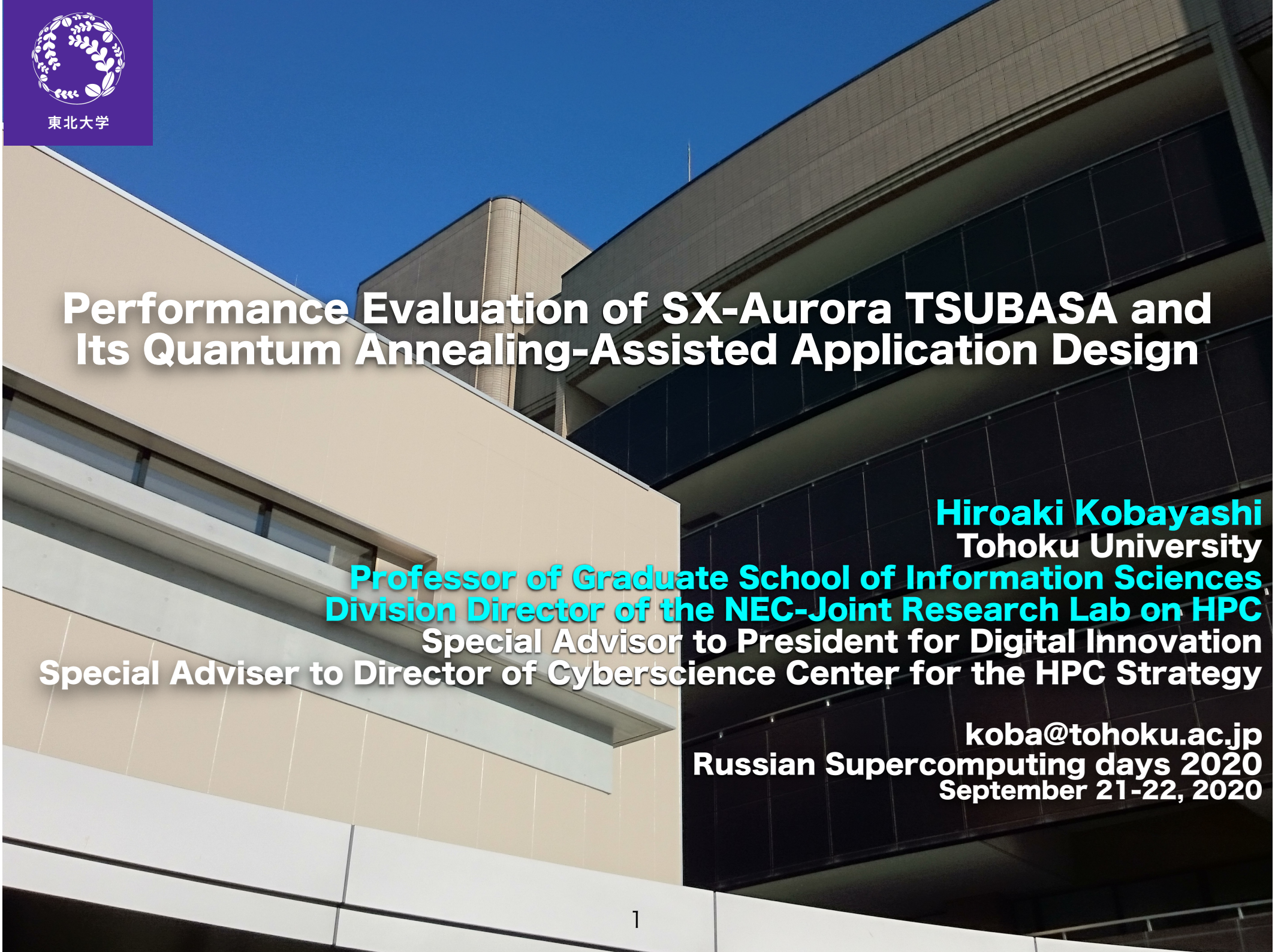


Performance Evaluation of SX-Aurora TSUBASA and Its Quantum Annealing-Assisted Application Design

Hiroaki Kobayashi
Tohoku University
Professor of Graduate School of Information Sciences
Division Director of the NEC-Joint Research Lab on HPC
Special Advisor to President for Digital Innovation
Special Adviser to Director of Cyberscience Center for the HPC Strategy

koba@tohoku.ac.jp
Russian Supercomputing days 2020
September 21-22, 2020

Today's Agenda

- ★ Toward the Realization of Society 5.0 in Japan
  - Realize the highly data-driven society to maximize performance, efficiency, safety, reliability and even comfortability of social systems.
  - Needs high-performance computing and data analysis platform to realize cyber-physical systems deployed in the society.
    - ✓ Examples of Digital Twins
- ★ SX-Aurora TSUBASA News Update
  - ✓ Performance Evaluation using HPCG, Himeno and HPL
  - ✓ SX-Aurora TSUBASA Roadmap
- ★ R&D of a Quantum Annealing-Assisted HPC Infrastructure
  - Realizes a Vector-Scalar and Quantum-Annealing Hybrid Simulation and Data Analysis Environment
  - Provides a transparent interface to deductive and inductive computing platforms over the vector-scalar and quantum-annealing hybrid environment
  - ✓ Application design and implementation of Data Clustering assisted by Quantum Annealing

Hiroaki Kobayashi, Tohoku University

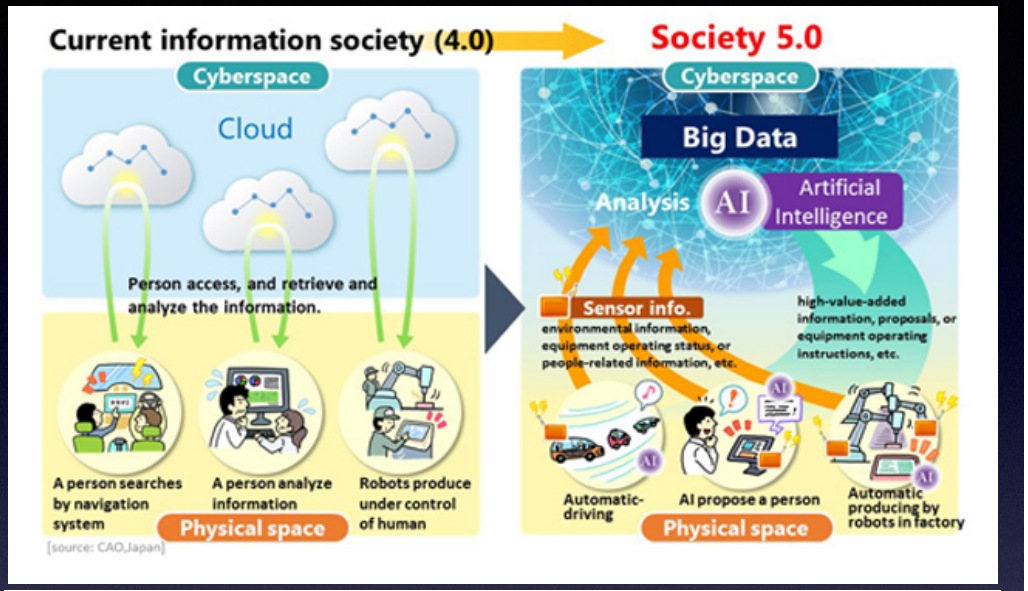
Toward the Realization of Society 5.0 in Japan: the Highly Data Driven Society Supported by Cyber-Physical System

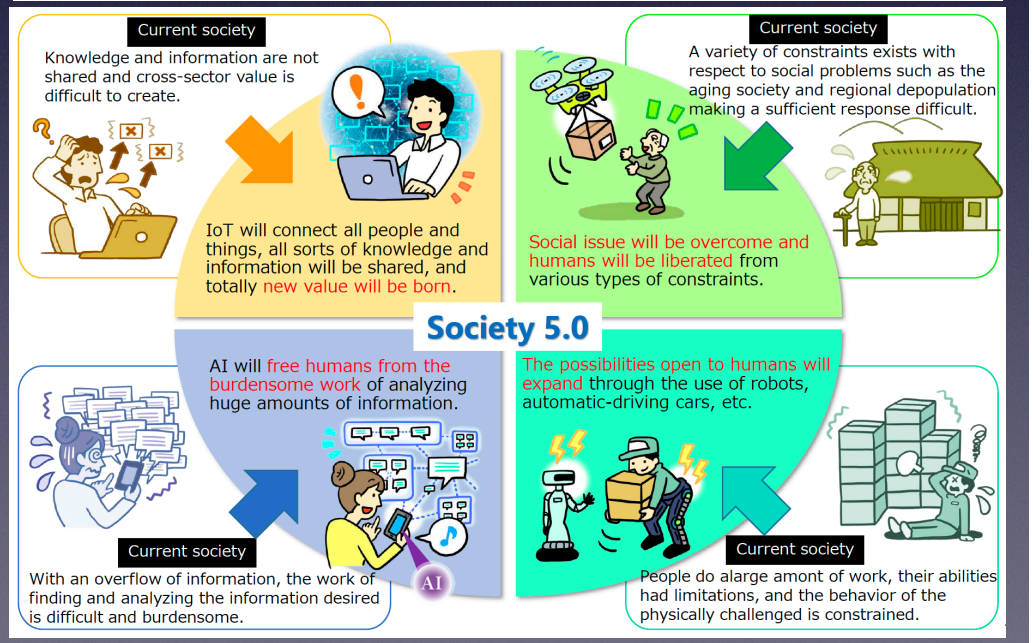
★ What is Society 5.0

- ✓ A human-centered society that balances economic advancement with the resolution of social problems by a system that highly integrates cyberspace and physical space.
- ✓ Highly data-driven society

★ Key Infrastructures and technologies to support the Society 5.0 world

- ✓ Cyber-Physical System, Close Interaction and convergence between Physical Space and Cyber Space, is a key infrastructure of Society 5.0
- ✓ High performance simulation is used to realize a digital twin of a real system
- ✓ High performance data analysis by AI/ML, exploits higher-order information from a huge amount of the cyber data (by simulation) and physical data (by IoT), and controls the cyber-systems and real-systems to maximize values, productivities, sustainability, safety… of any kinds of social activities, life, as well as advances in engineering and science.



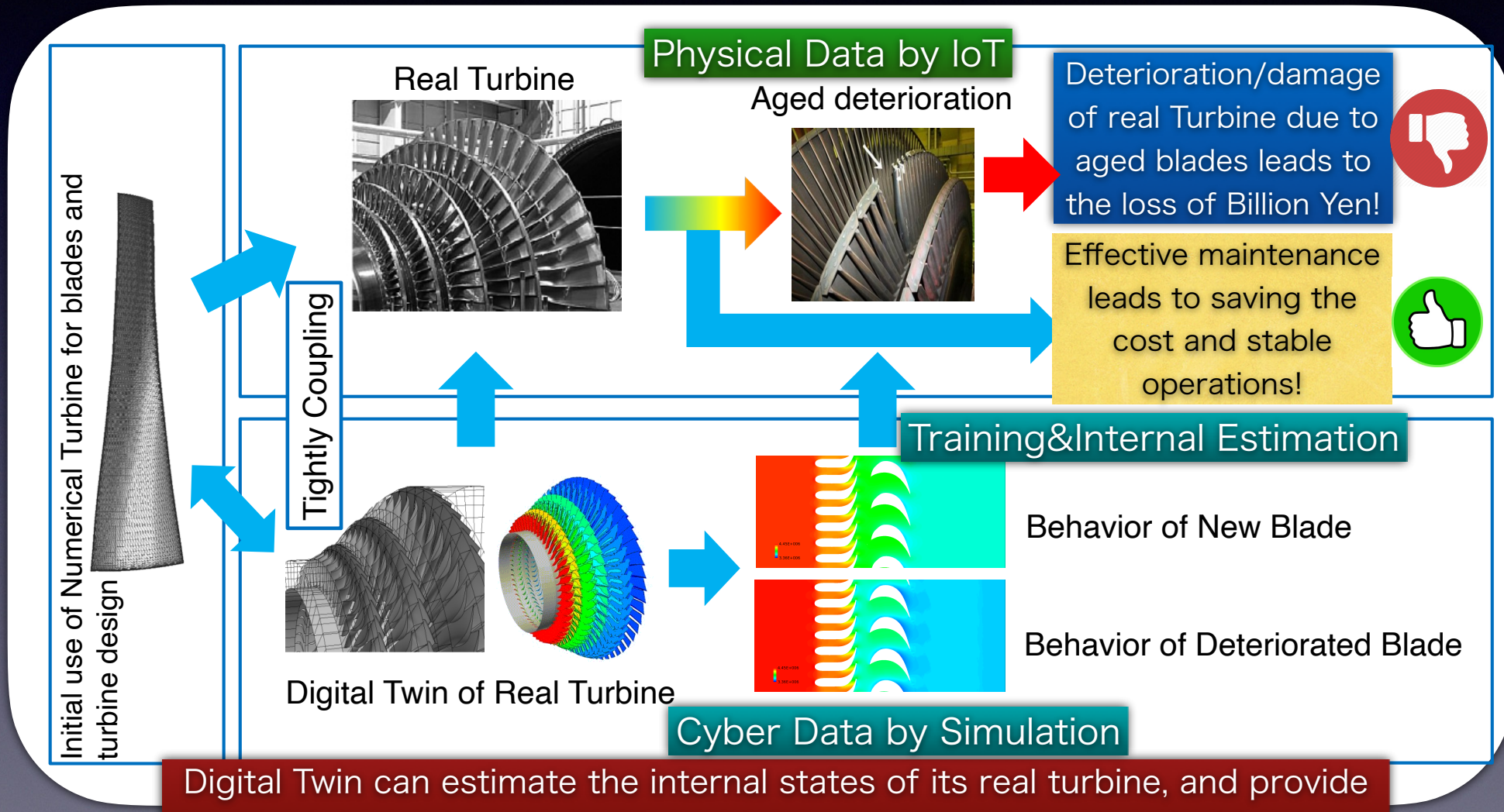


Source: Cabinet office of Japan

High Performance Computing is a Fundamental Infrastructure for Cyber-Physical System

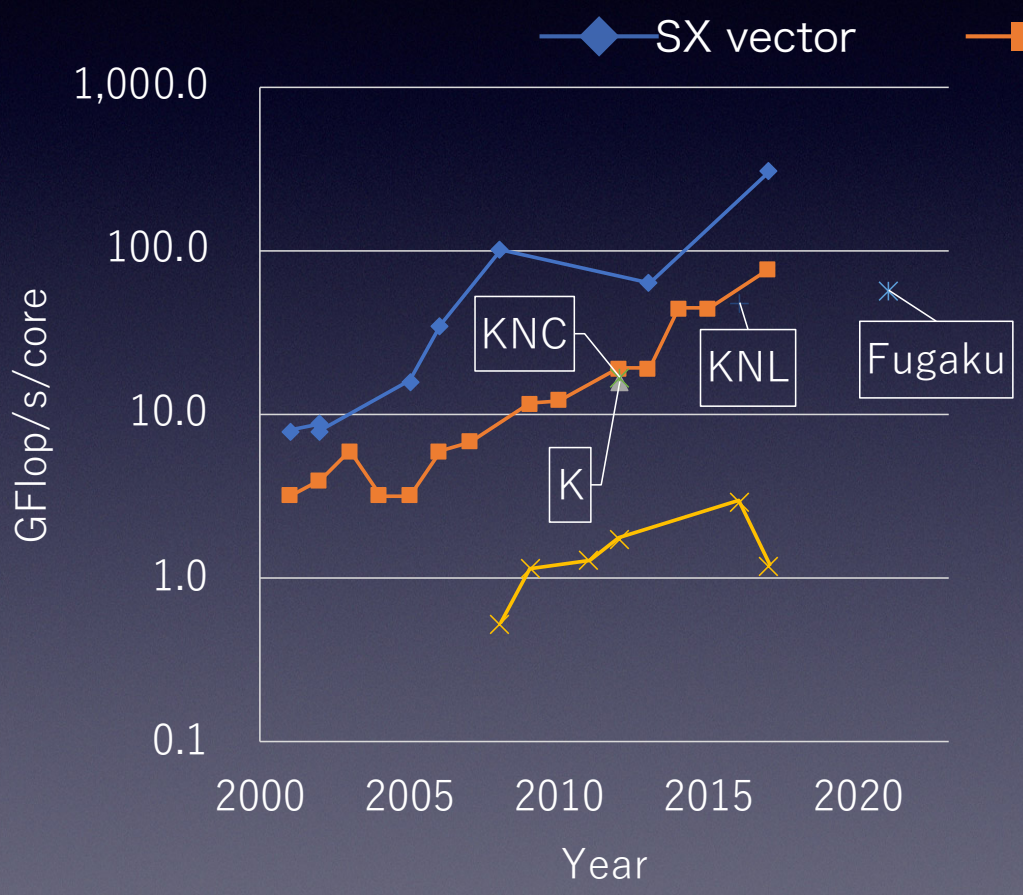
★ Digital Twin of Real Systems

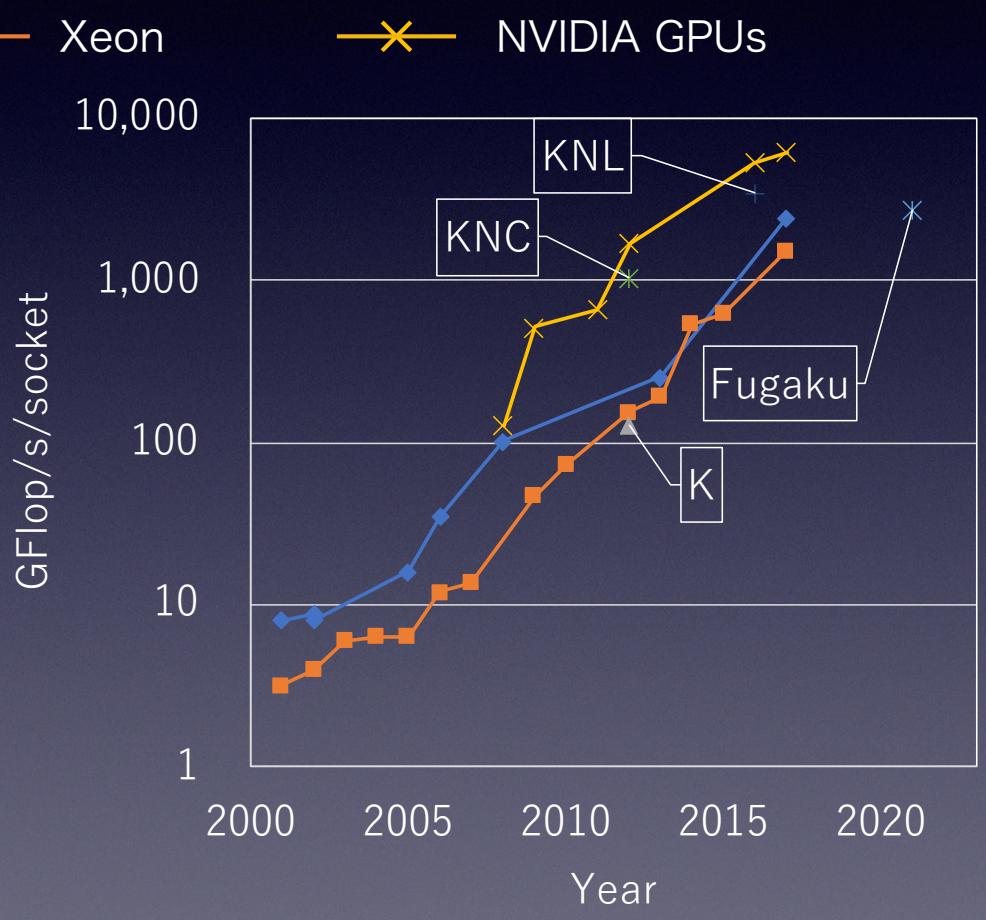
✓ Tightly-Coupled High Performance Simulation and Big Data Analysis



Digital Twin can estimate the internal states of its real turbine, and provide the information of effective maintenance to avoid serious incident!

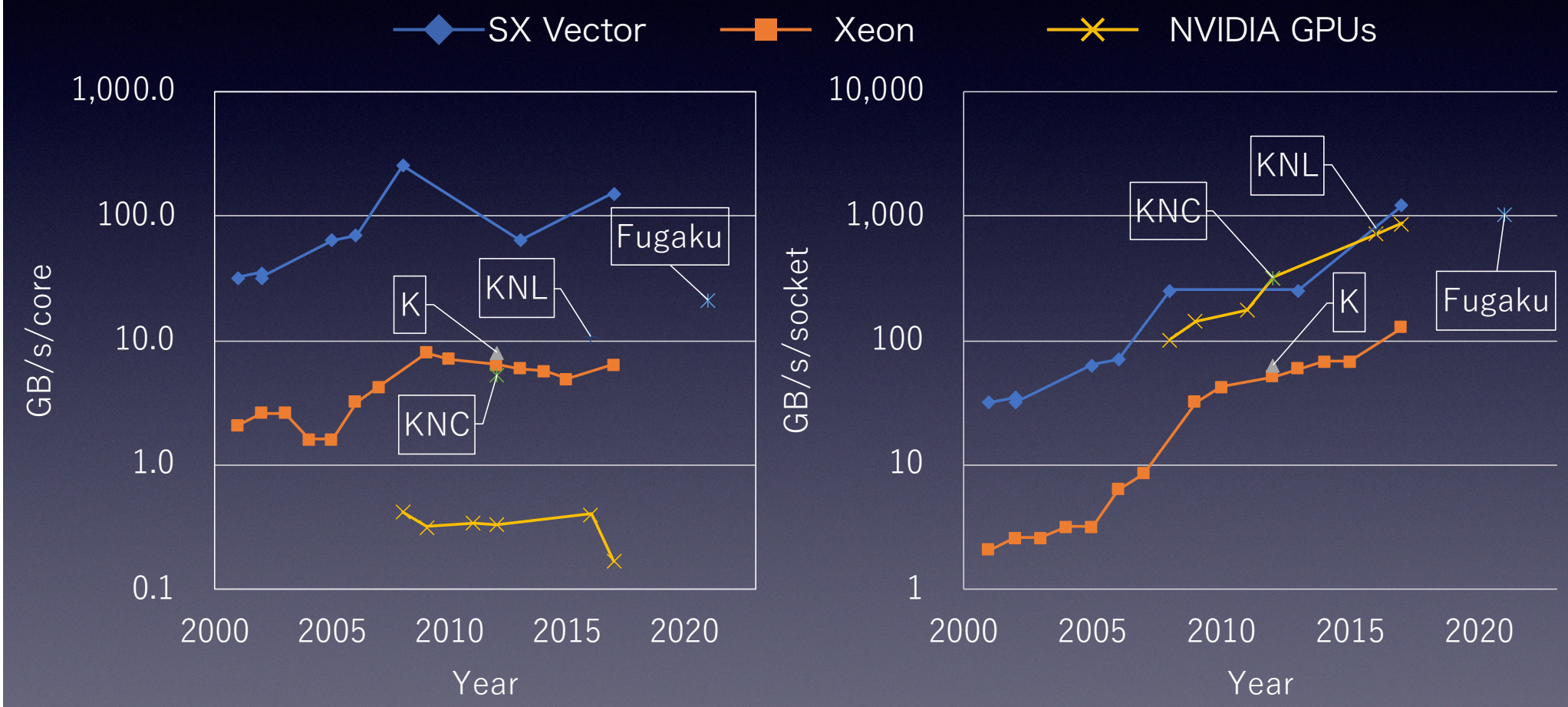
Vector Computing is common for boosting core/socket performance!





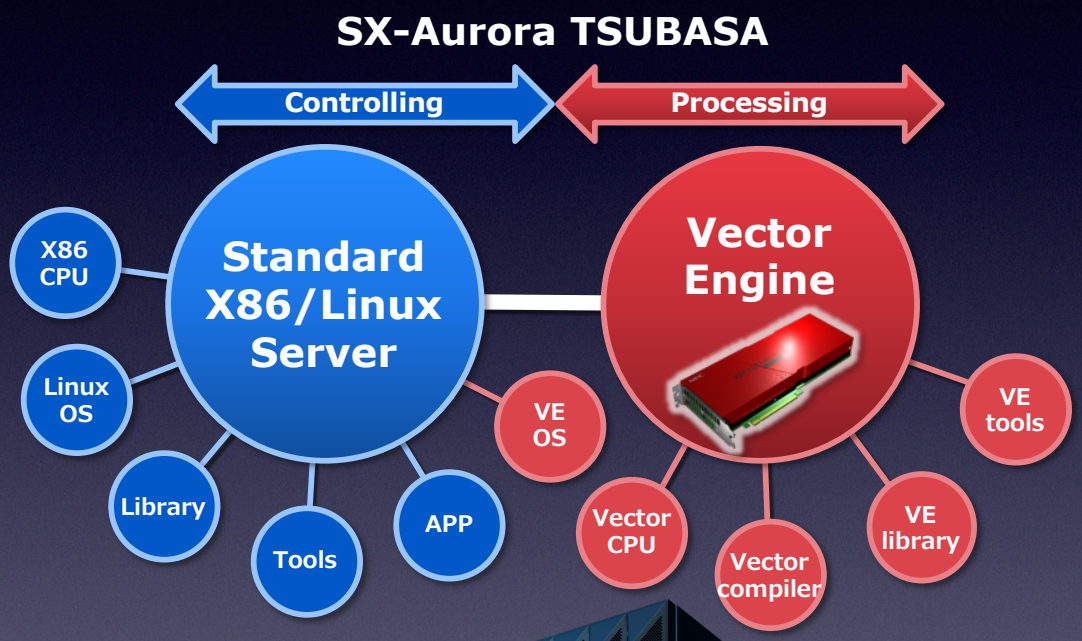
TOHOKU UNIVERSITY

Yes, Vector Computing is common, but memory bandwidth is a key for high sustained performance



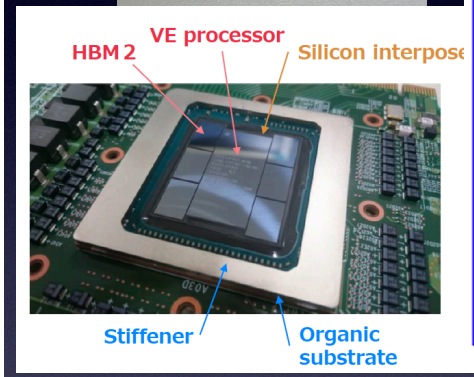
Why Vector System: SX-Aurora-TSUBASA?
~Balanced Architecture for High Sustained Performance~

Two types of balancing: computing performance and memory performance, and standardization and customization

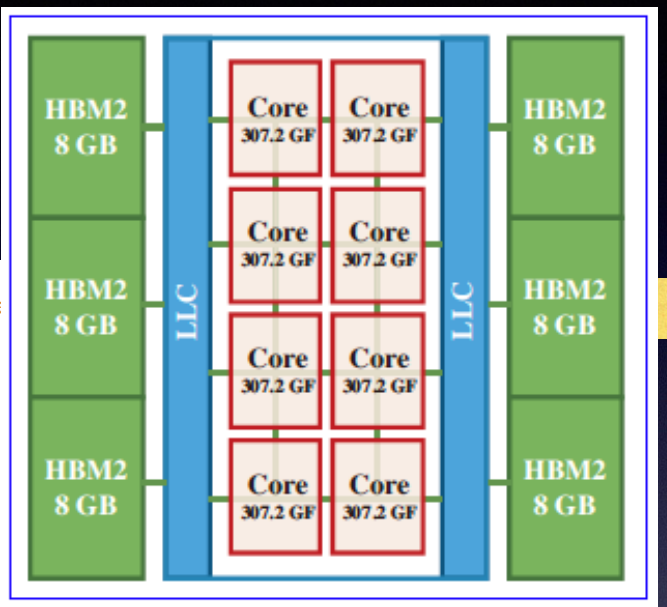


- ★ Customization for realization of the balanced vector architecture for memory-intensive apps
  - ✓ Highest Mem. BW
  - ✓ Largest Single Core Performance
- ★ Standardization for realization of the user-friendly environment and control-intensive apps.
  - ✓ x86 Linux Environment
  - ✓ New execution model centralized on vector computing

Hardware Specification of SX-Aurora TSUBASA (1st Gen in 2018)

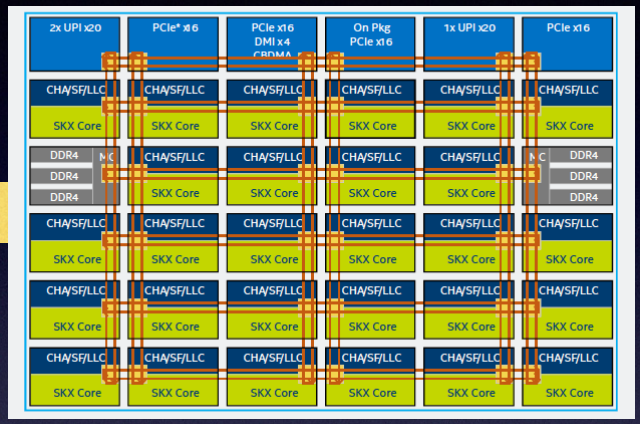


SX Vector Processor



PCIe

X86 Processor(Xeon)

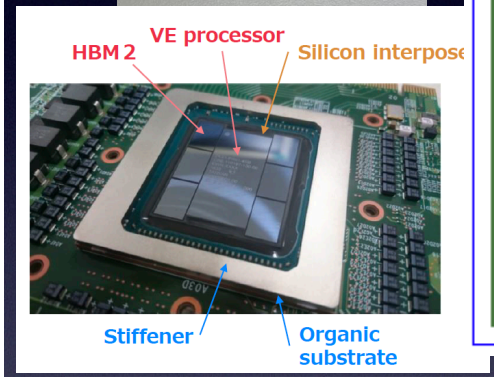


Vector Engine (VE)	Type 10B
Frequency	1.4GHz
Performance / Core	537.6 GF (SP), 268.8 GF (DP)
# cores	8
Performance / socket	4.30 TF (SP) 2.15 TF (DP)
Memory Subsystem	HBM2 8 GB x6
Memory Bandwidth	1.22 TB/s
Memory Capacity	48 GB

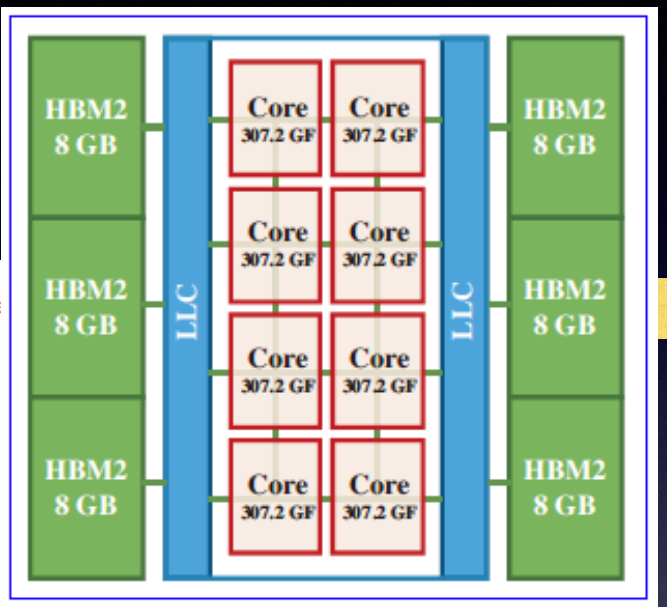
Vector Host (VH)	Intel Xeon Gold 6126
Frequency	2.60 GHz / 3.70 GHz (Turbo boost)
Performance/Core	166/236 GF(SP), 83/118 GF (DP)
No. of Cores	12
Performance/Socket	1,996/2,840 GF(SP) 998.4/1,420 GF(DP)
Memory Subsystem	DDR4-2666 DIMM 16GB x 6
Memory Bandwidth	128 GB/s
Memory Capacity	96 GB

Hiroaki Kobayashi, Tohoku University

Hardware Specification of SX-Aurora TSUBASA (2nd Gen in 2020)

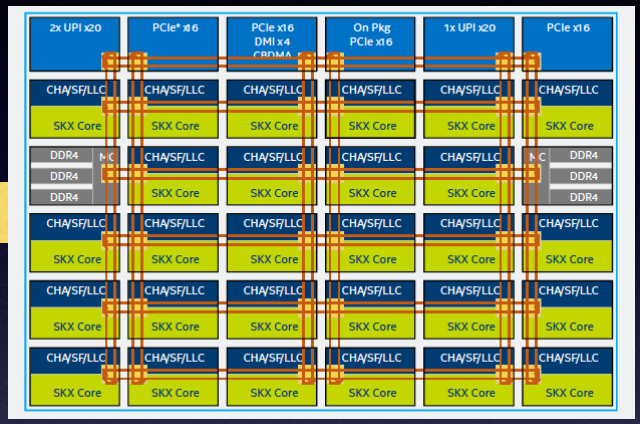


SX Vector Processor



PCIe

X86 Processor(Xeon)



Vector Engine (VE)	Type 20B
Frequency	1.6 GHz
Performance / Core	614 GF (SP), 307 GF (DP)
# cores	8
Performance / socket	4.91 TF (SP) 2.45 TF (DP)
Memory Subsystem	HBM2 8 GB x6
Memory Bandwidth	1.53 TB/s
Memory Capacity	48 GB

14% ↑

25% ↑

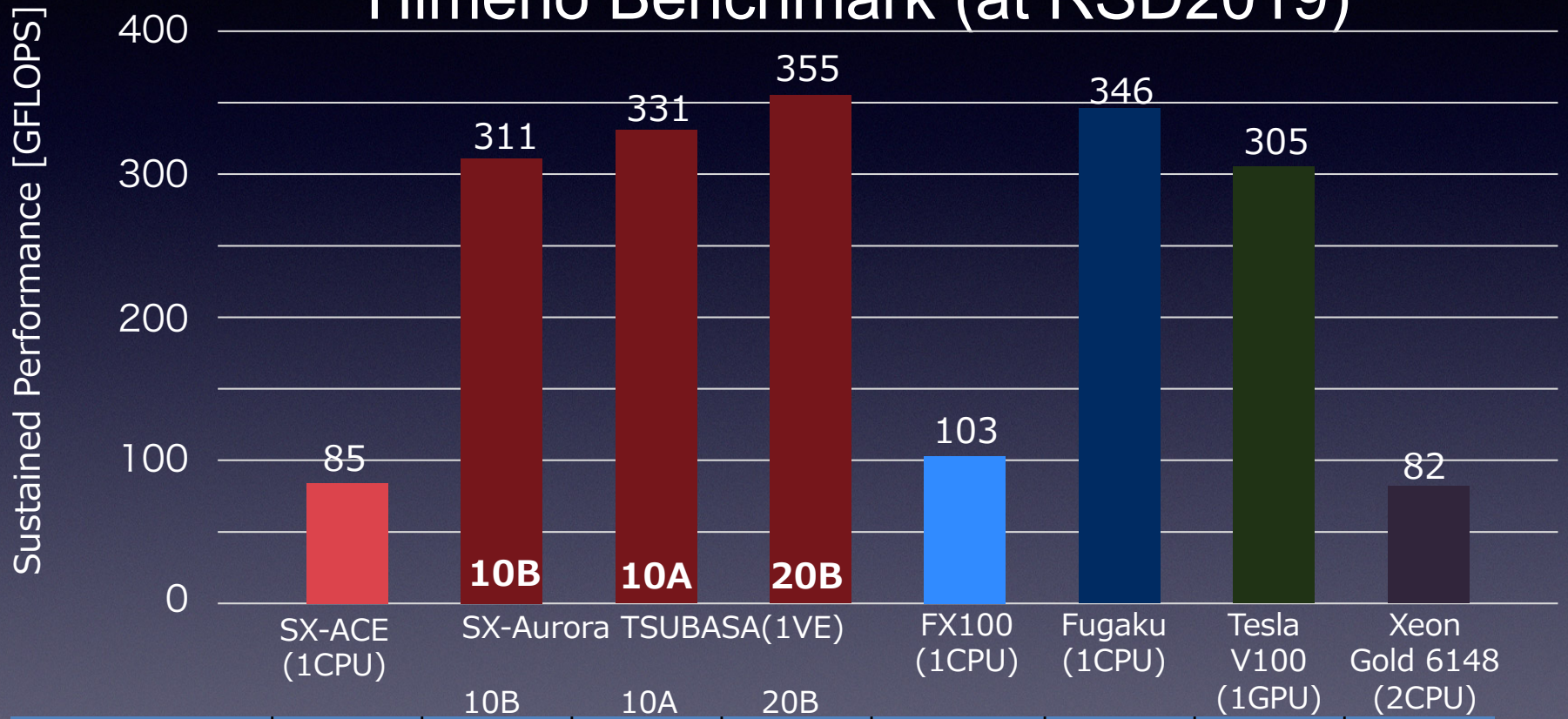
Vector Engine (VE)	Type 10B
Frequency	1.4GHz
Performance / Core	537.6 GF (SP), 268.8 GF (DP)
# cores	8
Performance / socket	4.30 TF (SP) 2.15 TF (DP)
Memory Subsystem	HBM2 8 GB x6
Memory Bandwidth	1.22 TB/s
Memory Capacity	48 GB

Benchmark Programs for Performance Evaluation

- ★ HIMENO benchmark
  - ✓ the Jacobi kernel with a 19-point Stencil on the 3D arrays that represents a memory-intensive application
- ★ HPCG benchmark
  - ✓ HPCG (High Performance Conjugate Gradient) measures the performance of a computer by solving the conjugate gradient method (CG method) with preprocessing using the Multi-Grid method for solving a simultaneous linear equation Ax = b with symmetric sparse matrices discretized by the finite element method. It is also a memory-intensive application

The Memory Performance: The dominant factor for a high-sustained performance

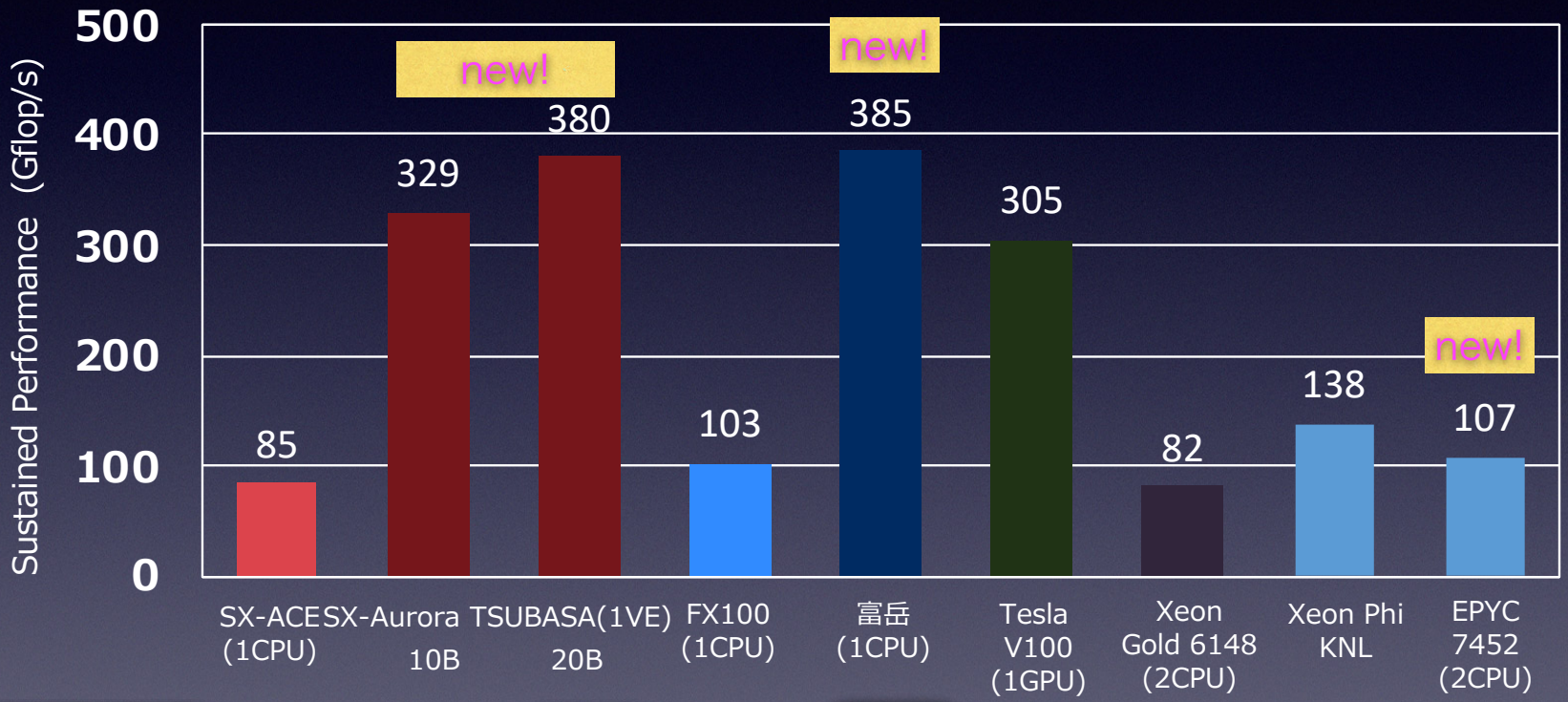
Himeno Benchmark (at RSD2019)



		10B	10A	20B				
Peak(DP)[TF]	0.256	2.15	2.45	2.45	1.12	2.7	7	3.07
MemBW[TB/s]	0.256	1.22	1.22	1.53	R:0.240 W:0.240	1.024	0.90	0.256

The Memory Performance: The dominant factor for a high-sustained performance

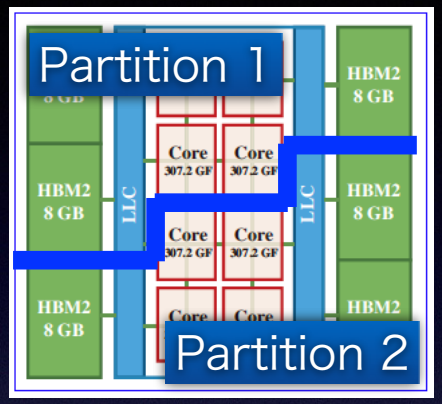
Himeno Benchmark (at RSD2020)

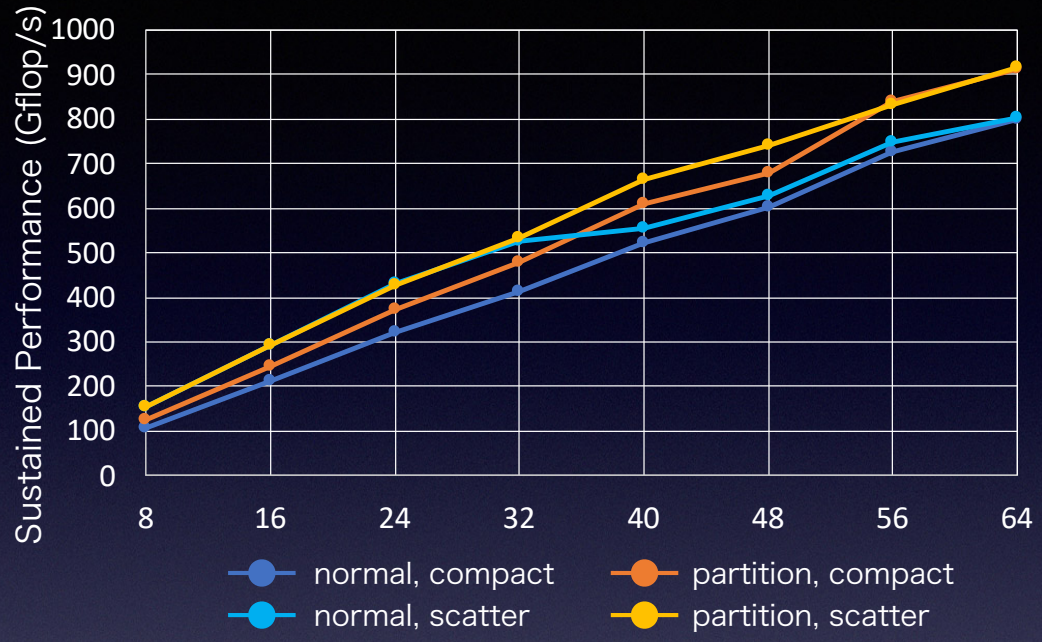


B/F	1.00	0.57	0.62	0.43	0.38	0.13	0.08	0.04	0.17
Efficiency (%)	33.2	15.3	15.5	9.2	14.3	4.4	2.7	4.0	4.5

HPCG Performance Results

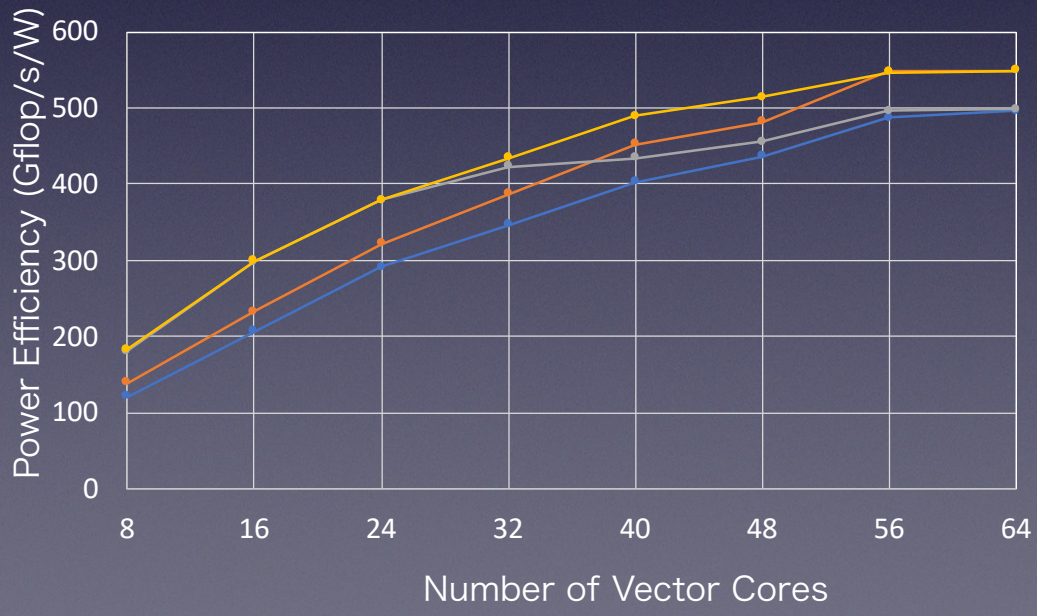
Partition mode



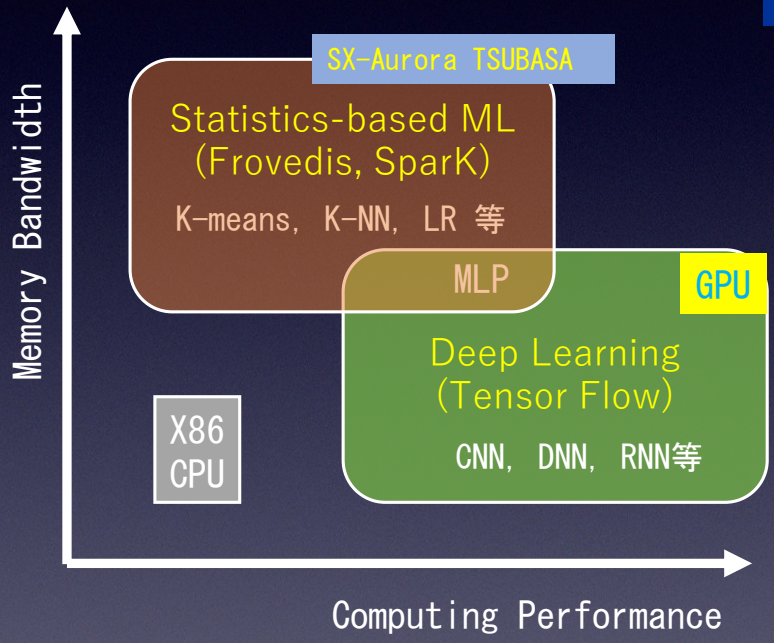


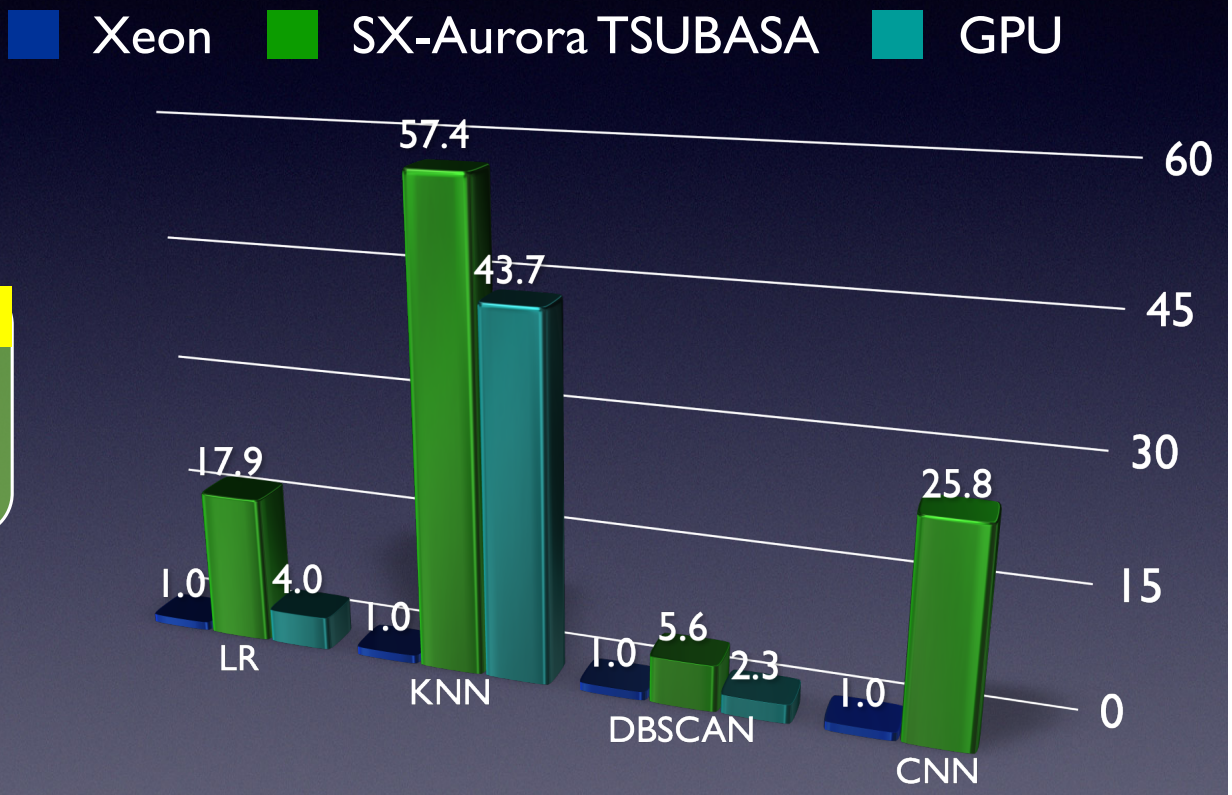
Compact mode

Scatter mode



SX-Aurora TSUBASA as an AI-ML Platform



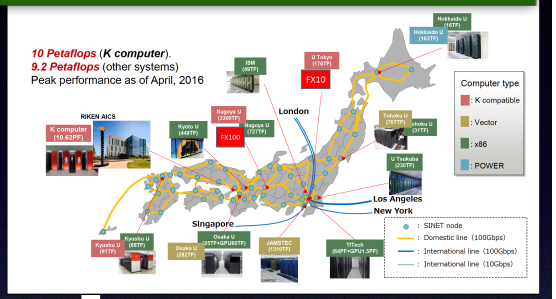


New Supercomputer System at Tohoku University

★Start Servicing in Oct. 2020

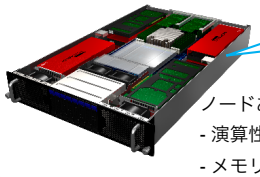
- Peak Performance of 1.8Pflop/s
- 20+ x performance-enhanced in 2022

HPCI for Nation-wide service



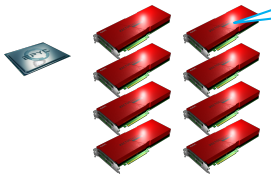
Vector Supercomputer SX-Aurora TSUBASA (2nd Gen VEs)

NEC SX-Aurora TSUBASA B4



- ノードあた
- 演算性能
- メモリ容
- メモリ帯

AMD EPYC 7402P x 1
NEC Vector Engine Type 20B



NEC Vector Engine Cores x 8
HBM2 Memory Module x 6

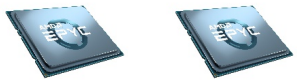
Tohoku Univ Campus net.

Interconnect Fabric (InfiniBand HDR)

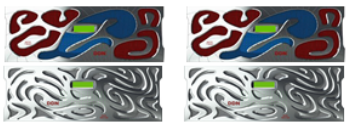
NEC LX 406Rz-2 x 68

- ノードあ
- 演算性
- メモリ
- メモリ帯

AMD EPYC 7702P x 2

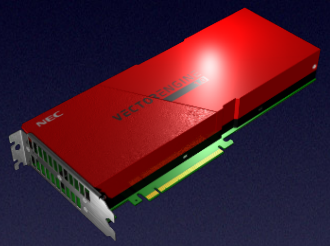


Storage 2PB
DDN SFA7990EX



X86 Cluster System(AMD EPYC 7720)

SX-Aurora TSUBASA Tick-Tock Roadmap



Memory bandwidth / VE

1st Gen
VE10
8C/2.45TF
1.22TB/s memory bandwidth

VE10E
8C/2.45TF
1.35TB/s memory bandwidth

2nd Gen
VE20
10C/3.07TF
1.53TB/s memory bandwidth

3rd Gen
VE30
2+TB/s memory bandwidth

New Architecture

4th Gen
VE40
2++TB/s memory bandwidth

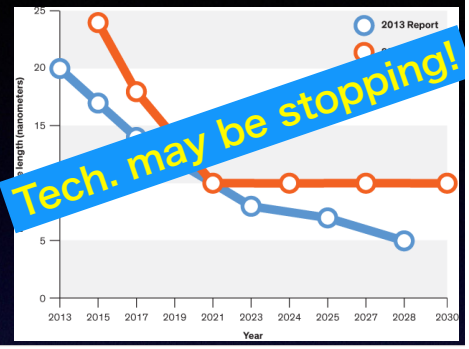
5th Gen
VE50
New Architecture

2019 2020 2021 2022 2023 2024

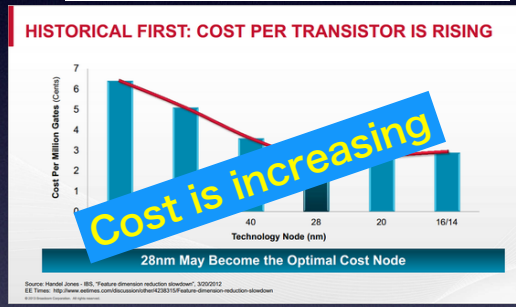
Source: ISC20 NEC Vender Showdown
Sep 21-22, 2020

TOHOKU UNIVERSITY

Challenges in Computer Systems Design: Scaling may be End, but Silicon is not End! And Use it Smart and Effective!



- ✓ We are facing the end of Moore's law due to the physical limitations, and the transistor cost is hard to reduce, however
- ✓ Tech. is slowing, cost is increasing, and efficiency is lowering!



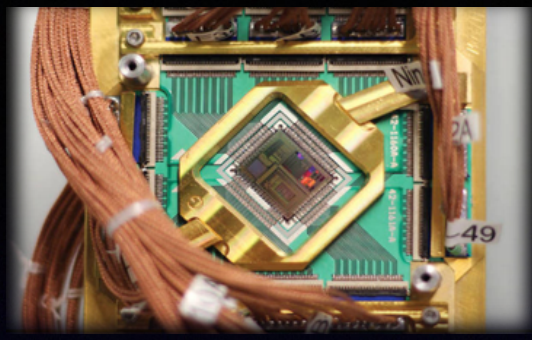
- ☆ Silicon is still fundamental constructing material for computing platforms just like plastic, steel and concrete for automobiles, buildings and home appliances.

Use precious silicon budget (+ advanced device technologies) to effectively design mechanisms that can maximize the sustained performance of individual applications.

Rank	Name	Cores	Rmax (Tflop/s)	Rpeak (Tflop/s)	Rmax/ Peak
1	Fugaku	7,299,072	415,530.00	513,854.70	80.87
2	Summit	2,414,592	148,600.00	200,794.90	74.01
3	Sierra	1,572,480	94,640.00	125,712.00	75.28
4	Sunway TaihuLight	10,649,600	93,014.60	125,435.90	80.87
	Tianhe	4,981,760	61,444.50	100,678.70	61.03
8	Frontera				.69

It's time to focus on Domain-Specific Architectures for computation-intensive, memory-intensive, I/O intensive, mixed-precision computing… etc applications to improve silicon/power efficiency, and their orchestration to satisfy the requirements from a wide variety of applications is required!

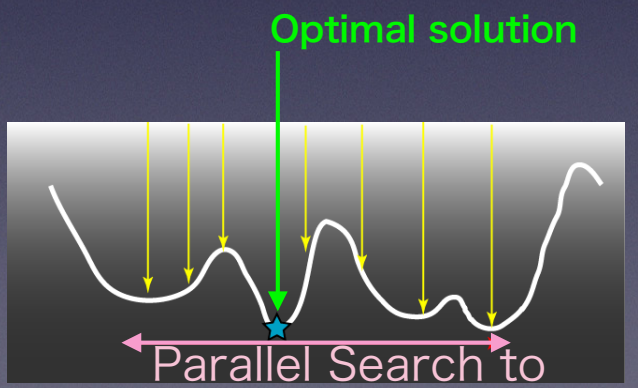
Quantum Computer: Emerging Domain Specific Architecture



- ★ Quantum computing is drawing much attention recently as an emerging technology in the era of post-Moore
  - ✓ In particular, quantum annealing machines are commercialized by the D-wave systems, and their applications are developed world-widely.
    - ✓ Google, NASA, Volkswagen, Lockheed, Denso…
  - ✓ The base model named the Ising model to design and implement the D-wave machines has been proposed by Prof. Nishimori et al of Tokyo Inst. Tech. In 1998.
- ★ The quantum annealing is a metaheuristic for finding the global minimum of a given objective function over a given set of candidate solutions (candidate states), by a physical process named quantum fluctuations

Source by D-Wave Sys.

Transverse magnetic field type quantum annealing Chip and System (D-Wave)



An ideal solver for combinatorial problems!

Toward Realization of Quantum Computing-Assisted HPC Infrastructure

★ Tohoku University has established an interdisciplinary priority research institute, named Q-HPC, for Quantum Computing-Accelerated HPC in 2018

★ We start a new 5-year research program named "R&D of Quantum Annealing-Assisted HPC Infrastructure", supported by MEXT, in collaboration with NEC and D-wave sys.

✓ provides transparent accesses to not only classical HPC resources but also Quantum Computing one in a unified fashion.

✓ Becomes an innovative infrastructure to develop next-generation applications in the fields of computational science, data sciences and their fusions

Innovative Applications in the fields of Computational Science, Data Science and their Fusion

QA-Assisted New Generation HPC infrastructure

Inductive Processing (for Data Science Apps.)

Deductive Processing (for Computational Science Apps.)

Tightly Coupling

QA Platform

QA-Assisted AI・ML Platform

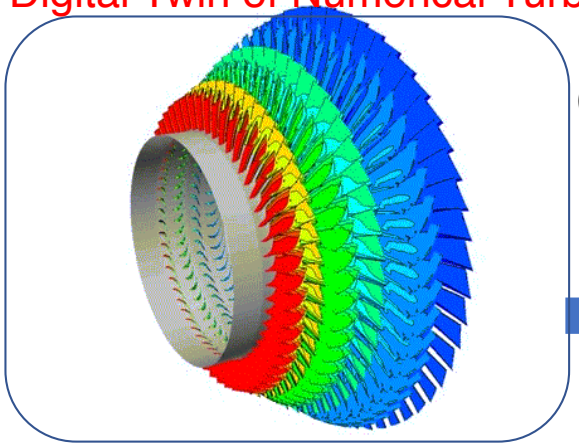
X86-Assisted Vector Computing Platform

D-Wave Machines and/or equvs.?

SX-Aurora TSUBASA and its Successors?

Target App: Realization of a Digital Twin for Real Turbine

Digital Twin of Numerical Turbine (Cyber Space)



Cyber Data

Parametric simulations results

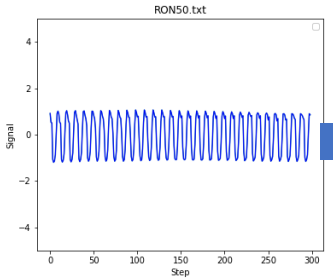
Simulation Database (SDB)

Pre-fault training and detection by AI/ML

Maintenance Planning

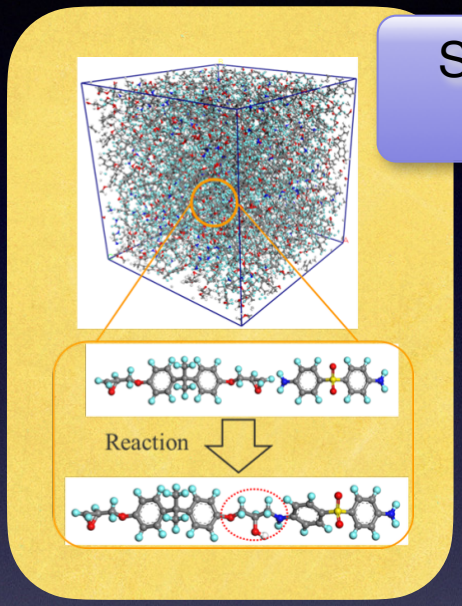
Real Machine (Physical Space)

Physical Data

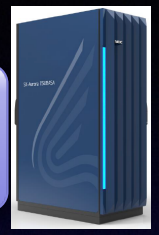


Measured acoustic data

Target App.: QA-Assisted Materials Integration System

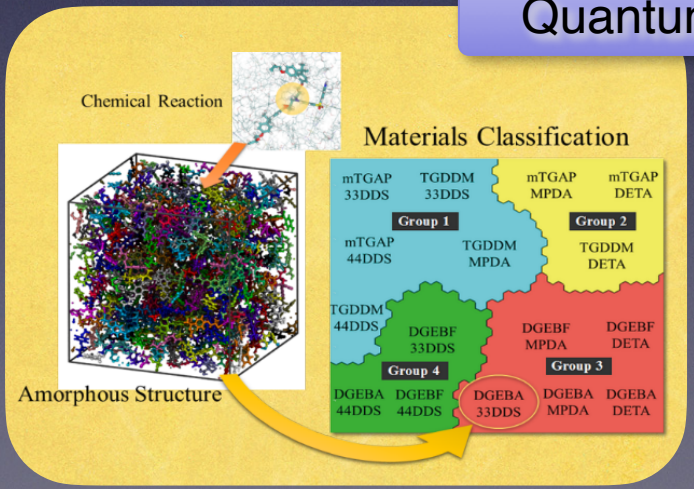


Simulation assisted by next-generation vector-type supercomputing

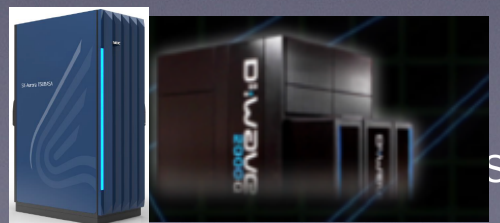


- ✓More accurate and faster reaction model incorporated into MD simulation for crosslinked network formation in thermosetting resins
- ✓Faster multiscale-simulation for predicting various thermomechanical properties

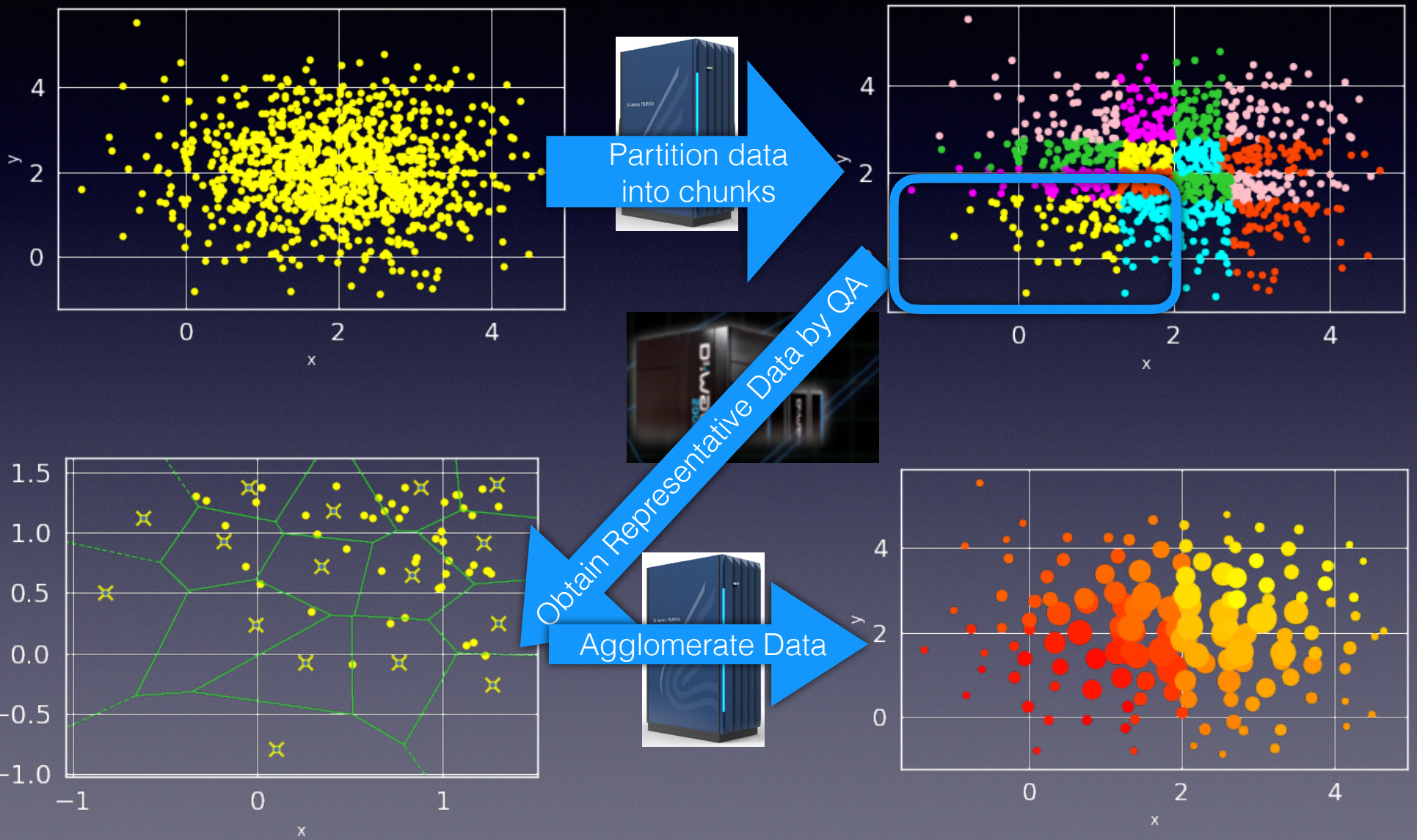
Quantum Annealing-assisted ML frameworks



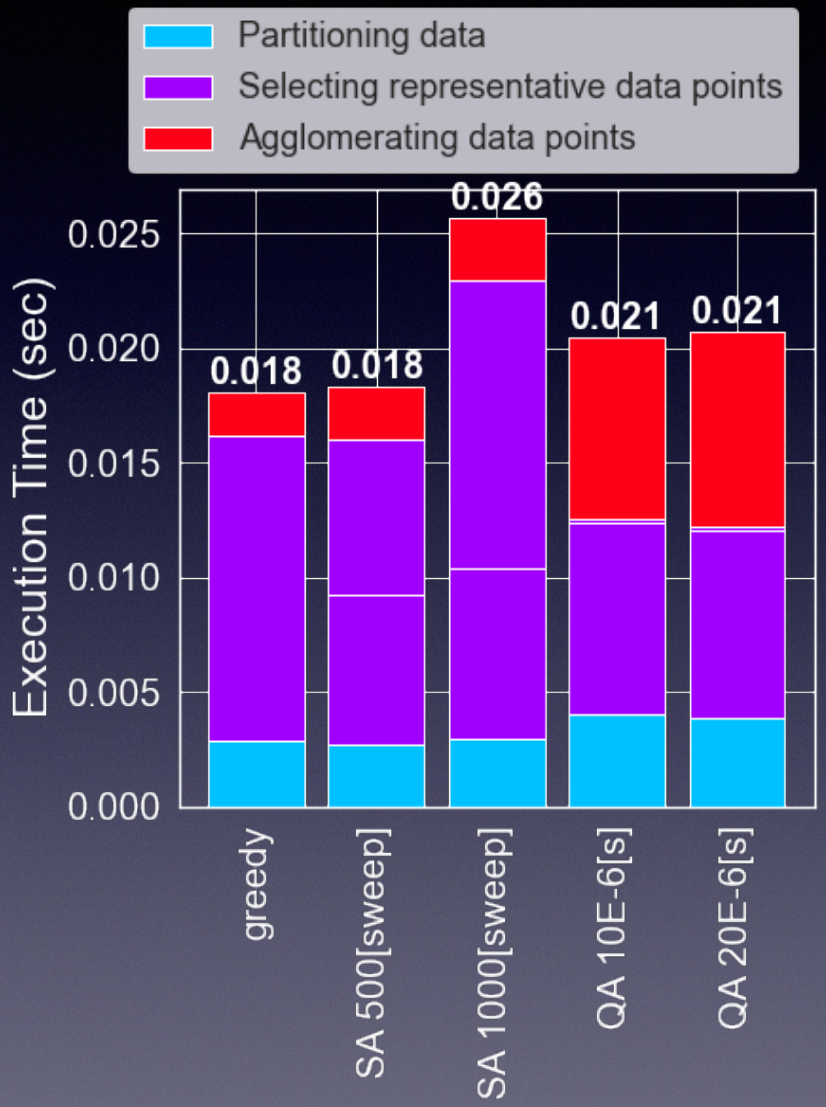
- ✓Hierarchical screening involving clustering approach
- ✓Highly accurate machine learning model based on polymer physics
- ✓Inverse problem-based optimum design for screening of polymeric materials

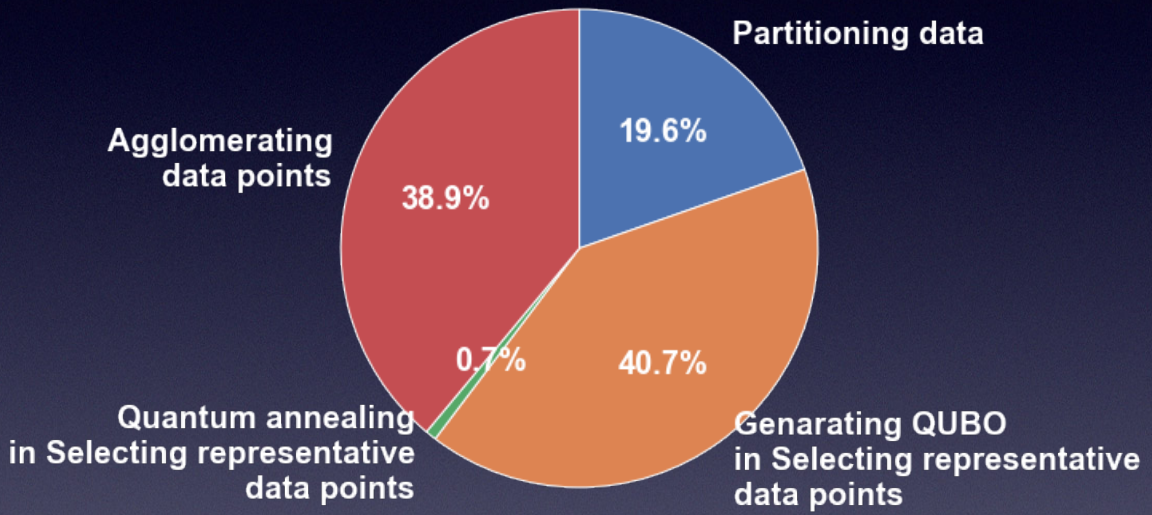


Hierarchical Agglomerative Clustering Using Quantum Annealing



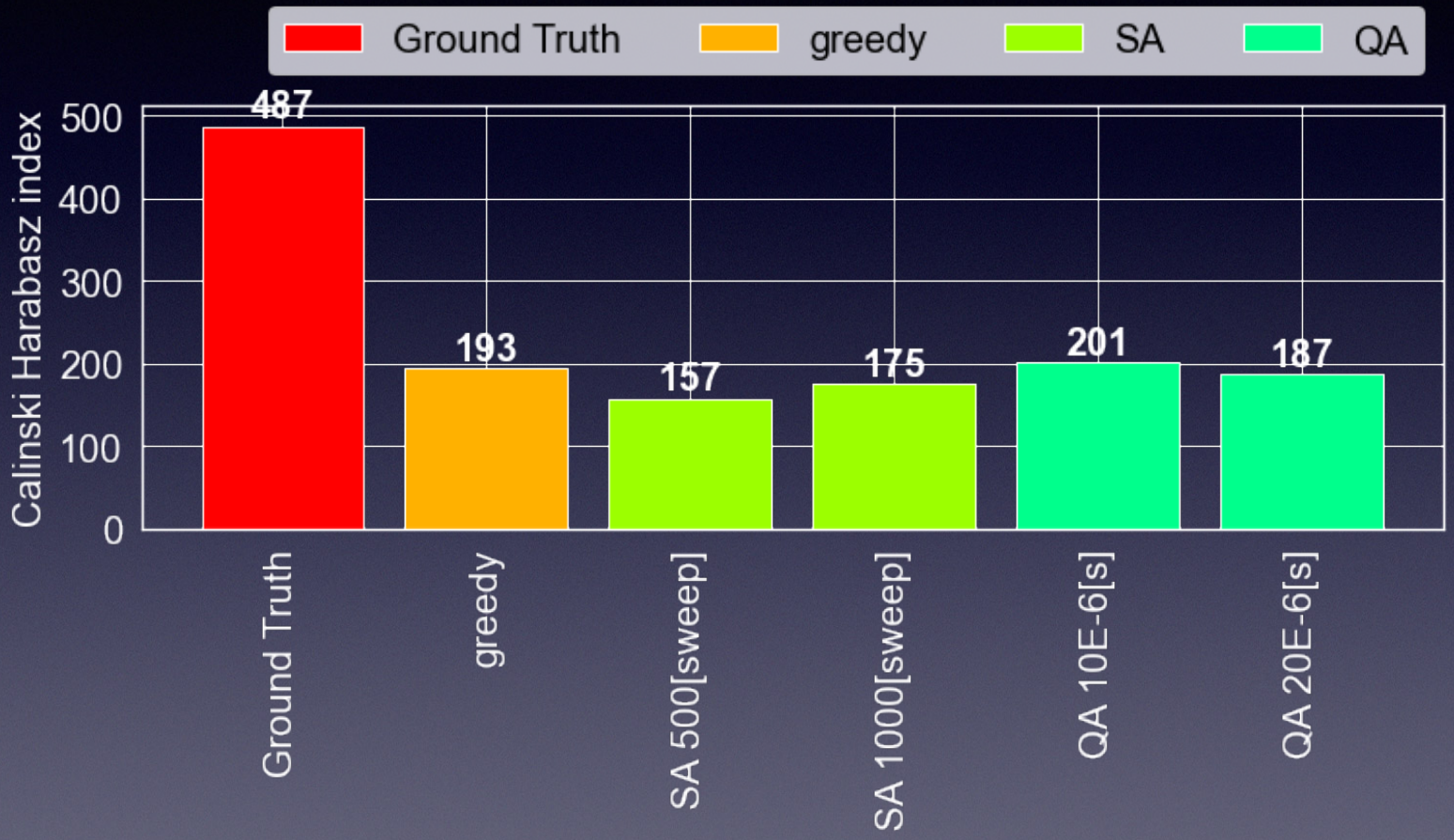
Design and Evaluation of QA-Assisted Clustering





The Iris data set of 150 data, four features each, whose clusters are three, is used.
QA is performed by D-Wave 2000Q and the others are by Xeon Gold of SX-Aurora TSUBASA.

Design and Evaluation of QA-Assisted Clustering



The Iris data set of 150 data, four features each, whose clusters are three, is used.
QA is performed by D-Wave 2000Q and the others are by Xeon Gold of SX-Aurora TSUBASA.

Summary

- Emerging applications that integrates high-performance simulation and big-data analysis for the realization of Society 5.0
  - ✓ Society 5.0 is a human-centered society that balances economic advancement with the resolution of social problems by a system that highly integrates cyberspace and physical space.
  - ✓ Simulation approach and data science approach work in a complementary style to realize Society 5.0.
- Realization of general-purpose computing by ensemble of domain specific architectures as the next generation computing infrastructure toward post Moore's era
  - ✓ Maximize computing performance per cost and/or power best suited for a specific domain
  - ✓ Best mix of domain specific architectures that satisfies the demands of a wide variety of applications
- ★ R&D of a next generation HPC infrastructure: Fusion of Quantum-Annealing and classical HPC in a unified way
  - ✓ SX-Aurora TSUBASA, combination of vector engine and X86 engine, has a great potential to achieve a high sustained performance because of its best mix of vector architecture for memory-intensive apps. and x86 architecture for complicated control-intensive apps.
  - ✓ D-wave machine, A Quantum annealing machine, is the best domain specific architecture for combinatorial problems in the post-Moore era
- ★ R&D of three innovative killer apps:
  - ✓ real-time optimal Tsunami inundation evaluation planning,
  - ✓ digital twin of a power generating Turbine for its effective operation and maintenance, and
  - ✓ material informatics for efficient carbon composite products design
- ★ Quantum annealing has a potential as a game changer toward the post-Moore era, but still is in its infancy
  - ✓ We are seeing The Dawn of Quantum Computing!?
  - ✓ Yes it needs more efforts and breakthrough to make it happen!

Acknowledgments

Members of Association for Real-time Tsunami Science (ARTS)

★ Tohoku University
- IRIDES Shun-ichi Koshimura, Takashi Abe
- Graduate School of Science: Ryota Hino, Yusaku Ota
- Cyberscience Center: Kenji Oizumi

★ KOKUSAI KOGYO Co, LTD.
- Yoichi Murashima (Visiting Prof. of Tohoku Univ.)
- Muneyuki Suzuki
- Takuya Inoue

★ NEC
- Akihiro Musa (Visiting Prof. of Tohoku Univ.)
- Osamu Watanabe (Visiting Researcher of Tohoku Univ.)

★ NEC Solution Innovator LTD.
- Yoshihiko Sato

★ Osaka University
- Cybermedia Center
  - Shinji Simojyo
  - Susumu Date

★ A2 Corp.
- Masaaki Kachi

Members of Research Division of High-Performance Computing (jointly organized with NEC), Tohoku University
- Hiroyuki Takizawa (Cyberscience Center)
- Akihiro Musa (Visiting Prof., NEC)
- Mitsuo Yokokawa (Visiting Prof., Kobe Univ.)
- Ryusuke Egawa (Visiting Prof, Tokyo Denki Univ)
- Shintaro Momose (Visiting Assoc. Prof., NEC)
- Kazuhiko Komatsu (Cyberscience Center)
- Masayuki Sato (GSIS)
- Technical Staff members (all from Cyberscience Center)
  - Kenji Oizumi
  - Satoshi Ono
  - Tsuyoshi Yamashita
  - Atsuko Saito
  - Tomoaki Moriya
  - Daisuke Sasaki
- Visiting Researchers (all from NEC)
  - Shigeyuki Aino
  - Kazuto Nakada (Research Prof, Tohoku Univ)
  - Noritaka Hoshi
  - Takashi Hagiwara
  - Osamu Watanabe
  - Yoko Isobe
  - Yasuhisa Masaoka
  - Takashi Soga
  - Yoichi Shimomura
  - Soya Fujimoto