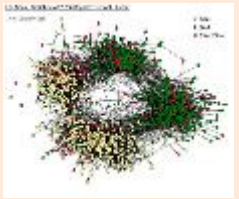
Kyushu University : Center of Innovation Program(COI)
Urban OS : Example - Transportation

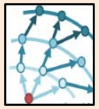
CESS

COI STREAM

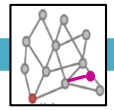
	Real World	Cyber Space	Real World
	Real World Digitization into Cyberspace	Large Scale Simulation & Optimization	Reflect the results
Macro Analysis	Digitizing City Traffic and Facilities Infrastructure	Bottleneck Analysis & Optimization	Long Term Infrastructure Plans
Mid Analysis	Digitizing roads, trains, cars, humans	Bottleneck Analysis & Optimization	Short-term traffic optimization
Micro Analysis	Digitize crowds and facilities	Optimized Flow Analysis	Compute personalized escape paths

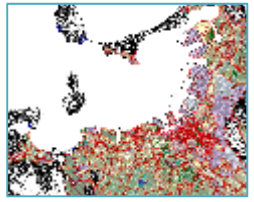




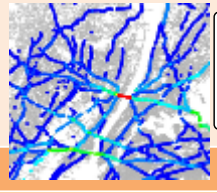


Macro Deep Analysis – Graphs, etc.

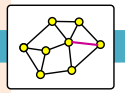




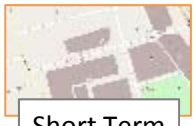
Optimized City Planning



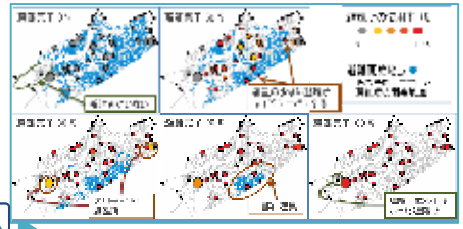
Mid On Demand Analysis



Controlling and Scheduling Traffic based on real-time events



Micro Real-time Analysis



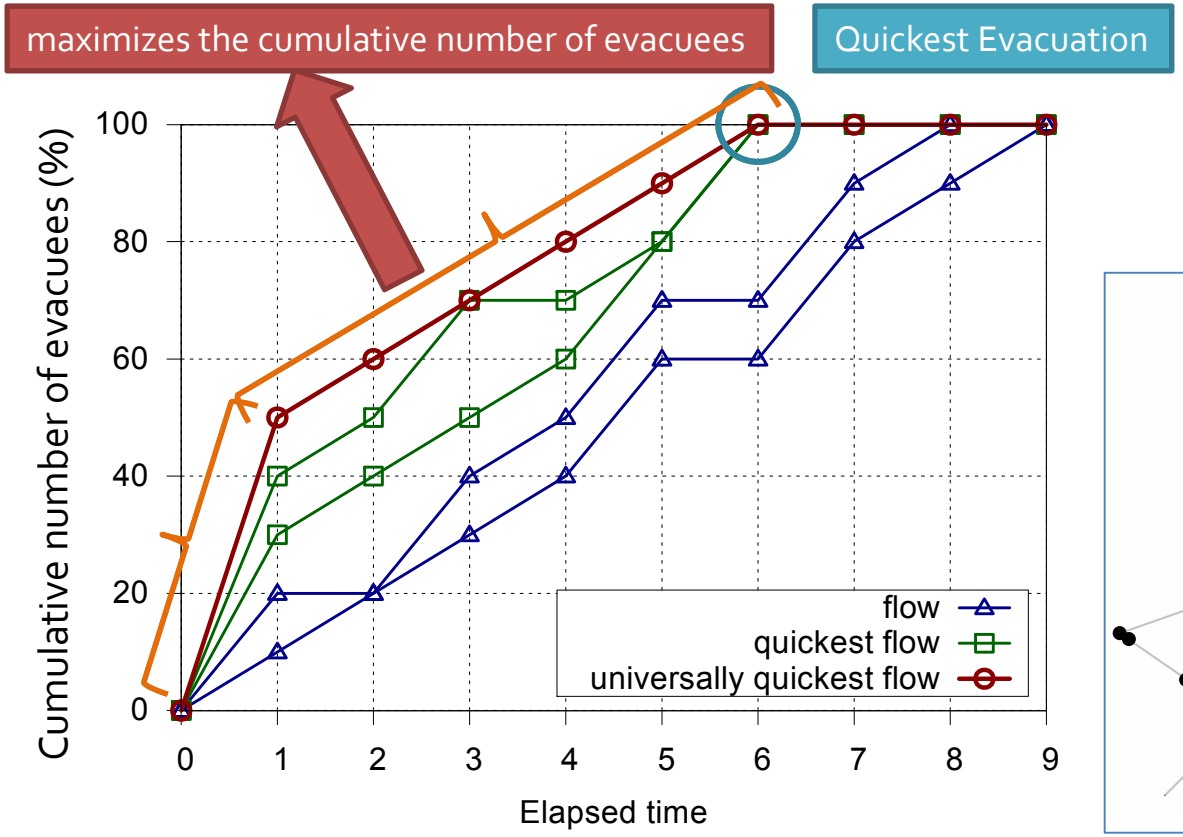
Real-time Emergency Evacuation Planning

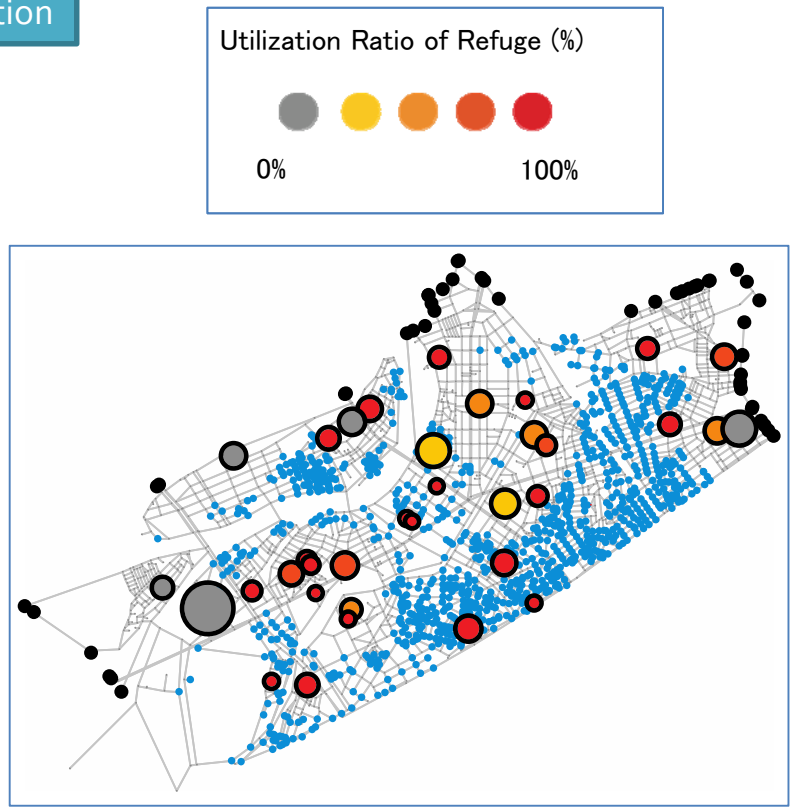
Real-time Emergency Evacuation Planning using the Universally Quickest Flow

- catastrophic disasters by massive earthquakes are increasing in the world, and disaster management is required more than ever

Universally Quickest Flow(UQF) → Not simulation But Optimization Problem

UQF simultaneously maximizes the cumulative number of evacuees at an arbitrary time.
Evacuation planning can be reduced to UQF of a given dynamic network.





TSUBAME4 2021~ K-in-a-Box (Golden Box)
Post-Moore Convergent Architecture

1/500 Size, 1/150 Power, 1/500 Cost, x5 DRAM+ NVM Memory

10 Petaflops, 10 Petabyte Hiearchical Memory (K: 1.5PB), 10K nodes

50GB/s Interconnect (200-300Tbps Bisection BW)

(Conceptually similar to HP “The Machine”)

Datacenter in a Box

Large Datacenter will become “Jurassic”

Tokyo Tech. GoldenBox Proto1 Post-Moore Convergent Architecture

- 36 Node Tegra K1, 11TFlops SFP
- ~700GB/s BW
- 100~700Watts
- Integrated mSata SSD, ~7GB/s I/O
- Ultra dense, Oil immersive cooling
- Same SW stack as TSUBAME

2022: x10 Flops, x10 Mem Bandwidth, silicon photonics, x10 NVM, x10 node density

IMPULSE: Initiative for Most Power-efficient Ultra-Large-Scale data Exploration

AIST IMPULSE

Non-Volatile Memory

- Voltage-controlled, magnetic RAM mainly for cache and work memories

High-Performance Logic

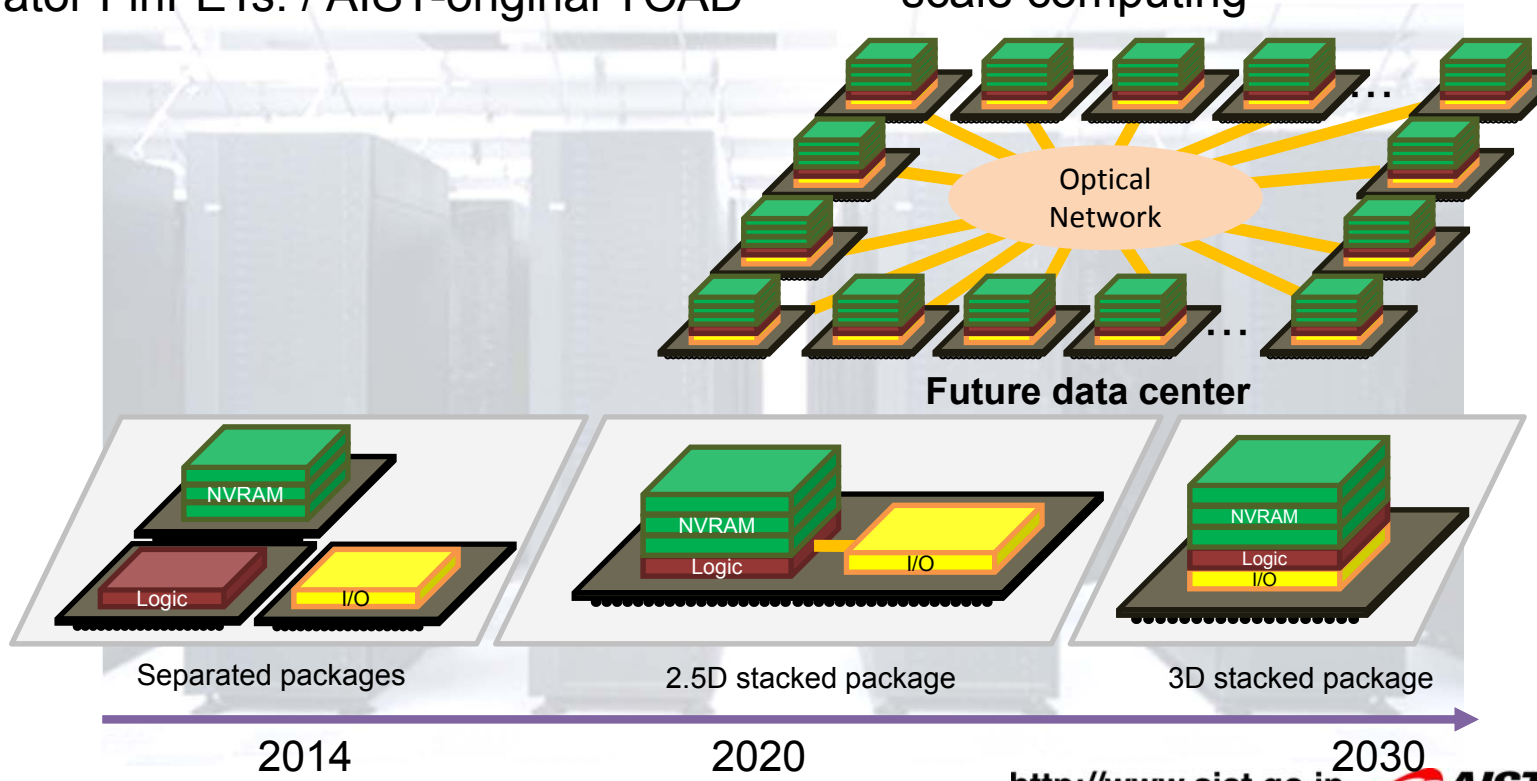
- 3D build-up integration of the front-end circuits including high-mobility Ge-on-insulator FinFETs. / AIST-original TCAD

Optical Network

- Silicon photonics cluster SW
- Optical interconnect technologies

Architecture

- Future data center architecture design / Dataflow-centric warehouse-scale computing



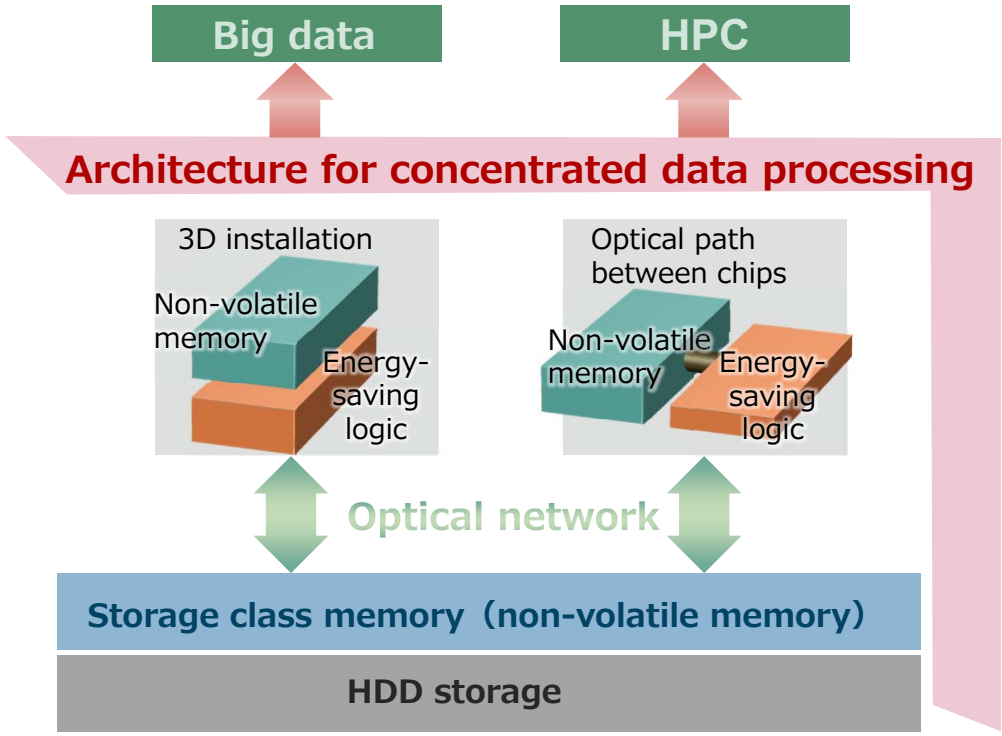
AIST's IMPULSE Program

IMPULSE | **STrategic AIST integrated R&D (STAR) program**
＊STAR program is AIST research that will produce a large outcome in the future.

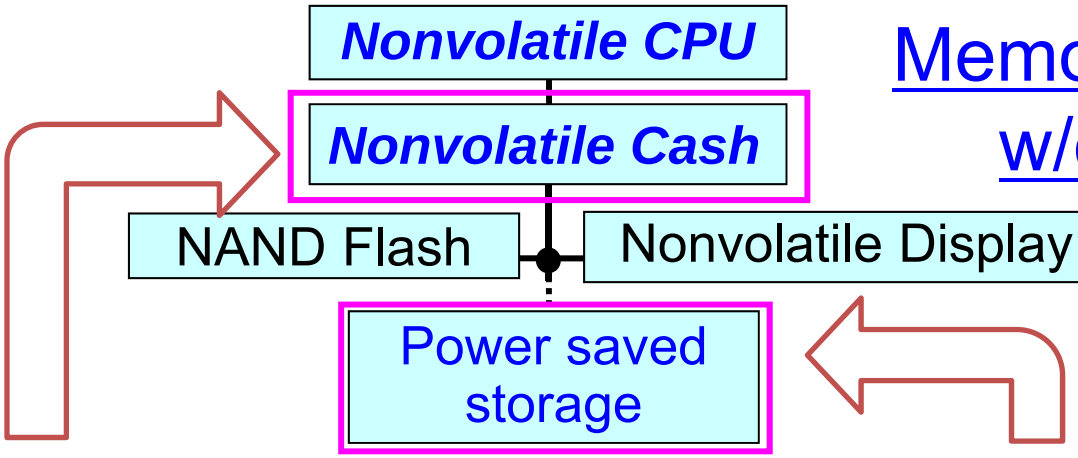
Initiative for Most Power-efficient Ultra-Large-Scale data Exploration

Creating a rich and eco-friendly society

- **High performance server module**
- **Energy-saving high-speed network**
- **Energy-saving large-capacity storage**

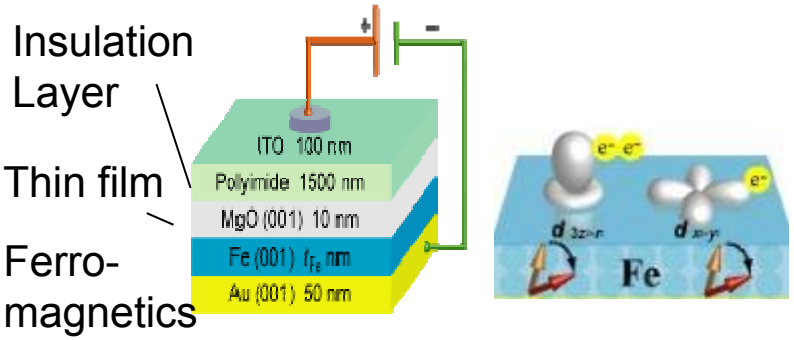


Voltage-controlled Nonvolatile Magnetic RAM



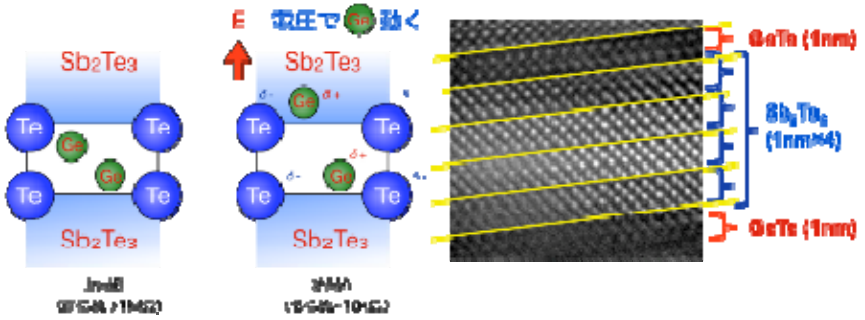
Memory keeping w/o power

Voltage Controlled Spin RAM



- voltage-induced magnetic anisotropy change
- Less than 1/100 rewriting power

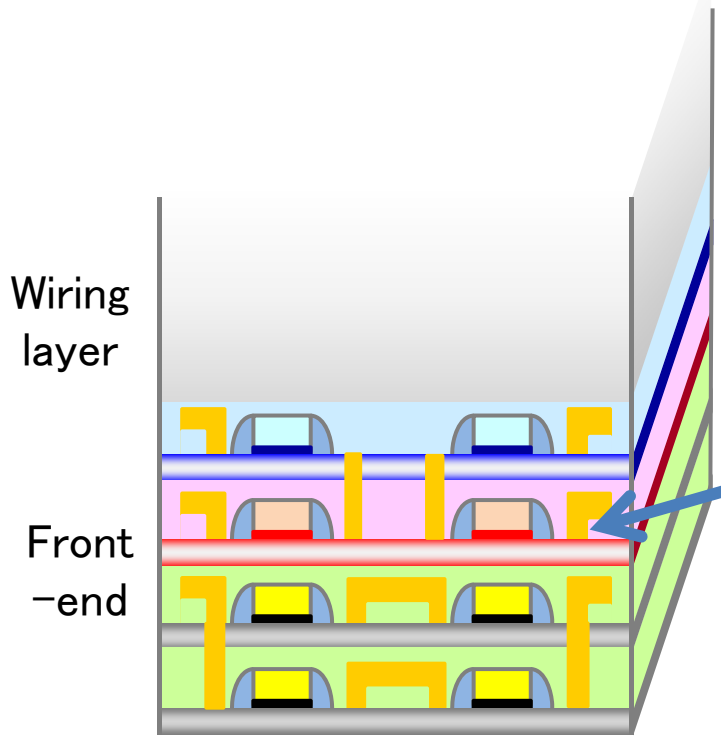
Voltage Controlled Topological RAM



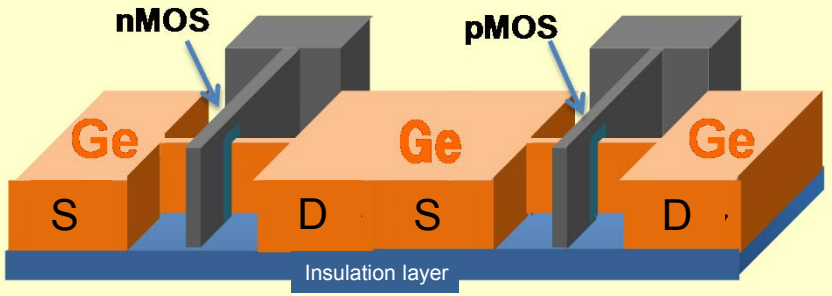
- Resistance change by the Ge displacement
- Loss by entropy: < 1/100

Low Power High-performance Logic

Front-end 3D integration



Ge Fin CMOS Tech.



- Low-power/high-speed by Ge
- Toward 0.4V – Ge Fin CMOS

- Dense integration w/o miniaturization
- Reduction of the wiring length for power saving
- Introduction of Ge and III-V channels by simple stacking process
- Innovative circuit by using Z direction

Optical Network Technology for Future Datacenters

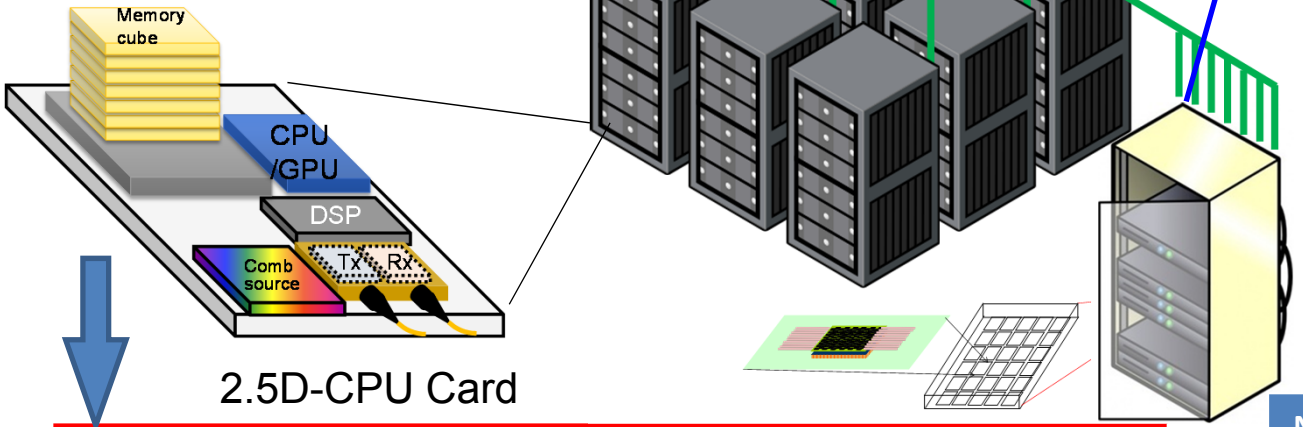
- Large-scale silicon photonics based cluster switches
- DWDM, multi-level modulation, highly integrated “elastic” optical interconnects
- Ultra-low energy consumption network by making use of optical switches

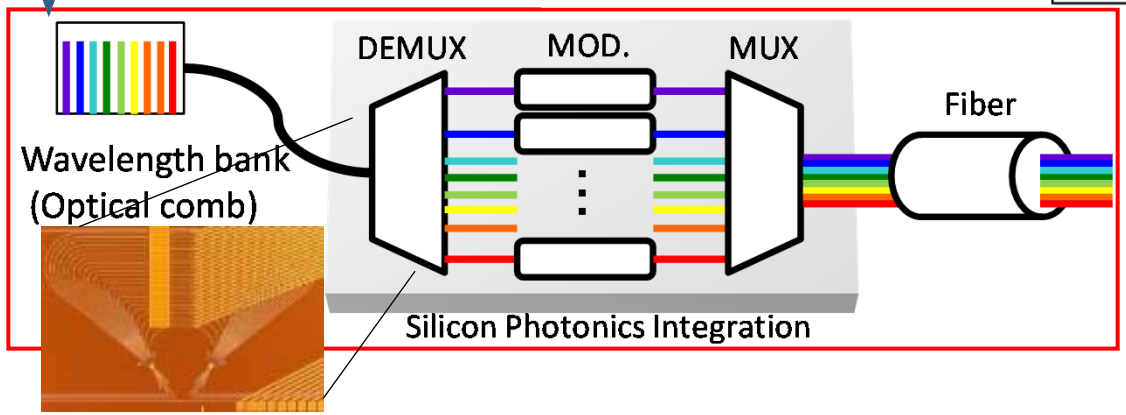
DWDM, multi-level modulation optical interconnects

Datacenter server racks

Silicon photonics cluster switches

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





Current electrical switches:
～130Tbps
↓
～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

Architecture for Big Data and Extreme-scale Computing

Warehouse Scale and data flow centric computing

1 - Single OS controls entire data center

2 - Guarantee the real time data processing by the priority controlled architecture for data flow