System and Interconnect of the Supercomputer Fugaku

Yuichiro Ajima
Global Fujitsu Distinguished Engineer
Principal Architect
Future Society & Technology Unit
Fujitsu Limited

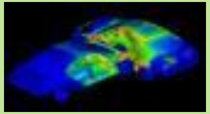
5/12/2022, ICEP2022

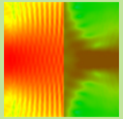
Supercomputing is the Key to the SDGs

FUJITSU

Evolution of supercomputing technology

Simulations
- Fluid Dynamics
- Collision Analysis
- Material Science
- etc.





Data Analytics
- Artificial Intelligence
- Big Data
- etc.

Advanced supercomputing is expected to bring innovation in many areas

Areas	Life Science	Energy	Manufacturing	Disaster Mgmt.

SDGs

3 GOOD HEALTH AND WELL-BEING

7 AFFORDABLE AND CLEAN ENERGY

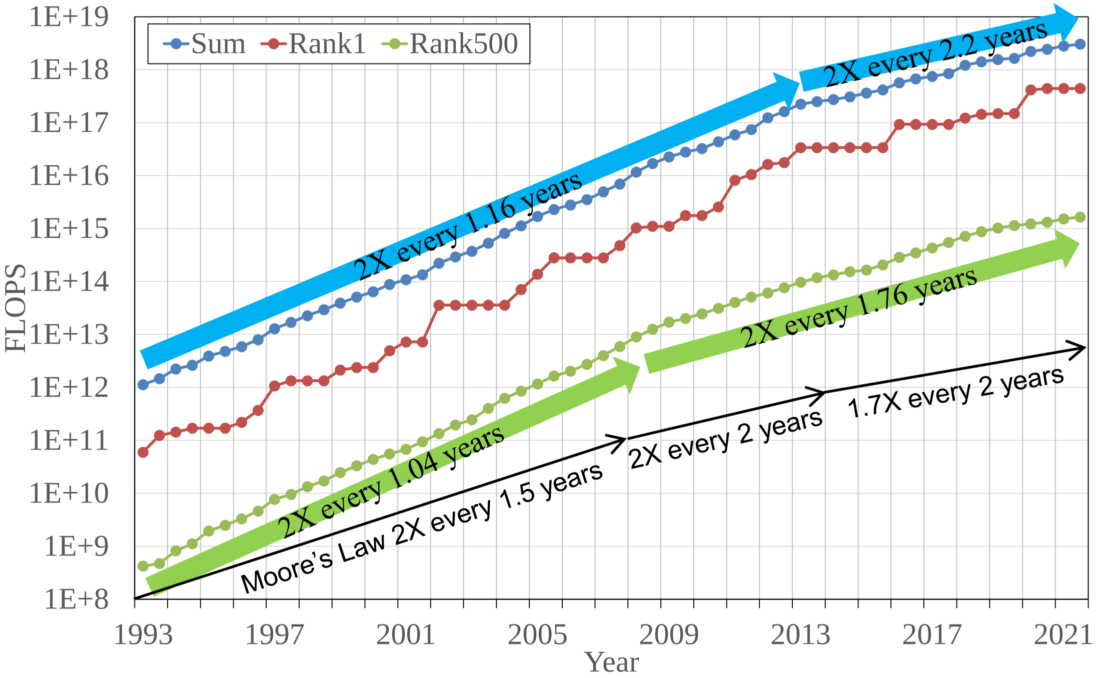
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

11 SUSTAINABLE CITIES AND COMMUNITIES

13 CLIMATE ACTION

Supercomputers are the key infrastructure for the SDGs

Supercomputer Performance Development



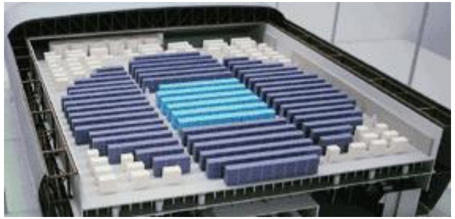
- Performance continues to improve at a rate that exceeds Moore's Law, due to increases in system scale and density

Increased Scale and Density

- Modern supercomputers are distributed memory parallel computers
  - Parallel: Interconnect connects many nodes
  - Distributed memory: OS runs per node (1-8 processor sockets)

(C) JAXA

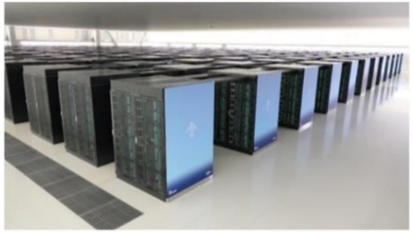
Numerical Wind Tunnel (1993)
124GFlops
140 CPU



(C) JAMSTEC

Earth Simulator (2002)
35.9TFlops
5,120 CPU

K computer (2011)
10.5PFlops
88,128 CPU

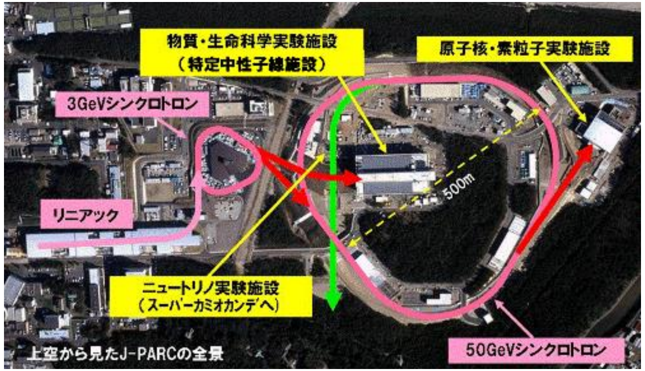


(C) RIKEN

Fugaku (2020)
442PFlops
158,976 CPU

National Large-Scale Research Facilities

- Fugaku is one of the government-designated shared research facilities, as was its predecessor, the K computer
  - Basically, free of charge for domestic researchers
  - Examples of other facilities
    - Radiation Facility (SPring-8), X-ray Free Electron Laser Facility (SACLA), Proton Accelerator Research Complex (J-PARC)



http://www.mext.go.jp/a_menu/kagaku/shisetsu/index.htm

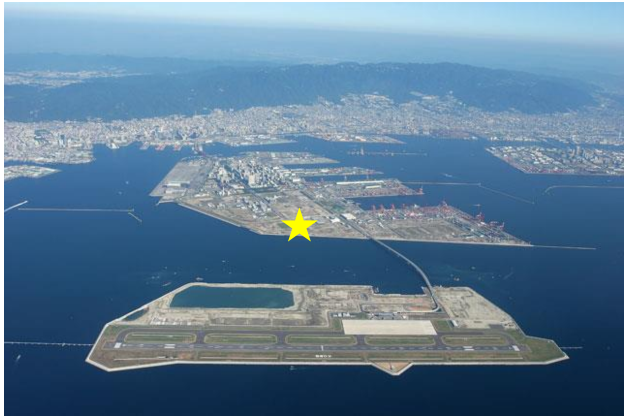
Organization and Location of Fugaku

FUJITSU

- Organization: RIKEN R-CCS
  - Responsible for operation
  - Also, the main developer

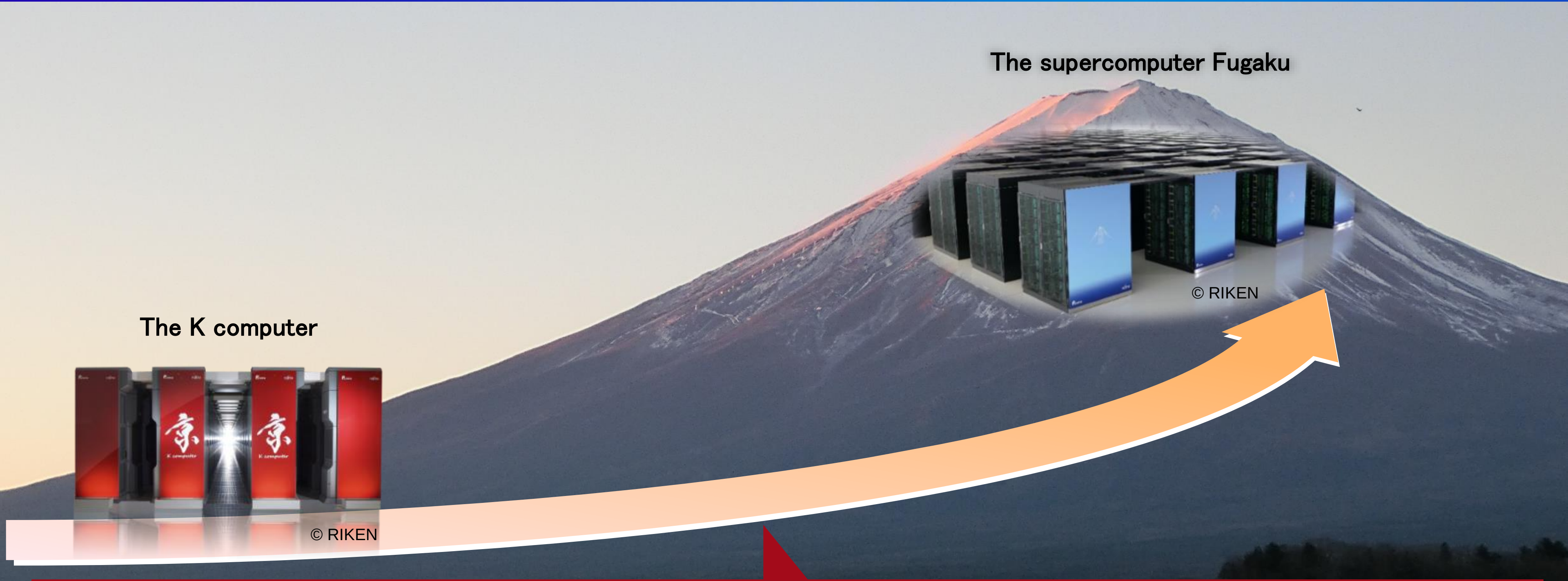
https://www.r-ccs.riken.jp/jp/overview/

- Location: Kobe City
  - Port Island
  - Opposite shore from Kobe Airport



https://www.mlit.go.jp/koku/koku_fr14_000081.html

Migration from the K Computer to Fugaku



- The K computer shut down in August 2019 after seven years of full operation
- Five months later, the supercomputer Fugaku started partial shared use in March 2020

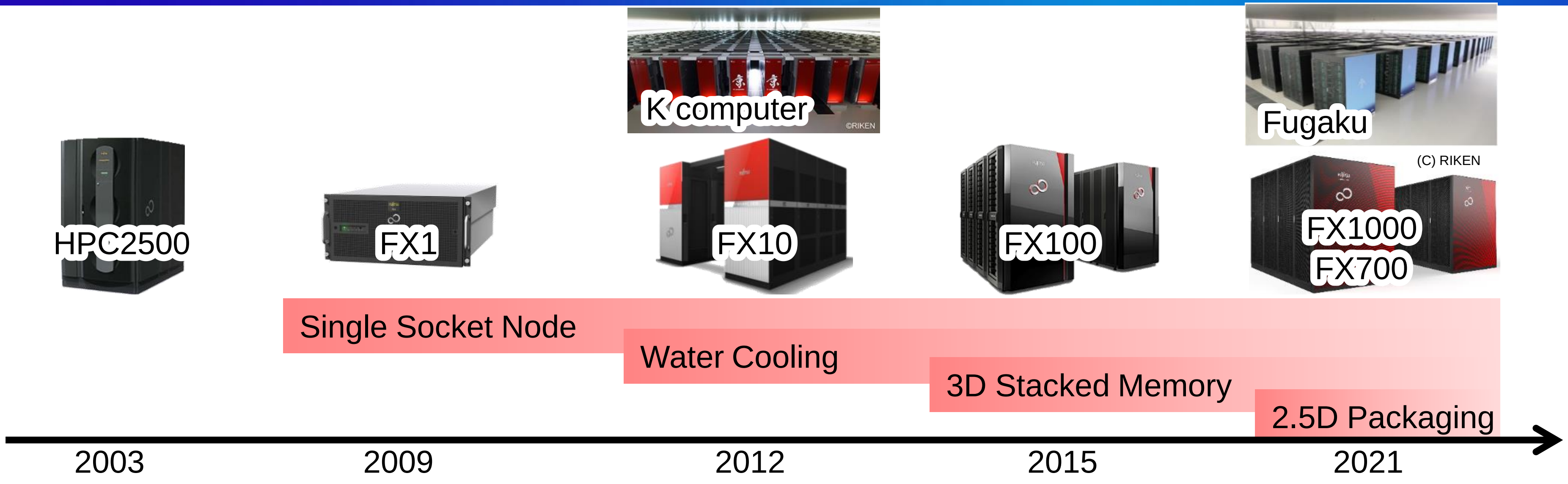
Benchmark Results

FUJITSU

	K computer	Fugaku
TOP500 (Petaflops)	10.51 (#1 Jun/Nov 2011)	442.01 (#1 Jun 2020~)
HPCG (Petaflops)	0.60274 (#1 Nov 2011~Nov 2017)	1.60045 (#1 Jun 2020~)
HPL-AI (Petaflops)	–	2,000 (#1 Jun 2020~)
Graph500 (GTEPS)	31,302.4 (#1 Jun 2014, Jul 2015~Jun 2019)	102,956 (#1 Jun 2020~)

- TOP500: Performance of solving the linear equation Ax = b with dense coefficient matrix. This benchmark has been measured for 30 years and is the most popular performance indicator
- HPCG: Performance of solving the linear equation with sparse coefficient matrix using Conjugate Gradient method. This performance will be dominated by memory system performance
- HPL-AI: Performance of solving the linear equation utilizing lower precision floating point calculation, such as fp16, which AI applications often use
- Graph500: Performance of big data processing. This performance will be dominated by interconnect performance

Packaging of Fujitsu Supercomputers



K computer

©RIKEN

Fugaku

(C) RIKEN

HPC2500

FX1

FX10

FX100

FX1000
FX700

Single Socket Node

Water Cooling

3D Stacked Memory

2.5D Packaging

2003

2009

2012

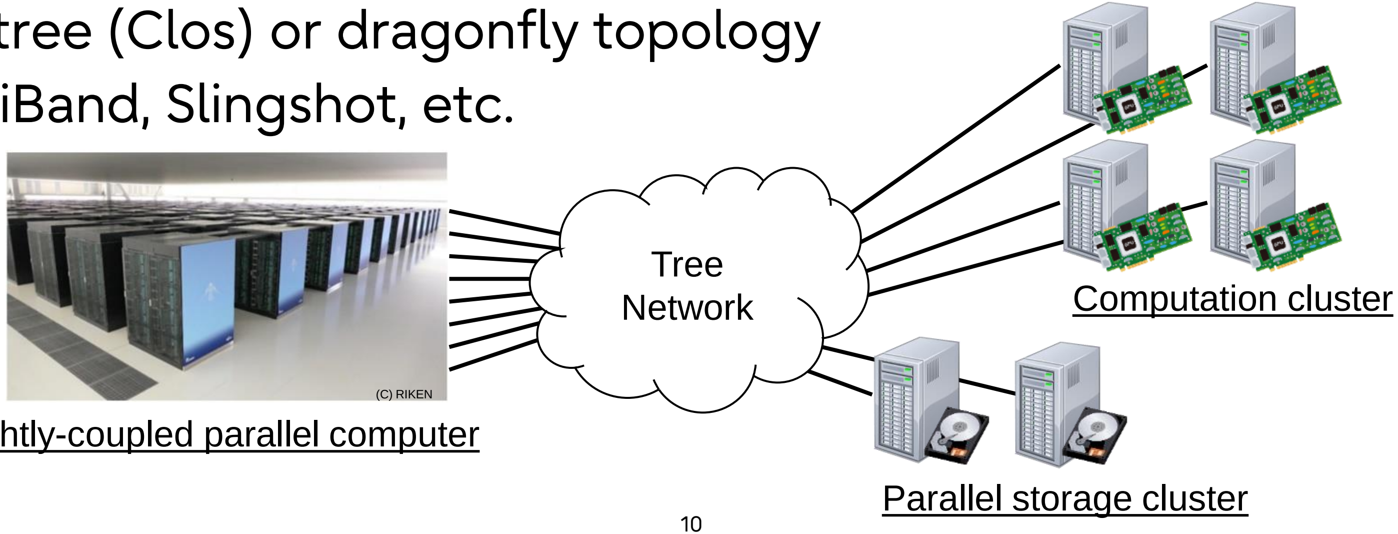
2015

2021

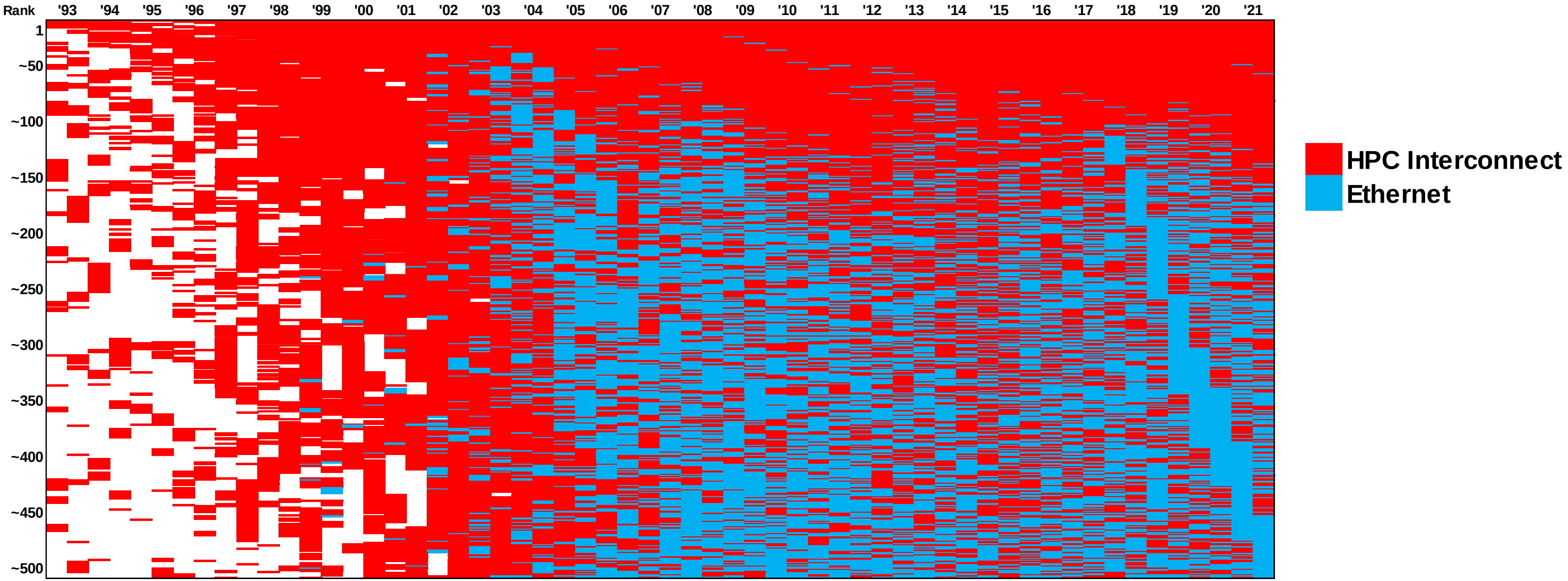
- Fujitsu has developed single-socket node, water-cooled supercomputers using 3D stacked memory
- Fugaku integrates memory into the CPU package

Two Types of HPC System and Interconnect

- Tightly-coupled parallel computer
  - Torus or multi-plane topology
  - K computer, Fugaku, Google TPU, NVIDIA DGX, etc.
- Cluster and parallel storage
  - Fat-tree (Clos) or dragonfly topology
  - InfiniBand, Slingshot, etc.



Interconnect Types of Supercomputers



- Top-level supercomputers use HPC interconnects such as InfiniBand and proprietary interconnects
- Ethernet-only systems have increased since 2000s

System Architecture and Structure of Fugaku

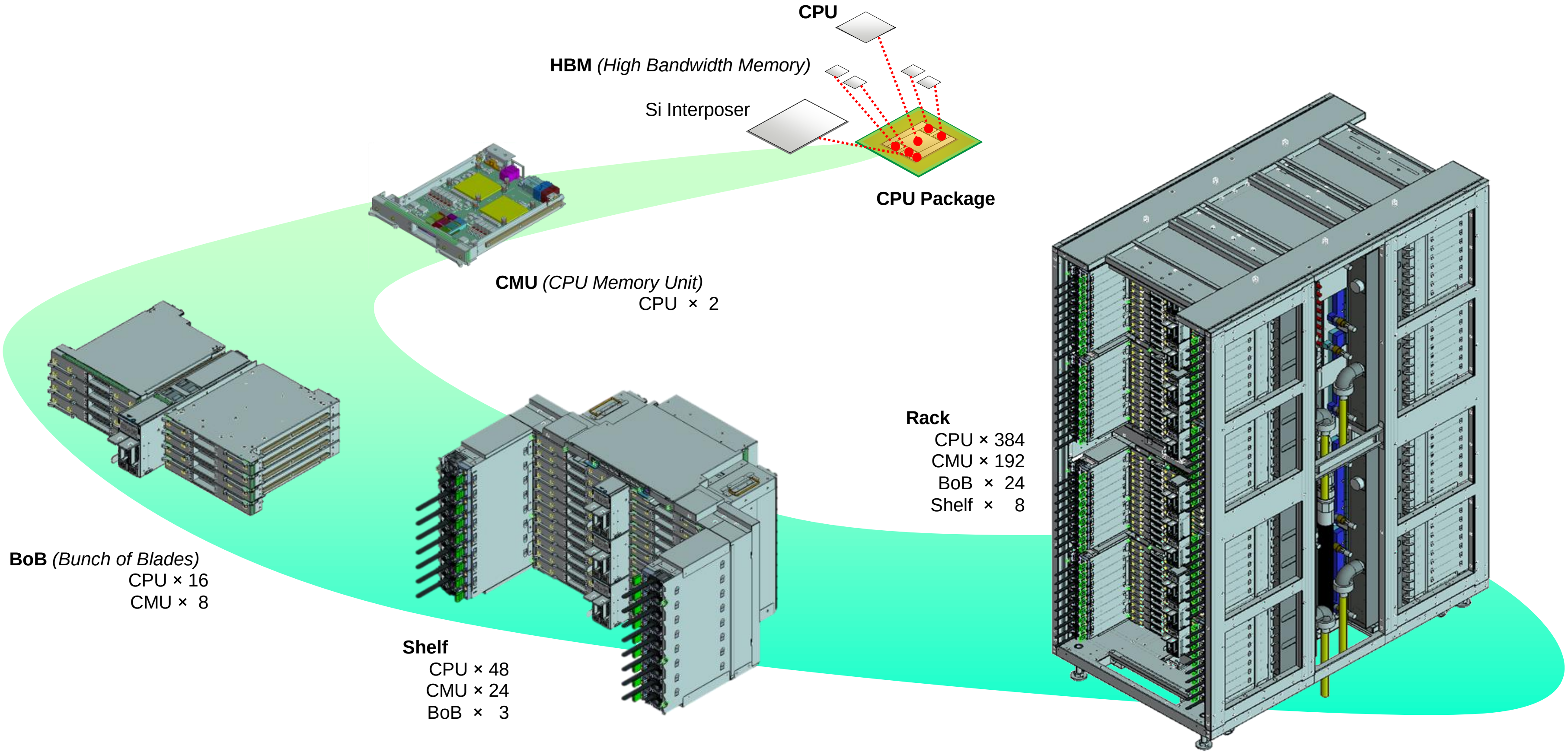
Configuration and Topology of Fugaku

- Configuration: 158,976 nodes x 1 CPU
- Topology: 6D-mesh/torus (24x23x24x2x3x2)
- Link: 97,632 AOCs, 4X 28 Gbps

"Report on the Fujitsu Fugaku System," Jack Dongarra (2020)

https://www.r-ccs.riken.jp/intro-hpc/hellosc-fugaku/04.html

Packaging Structure of Fugaku Rack



CPU

HBM *(High Bandwidth Memory)*

Si Interposer

CPU Package

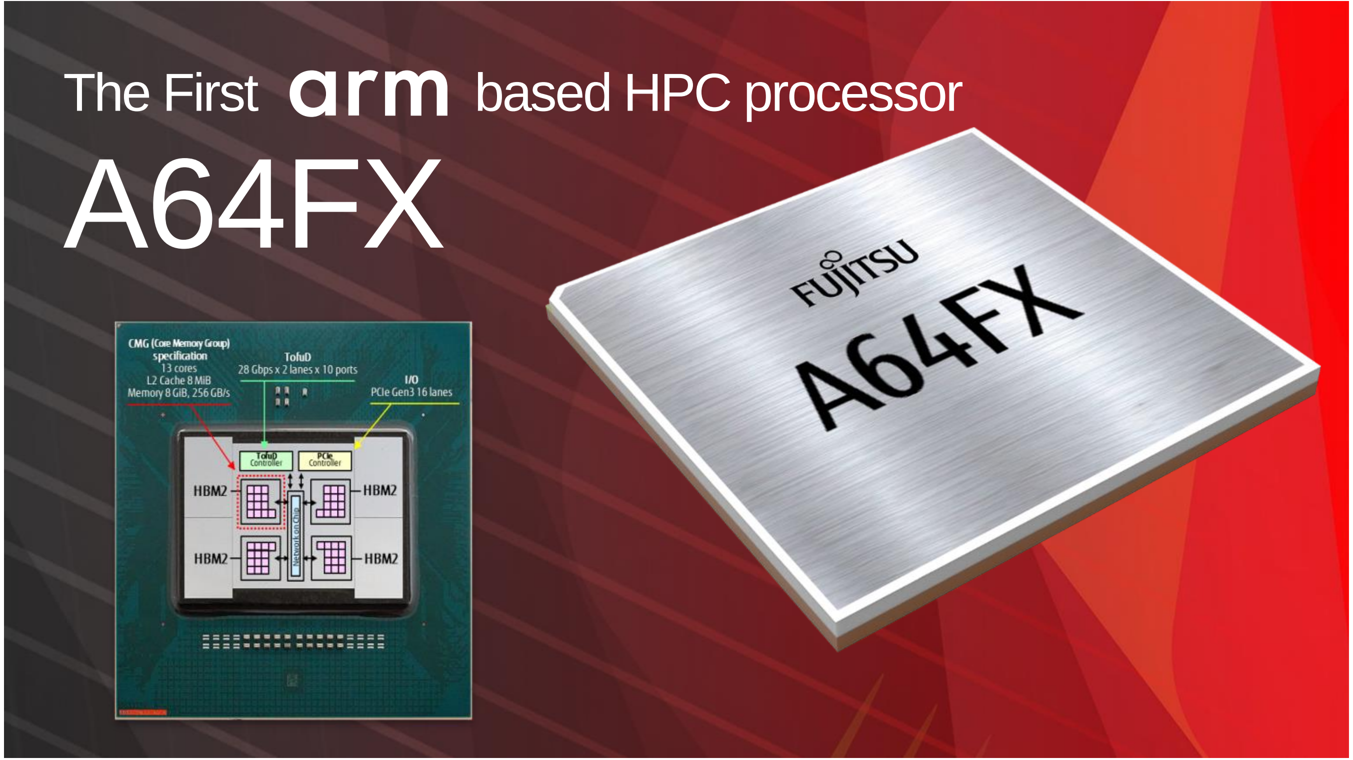
CMU *(CPU Memory Unit)*
CPU × 2

BoB *(Bunch of Blades)*
CPU × 16
CMU × 8

Shelf
CPU × 48
CMU × 24
BoB × 3

Rack
CPU × 384
CMU × 192
BoB × 24
Shelf × 8

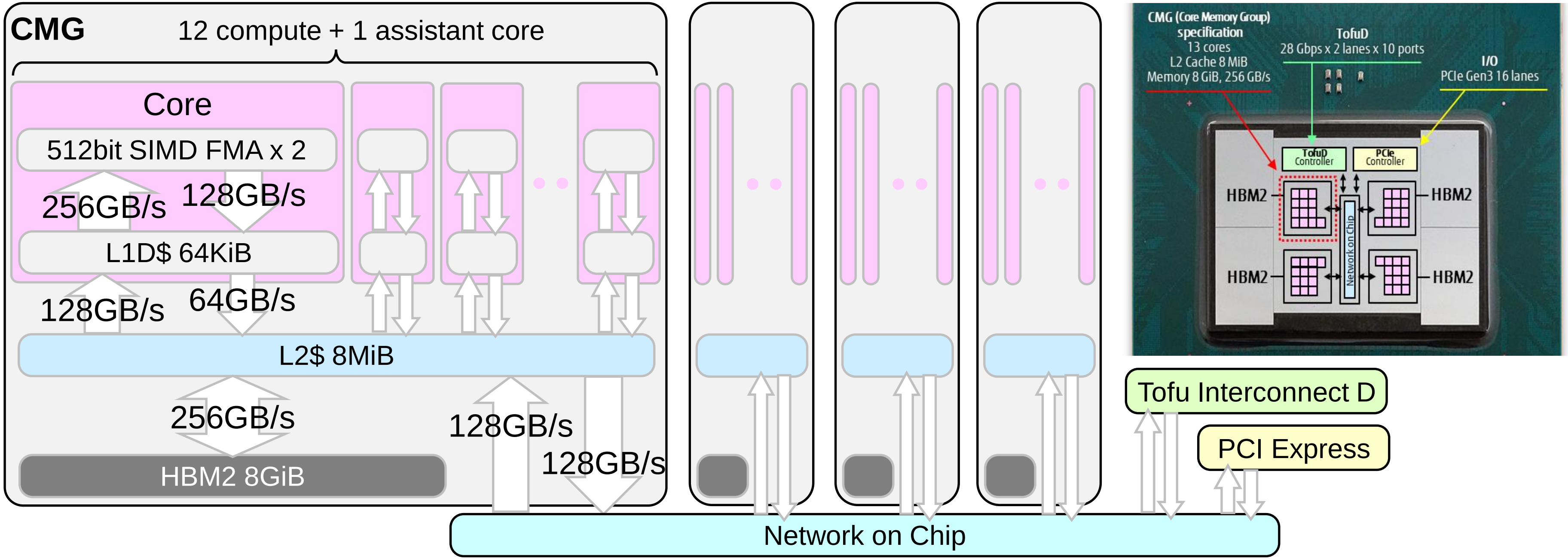
A64FX Processor Developed for Fugaku



- A64FX integrated memory into the package

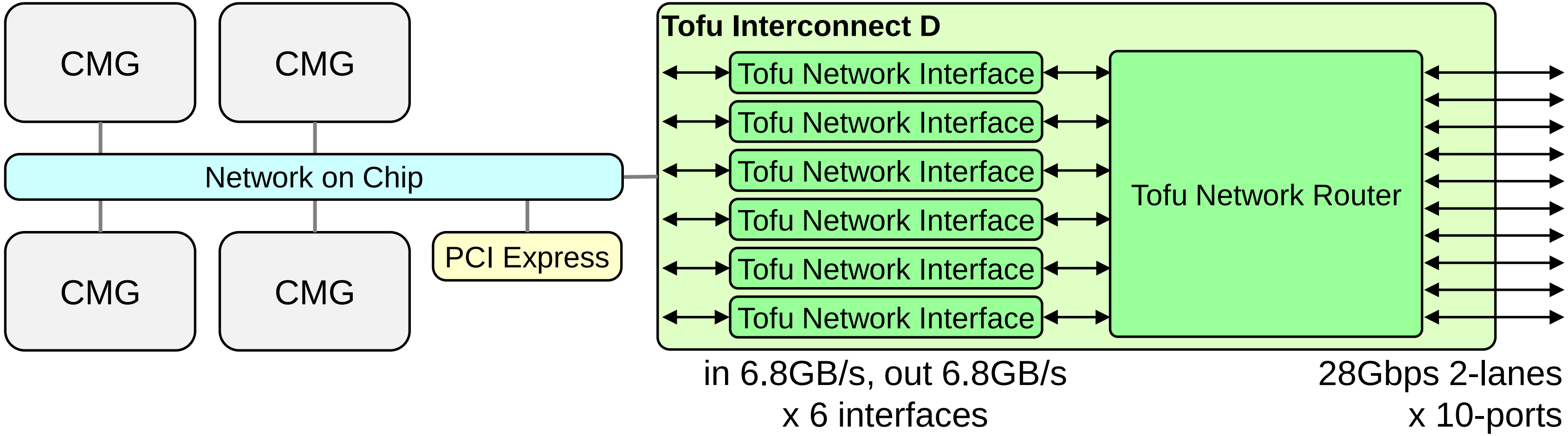
Configuration and Bandwidth of A64FX

- A64FX consists of 4 Core Memory Groups (CMGs)
  - Each CMG consists of 13 cores, shared L2 cache, and one HBM



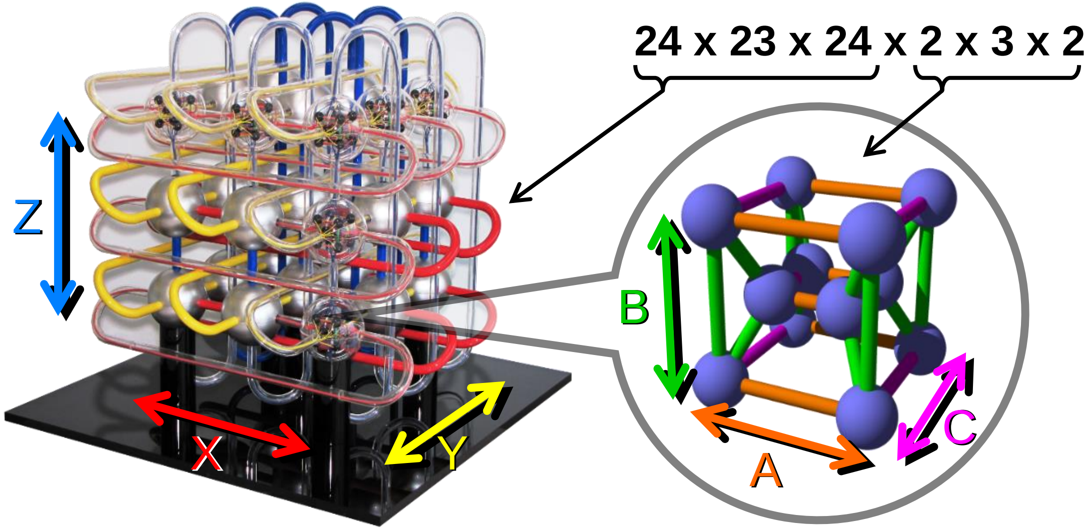
Tofu Interconnect D

- Six network interfaces per node
- The network router has 10 ports with 56 Gbps links
    - Each port is directly interconnected with other CPUs
    - Forming a 6-dimensional mesh/torus network



6D Mesh/Torus Network

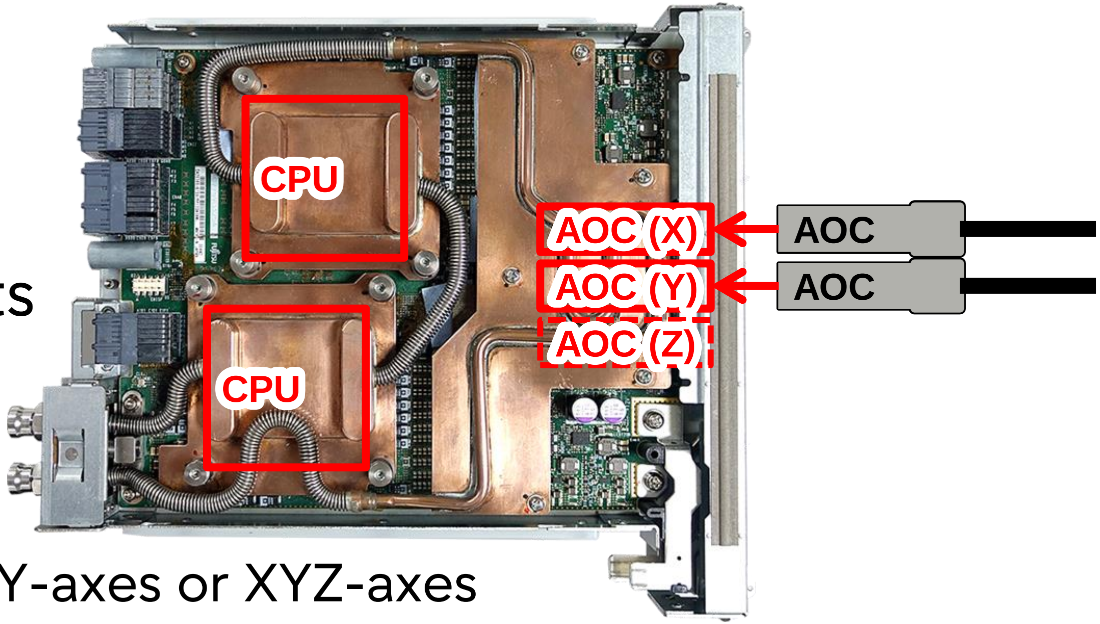
- XYZ dimensions are scalable, and ABC have fixed size



- The B-axis with length 3 improves the system availability
  - Virtual torus takes advantage of its redundancy

Structure and Topology of CMU

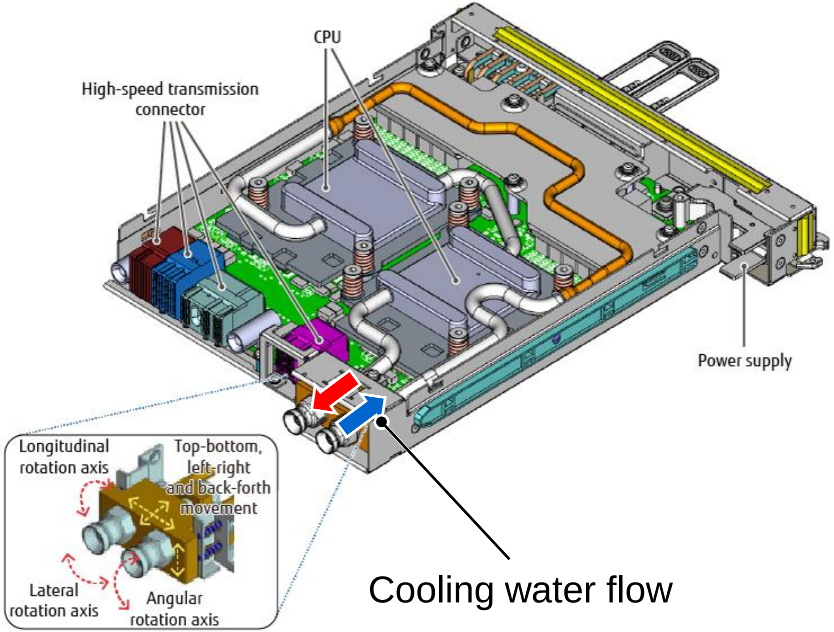
- Two CPUs
  - Connected by the C axis



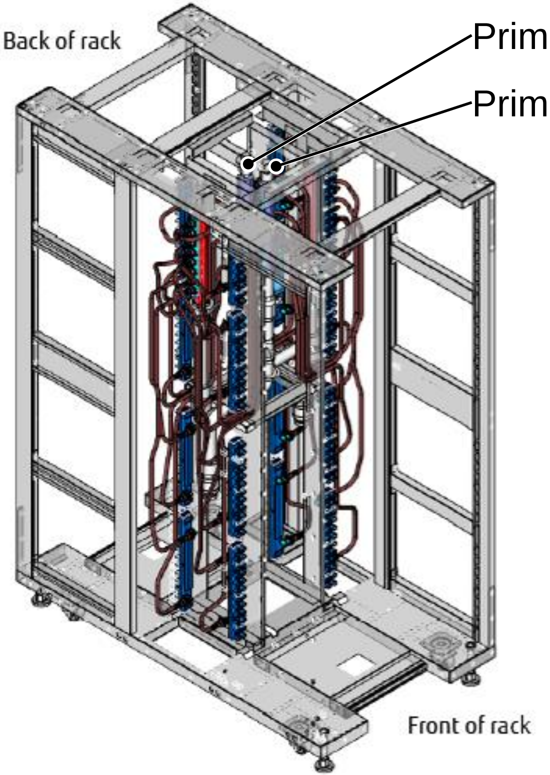
- Two or Three AOC ports
  - Each AOC, 4X 28 Gbps
    - Shared by two CPUs
    - Each CPU uses 2 lanes
  - AOCs are connected in XY-axes or XYZ-axes
  - The average number of AOC ports per CMU is 2.5
    - The average number of AOC per CMU is 1.25, and per CPU is 0.625

Water-Cooling of CMU

- Cooling water is supplied and discharged from the central block of the rack


CPU
High-speed transmission connector
Power supply
Longitudinal rotation axis
Top-bottom, left-right and back-forth movement
Lateral rotation axis
Angular rotation axis

Cooling water flow



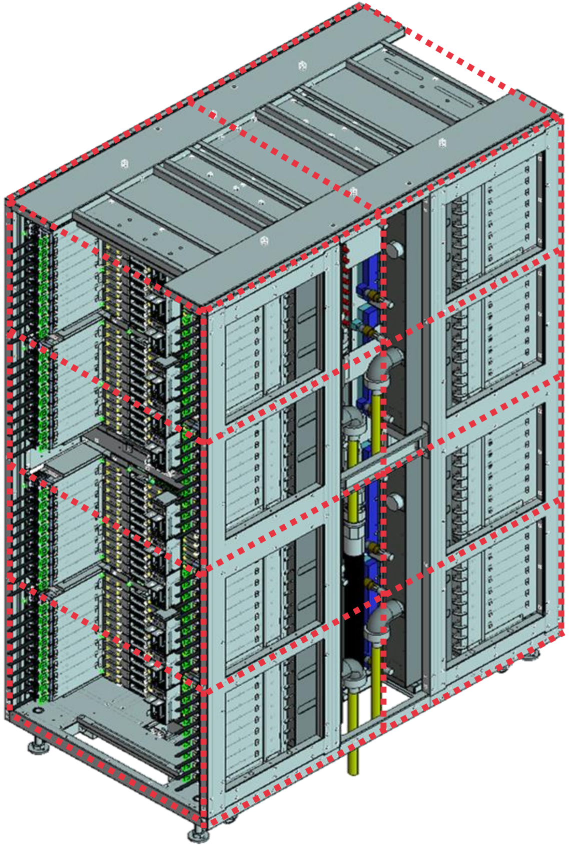
Primary supply pipe

Primary discharge pipe

Structure and Dimension of Shelf and Rack

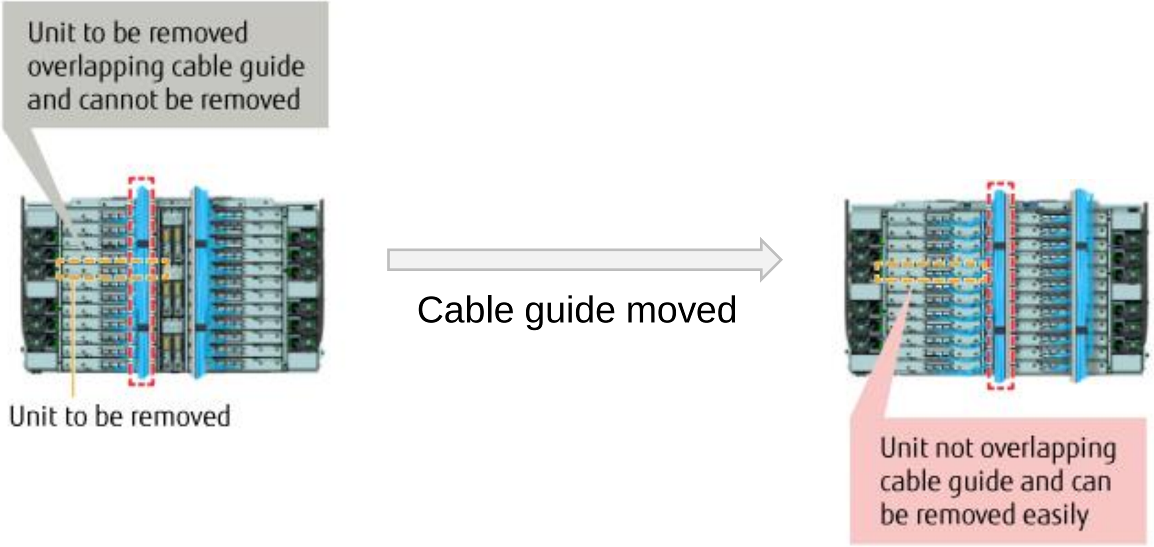
FUJITSU

- Shelf
  - 24 CMUs connected in ZAB-axes (4x2x3)
  - with backplane and electrical cables
- Half-rack (top or bottom)
  - Four shelves connected in XY-axes (2x2)
  - with electrical cables
- Rack
  - In most cases, two half-racks are connected
  - 96 out of 480 AOC ports in a rack are used



Movable Cable Guides

- We developed a movable cable guide to achieve high packaging density while avoiding interference with the replacement work

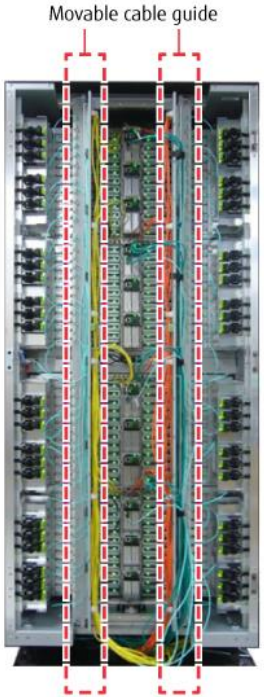


Unit to be removed overlapping cable guide and cannot be removed

Unit to be removed

Cable guide moved

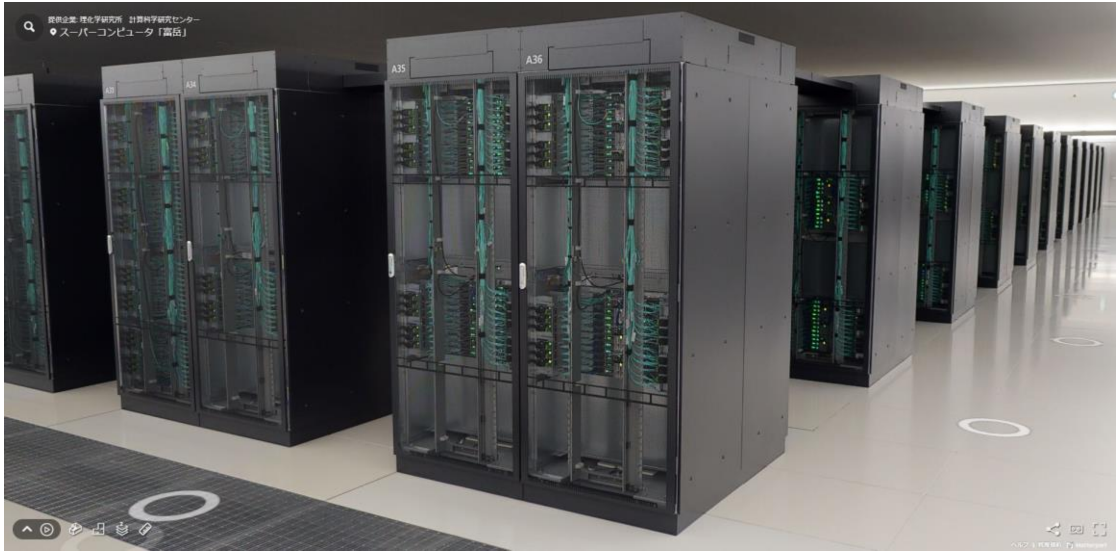
Unit not overlapping cable guide and can be removed easily



Movable cable guide

Half-Mount Racks of Fugaku

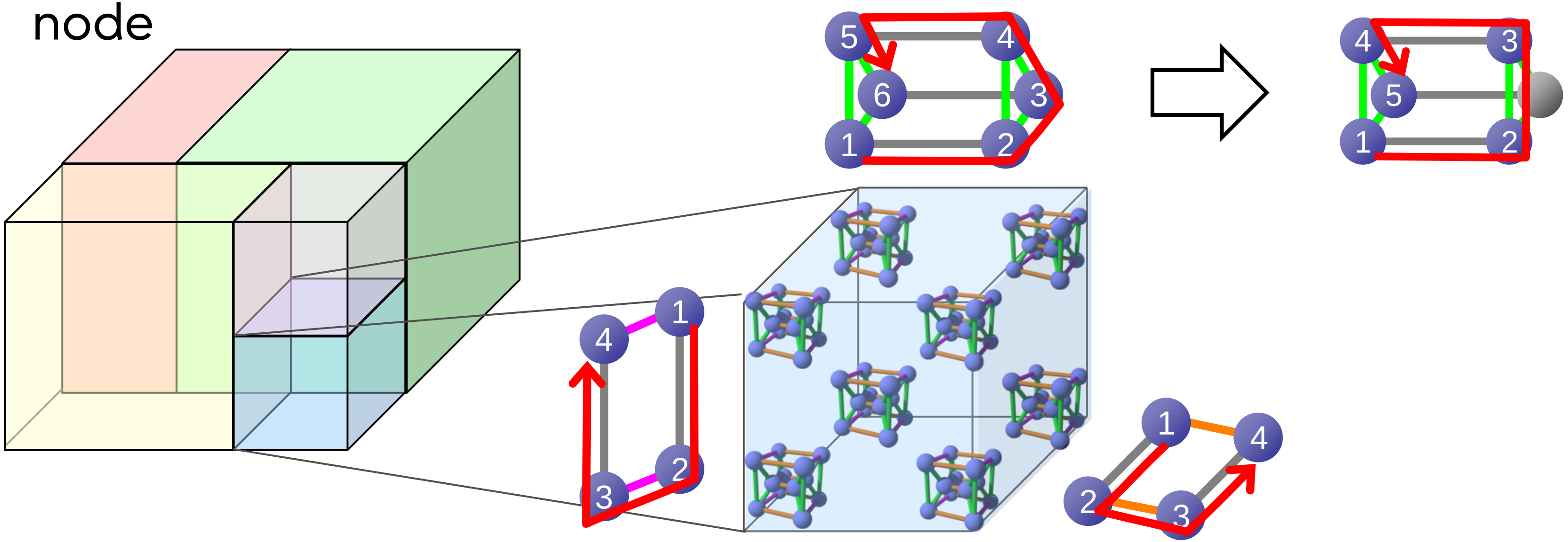
- The network size of Fugaku is 24 x 23 x 24 x 2 x 3 x 2
  - Some half-racks contain only half of the nodes
  - The network size of a half-mounted half-rack is 2 x 1 x 4 x 2 x 3 x 2


提供企業: 理化学研究所　計算科学研究センター
スーパーコンピュータ「富岳」
A33
A34
A35
A36

https://www.r-ccs.riken.jp/en/fugaku/3d-models/, https://my.matterport.com/show/?m=mnpGYx1pQtx&sr=-.23,-.99&ss=176

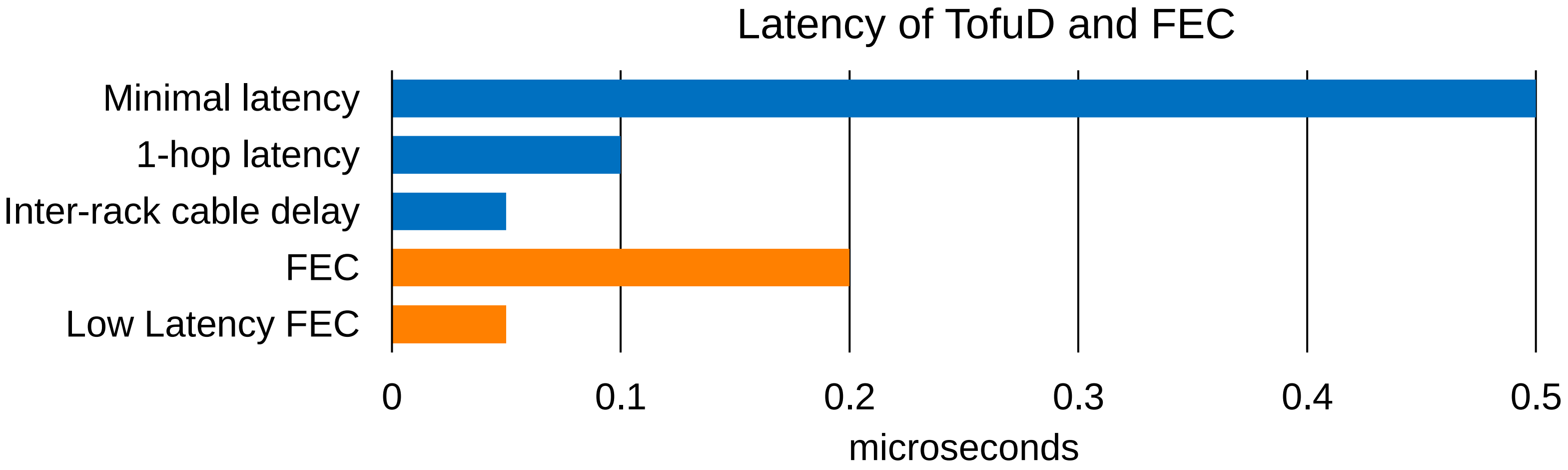
Fault Isolation and Virtual Torus

- A failed node is isolated in a rectangular partition
- Virtual torus can use a partition containing a failed node
  - Shortening the length of a virtual axis by one to exclude the failed node



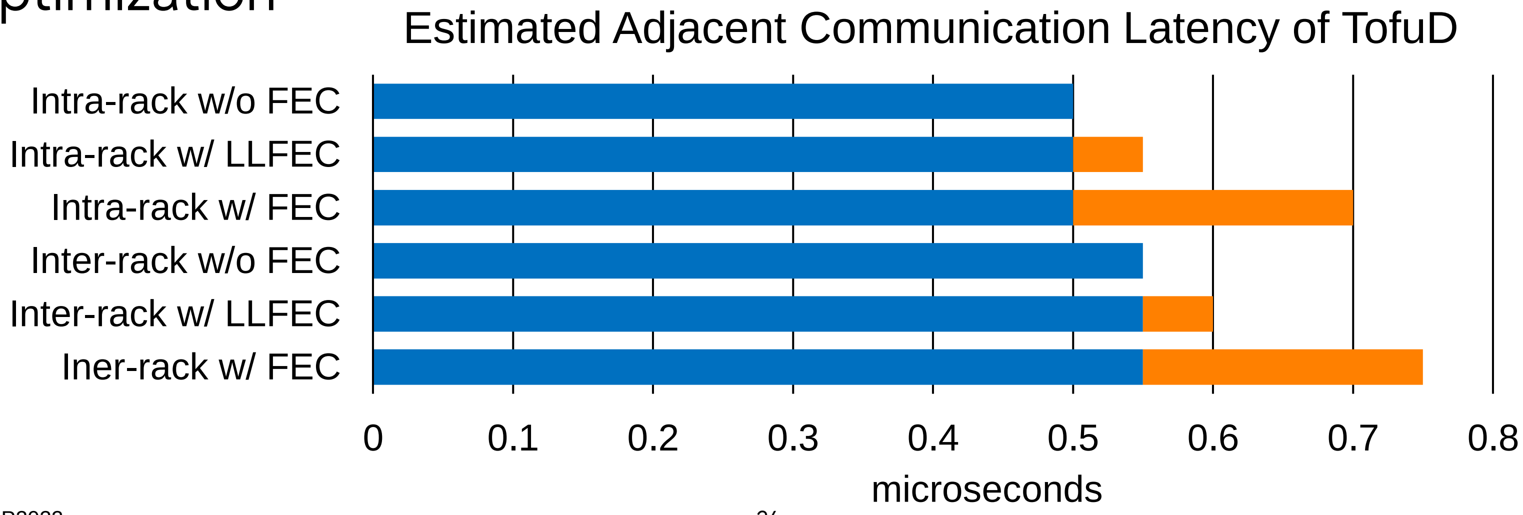
Discussion on Forward Error Correction (FEC)

- HPC Interconnects are designed for low latency
  - For example, minimal latency of TofuD is about 0.5 μsec
- FEC will add an additional delay of about 0.05 to 0.2 μsec
- TofuD was able to transmit without FEC, so it is disabled



Adjacent Communication Latency

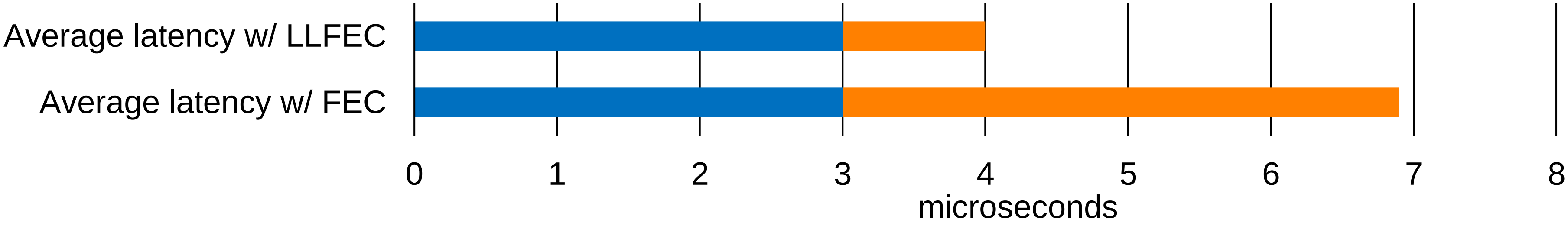
- FEC increased latency by 9-40%
  - Parallel efficiencies will be reduced in some application areas
    - Such as molecular dynamics and lattice quantum chromodynamics
  - For most apps, performance degradation can be avoided by optimization



Average Communication Latency

- FEC increased latency by 33% to 130%
  - Collective communication with short message sizes will be affected
    - Frequently used in insufficiently parallelized programs
      - This can be a limiting factor in expanding the range of parallel applications
    - Well-optimized applications can hide the latency by optimization
      - Such as shared parameter updates in multi-physics simulation

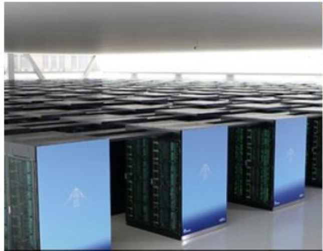
Estimated Average Latency of TofuD in 24x24x24x2x3x2 system

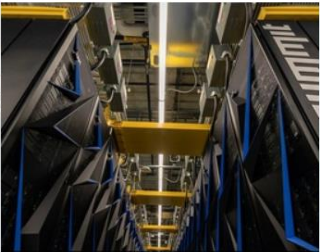


Other Top-Level Supercomputers in Recent Years

World Rank One System in Recent Years

- Number one systems located in the U.S., China, and Japan
  - Average time for a system to stay number one is about 1.7 years


Supercomputer Fugaku: RIKEN Center for Computational Science
No.1 in Jun 2020


Summit: DOE/SC/Oak Ridge National Laboratory
No.1 from Jun 2018 until Nov 2019

Sunway TaihuLight: National Supercomputing Center in Wuxi
No.1 from Jun 2016 until Nov 2017

Tianhe-2 (MilkyWay-2) : National University of Defense Technology
No.1 from Jun 2013 until Nov 2015

Titan: Oak Ridge National Laboratory
No.1 in Nov 2012

Sequoia: Lawrence Livermore National Laboratory
No.1 in Jun 2012

K Computer: RIKEN Advanced Institute for Computational Science
No.1 from Jun 2011 until Nov 2011

Tianhe-1A: National Supercomputing Center in Tianjin
No.1 in Nov 2010

Jaguar: Oak ridge National Laboratory
No.1 from Nov 2009 until Jun 2010

Roadrunner: Los Alamos National Laboratory
No.1 from Jun 2008 until Jun 2009

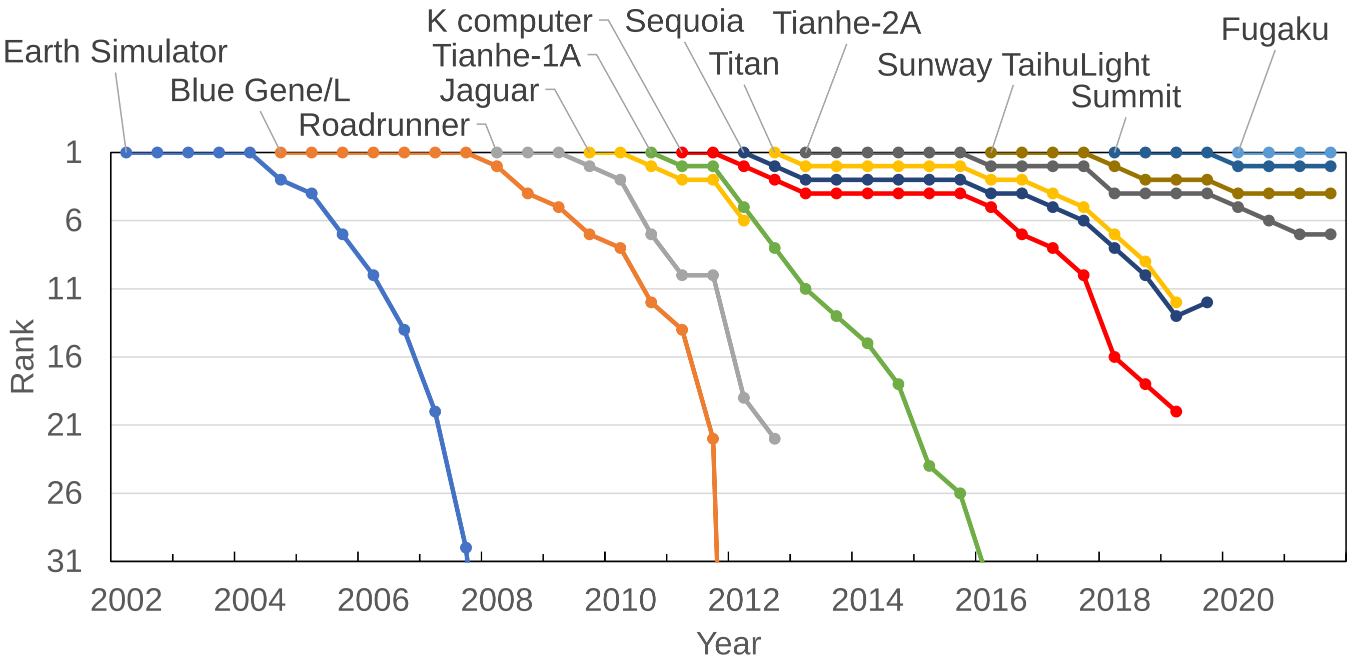
BlueGene/L: Lawrence Livermore National Laboratory
No.1 from Nov 2004 until Nov 2007

The Earth Simulator: Earth Simulator Center
No.1 from Jun 2002 until Jun 2004

https://www.top500.org/resources/top-systems/

Lifetime Ranking of the Rank 1 Systems

- The rank 1 systems after the K computer have remained in the top group for years due to their large scale



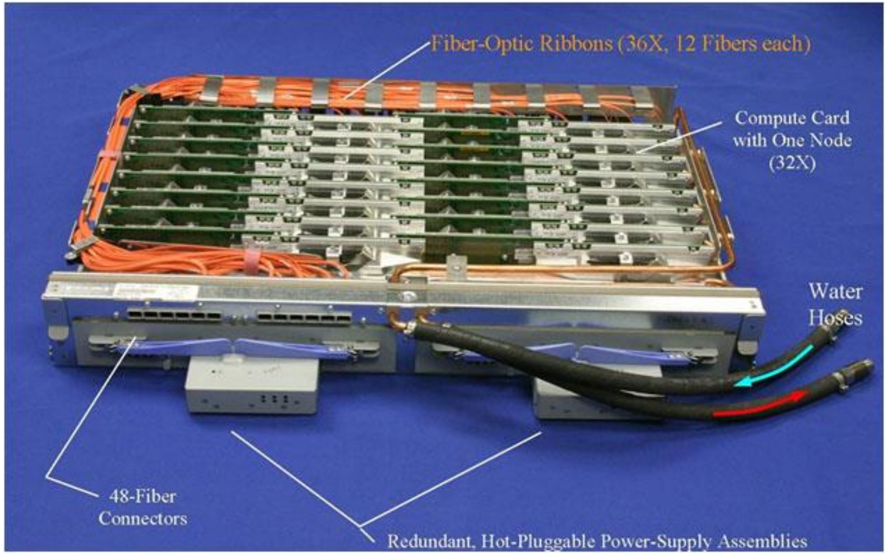
K computer (2011)

- Configuration: 82,944 nodes x 1 CPU
- Topology: 6D-mesh/torus (24x18x16x2x3x2)
- Link: about 200,000 electrical, 8X 6.25 Gbps

http://www.s.u-tokyo.ac.jp/ja/story/rigakuru/03/interview/info.html

Sequoia (2012)

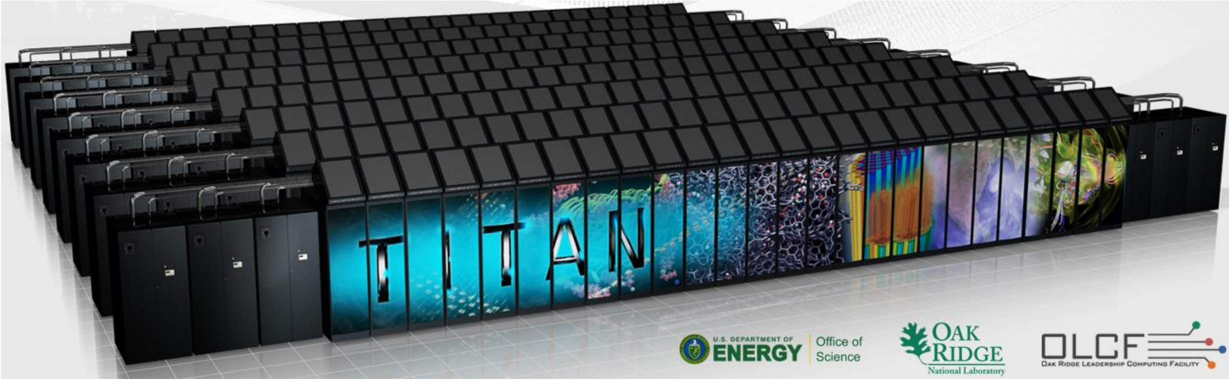
- Configuration: 98,384 nodes x 1 CPU
- Topology: 5D-torus (16x12x16x16x2)
- Link: about 12,000 optical, 24X 10 Gbps


Fiber-Optic Ribbons (36X, 12 Fibers each)
Compute Card with One Node (32X)
Water Hoses
48-Fiber Connectors
Redundant, Hot-Pluggable Power-Supply Assemblies

https://computing.llnl.gov/tutorials/bgq/, https://www.top500.org/featured/systems/sequoia-lawrence-livermore-national-laboratory/

Titan (2012)

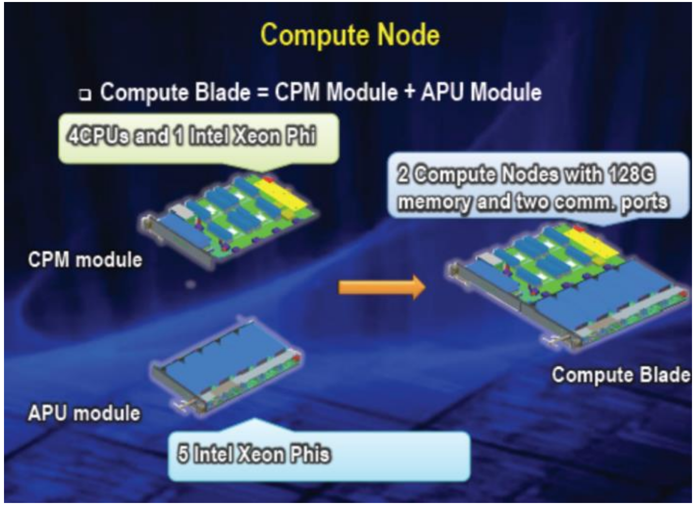
- Configuration: 18,688 nodes x (1 CPU + 1 GPU)
- Topology: 3D-torus (25x16x24)
- Link: about 20,000 electrical, 12X 3.125 Gbps


TITAN
U.S. DEPARTMENT OF ENERGY
Office of Science
OAK RIDGE National Laboratory
OLCF
Oak Ridge Leadership Computing Facility

“Hardware concepts and terminology relevant to the programmer (Magny Cours, Gemini interconnect, architecture of XE6), Launch of parallel applications/batch system, User Environment, Compilers of the XE6 (PGI, Pathscale, GNU, Cray),” https://www.nersc.gov/users/NUG/annual-meetings/NUG-2010/presentations/

Tianhe-2 (2013)

- Configuration: 16,000 nodes x (2 CPU + 3 Accelerator)
- Topology: 5-tiers tapered fat-tree
- Link: about 7,000 (estimated) optical, 8X 10 Gbps


Compute Node
□ Compute Blade = CPM Module + APU Module
4CPUs and 1 Intel Xeon Phi
2 Compute Nodes with 128G memory and two comm. ports
CPM module
Compute Blade
APU module
5 Intel Xeon Phis

Jack Dongarra, "Visit to the National University for Defense Technology Changsha, China"

Sunway TaihuLight (2016)

- Configuration: 40,960 nodes x 1 CPU
- Topology: 4-tiers tapered fat-tree (estimated)
- Link: about 6,000 (estimated) AOCs, 4X 25 Gbps

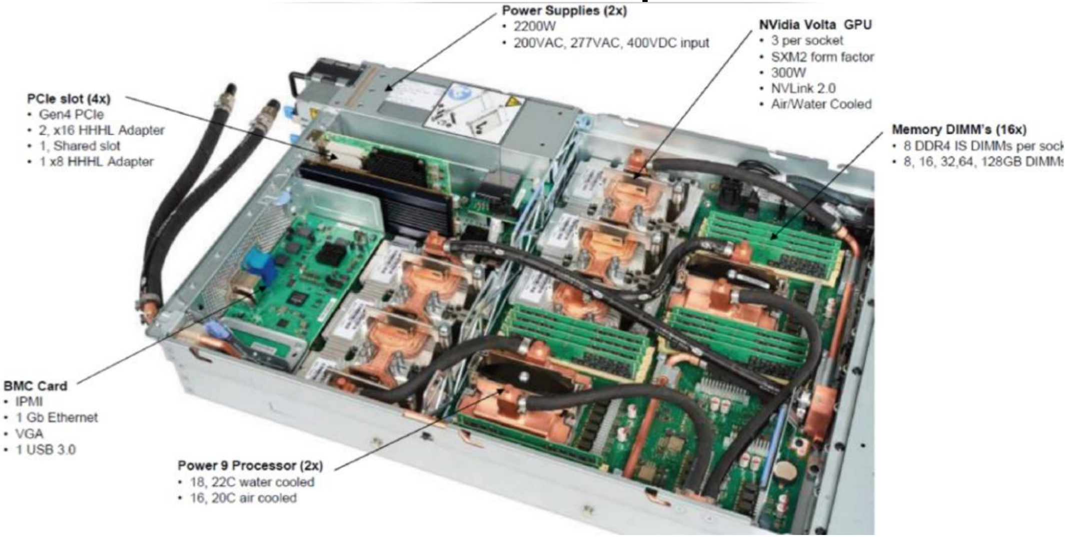
Haohuan Fu, "The Sunway TaihuLight Supercomputer System and Application"

https://www.top500.org/resources/top-systems/sunway-taihulight-national-supercomputing-center-i/

Summit (2018)

- Configuration: 4,608 nodes x (2 CPU + 6 GPU)
- Topology: 3-tiers full-bisectional fat-tree
- Link: about 9,000 (estimated) AOCs, 4X 25 Gbps

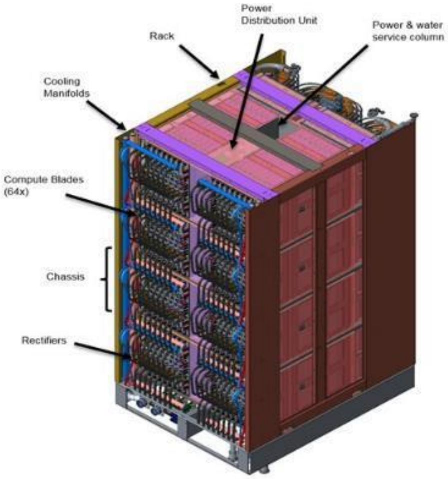
https://www.ornl.gov/news/ornl-launches-summit-supercomputer



https://www.olcf.ornl.gov/wp-content/uploads/2018/05/Intro_Summit_System_Overview.pdf

Frontier (2022 planned)

- Configuration: 9,408 nodes x (1 CPU + 4 GPU)
- Topology: dragonfly (> 73+ groups x 512 terminals)
- Link: about 10,000 (estimated) AOCs, 8X 50 Gbps

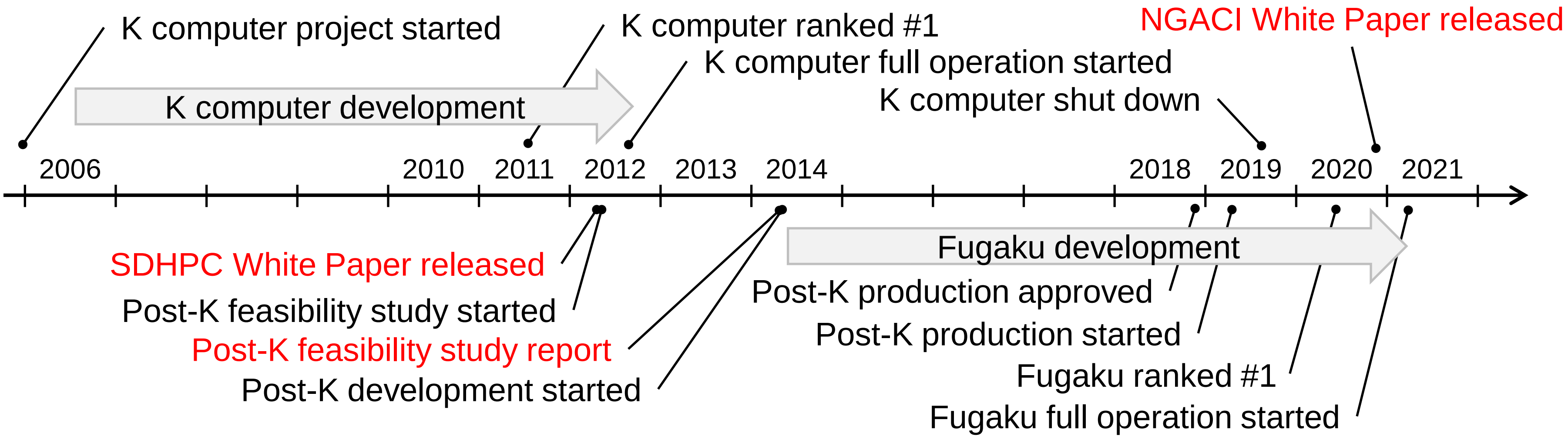


https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/202203/ASCAC_202203-Geist.pdf

Timeline of Fugaku Development and NGACI White Paper

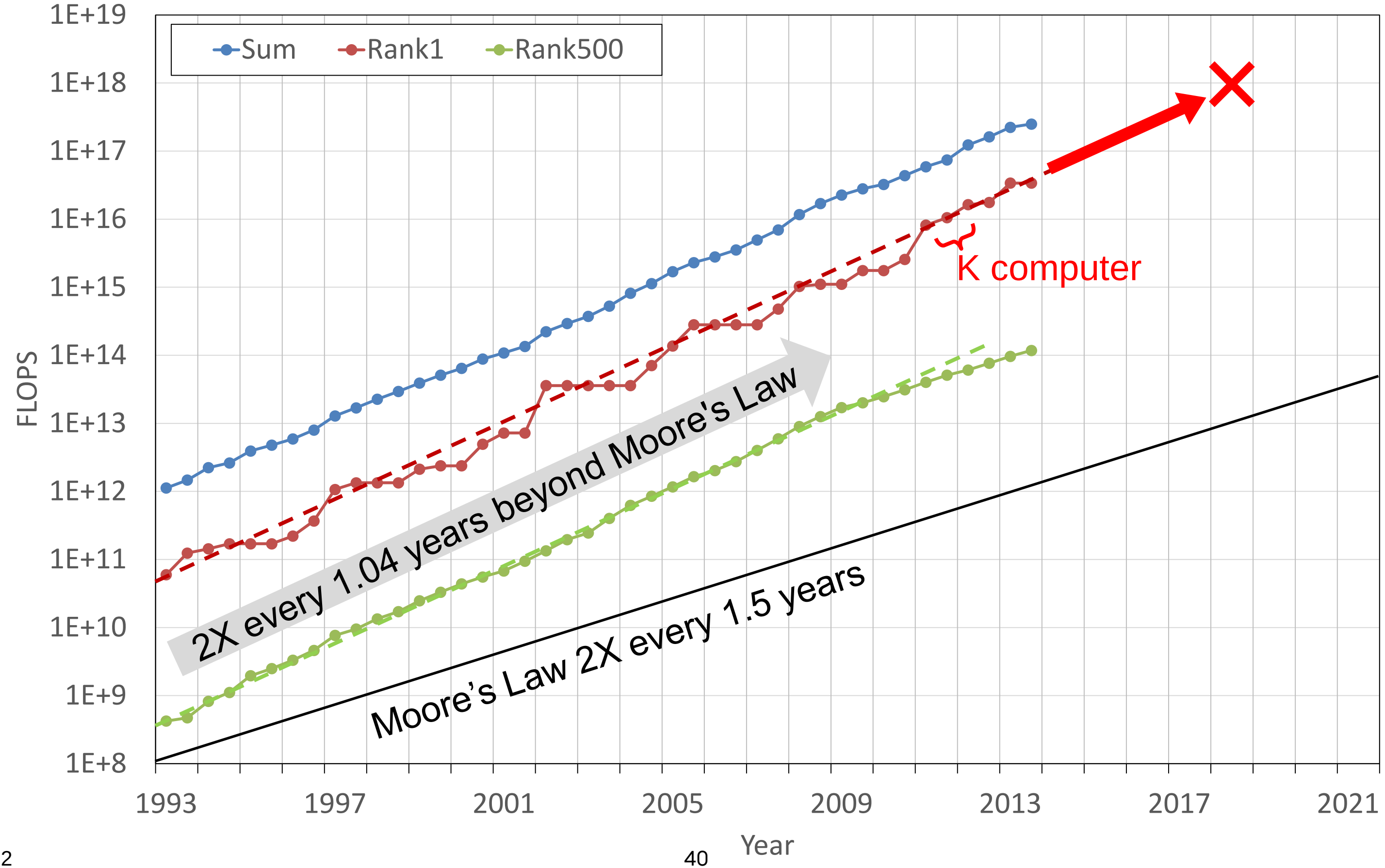
Timeline of Fugaku Development

- After the SDHPC White Paper and the feasibility study, Fugaku development started in 2014
  - SDHPC was a community activity to discuss strategies for developing HPC systems

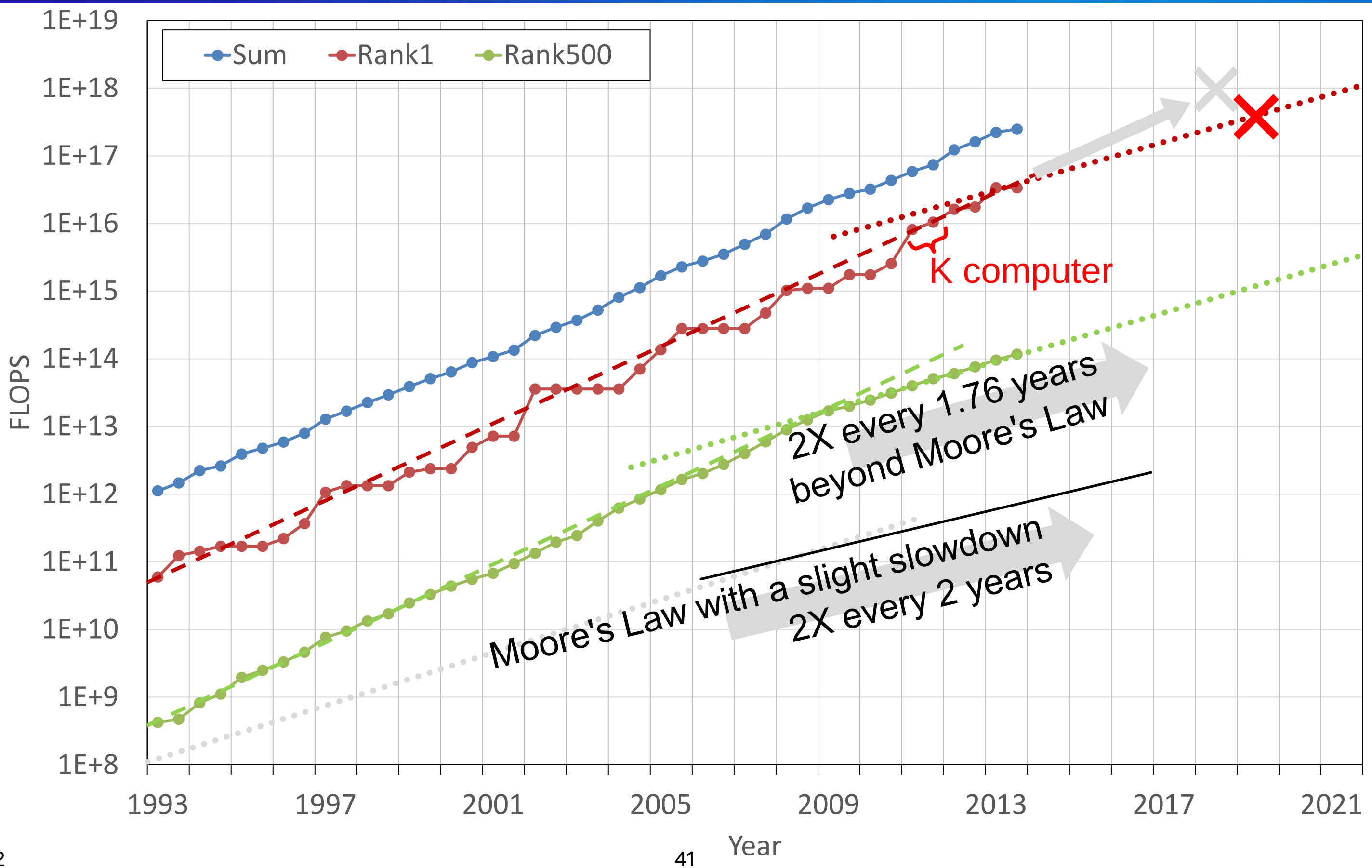


K computer project started
K computer ranked #1
K computer full operation started
K computer shut down
NGACI White Paper released
K computer development
2006
2010
2011
2012
2013
2014
2018
2019
2020
2021
SDHPC White Paper released
Post-K feasibility study started
Post-K feasibility study report
Post-K development started
Fugaku development
Post-K production approved
Post-K production started
Fugaku ranked #1
Fugaku full operation started

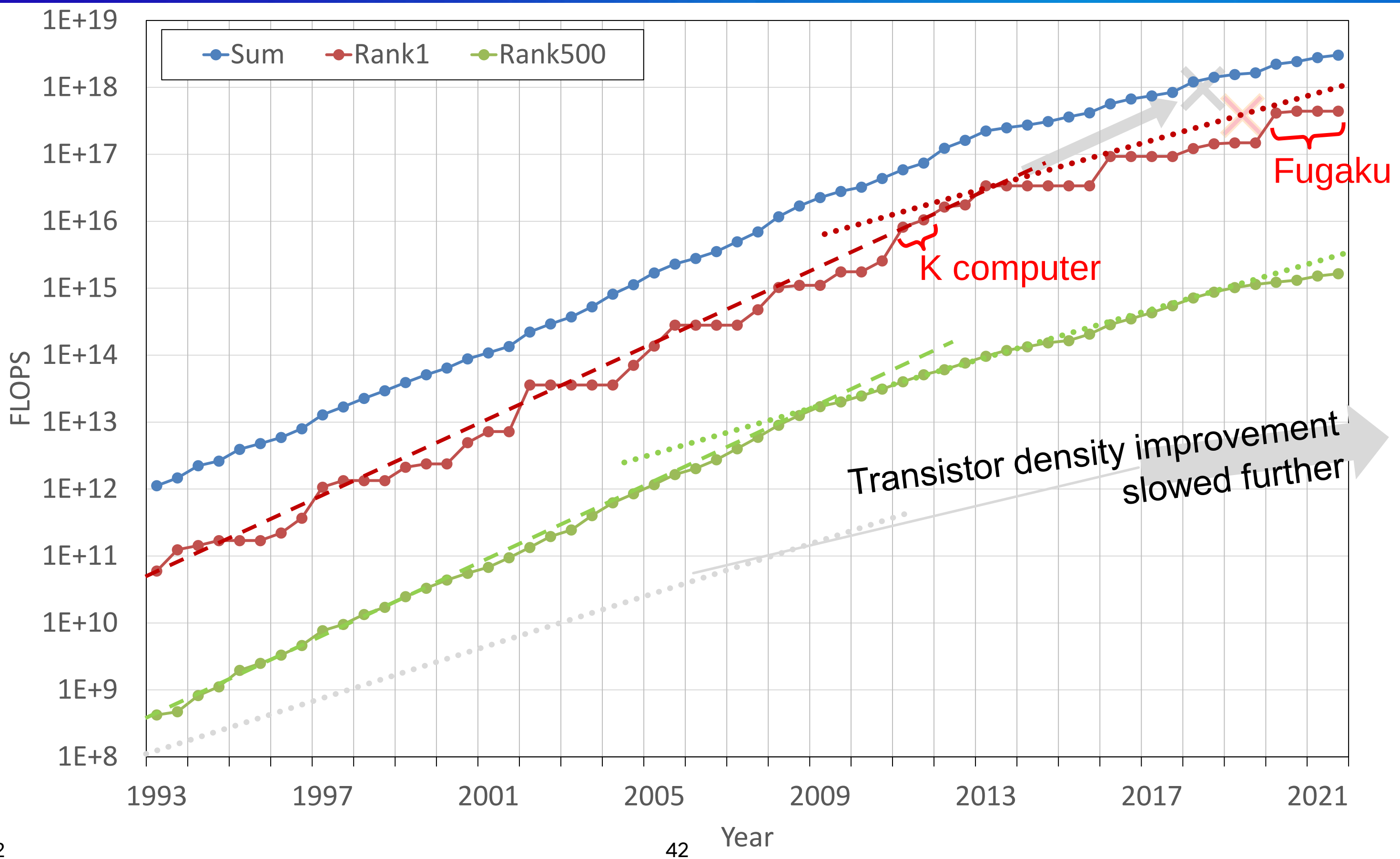
Projection at the Time of Feasibility Study



Prediction in SDHPC White Paper



Target Achieved with a One-Year Delay



Transistor density improvement slowed further

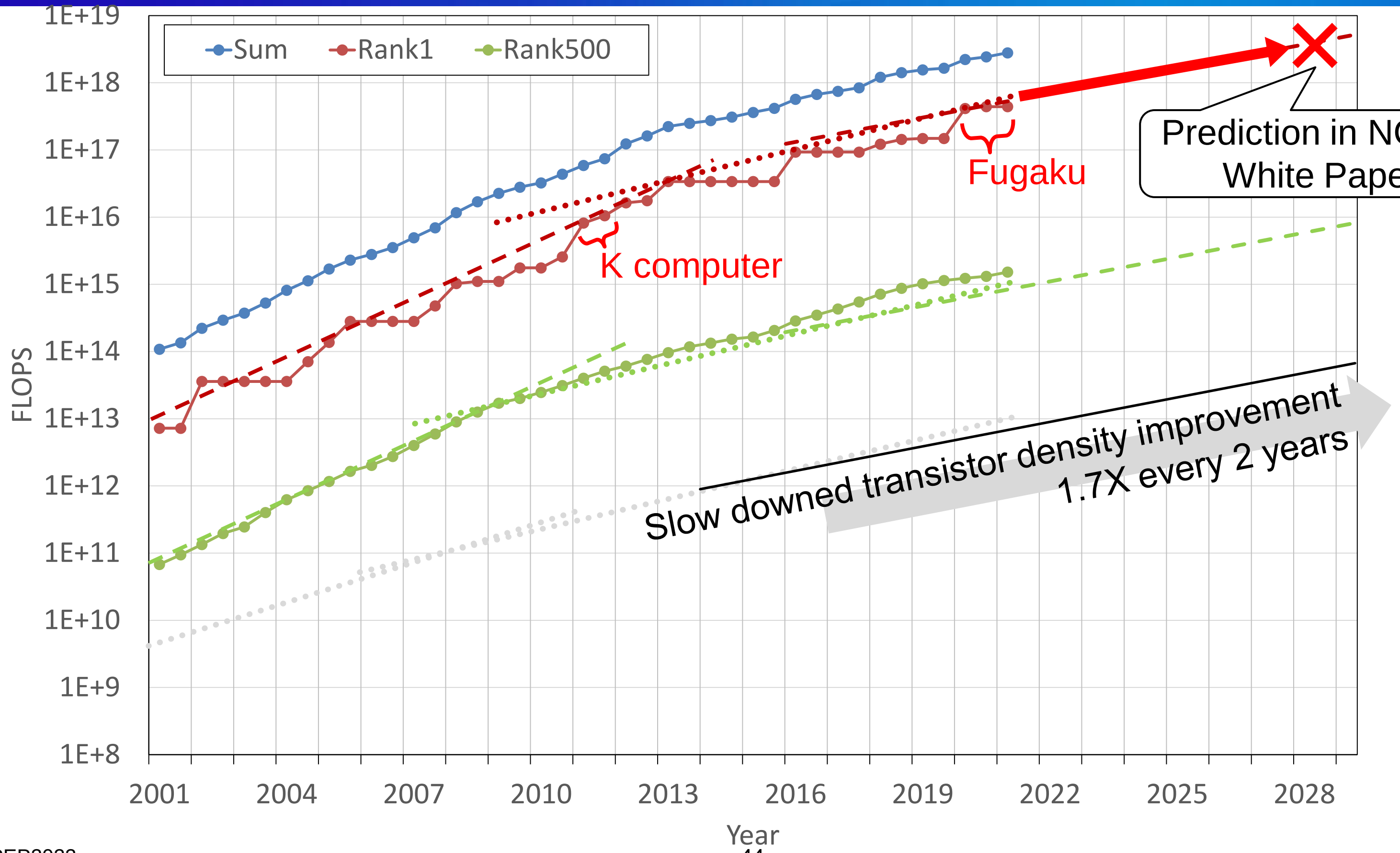
NGACI White Paper

- Next-Generation Advanced Computing Infrastructure
  - Overview and Objectives

In considering the sustainable development of high-performance computing in the future, we can expect further developments such as further integration with AI and Big Data technologies and deployment in new application fields such as Society 5.0, but it is also true that many technological challenges, such as the end of Moore's Law, lie ahead. This activity (NGACI) is a forum for open exchange of ideas and opinions on the technical issues that need to be addressed for future high-performance computing environments and for shared computing infrastructure, what kind of research and development is needed, and what kind of activities should be conducted as a community, and to summarize these ideas as White The purpose is to contribute to the development of this field by exchanging opinions openly and summarizing them as White Papers.

- Activities
  - Number of registered community members: > 100
  - Four working groups discussed future system vision and issues
  - White Paper 1.0.0 (164 pages) released https://sites.google.com/view/ngaci/home

Prediction in NGACI White Paper



Feasibility Study to Begin This Year

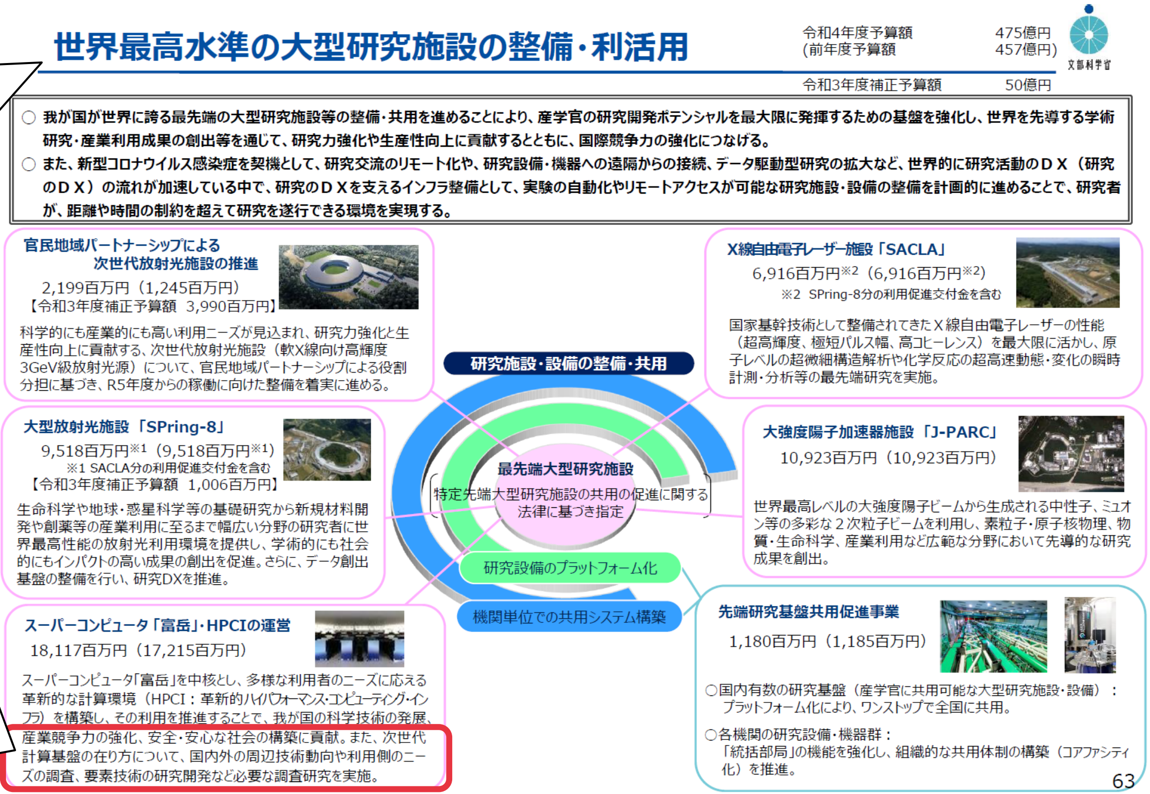
Ministry of Education, Culture, Sports, Science and Technology

Development and Utilization of World-Class Large-Scale Research Facilities

FY2022 Budget 45.7B JPY (376M USD)

Conduct necessary research and studies on the ideal **next-generation computing infrastructure**, including surveys of domestic and international trends in technologies and user needs, and research and development of elemental technologies

FY2022 Budget 0.4B JPY (in page 3)



世界最高水準の大型研究施設の整備・利活用

令和4年度予算額 475億円
(前年度予算額 457億円)

令和3年度補正予算額 50億円

文部科学省

○ 我が国が世界に誇る最先端の大型研究施設等の整備・共用を進めることにより、産学官の研究開発ポテンシャルを最大限に発揮するための基盤を強化し、世界を先導する学術研究・産業利用成果の創出等を通じて、研究力強化や生産性向上に貢献するとともに、国際競争力の強化につなげる。

○ また、新型コロナウイルス感染症を契機として、研究交流のリモート化や、研究設備・機器への遠隔からの接続、データ駆動型研究の拡大など、世界的に研究活動のＤＸ（研究のＤＸ）の流れが加速している中で、研究のＤＸを支えるインフラ整備として、実験の自動化やリモートアクセスが可能な研究施設・設備の整備を計画的に進めることで、研究者が、距離や時間の制約を超えて研究を遂行できる環境を実現する。

官民地域パートナーシップによる次世代放射光施設の推進

2,199百万円（1,245百万円）
【令和3年度補正予算額 3,990百万円】

科学的にも産業的にも高い利用ニーズが見込まれ、研究力強化と生産性向上に貢献する、次世代放射光施設（軟X線向け高輝度3GeV級放射光源）について、官民地域パートナーシップによる役割分担に基づき、R5年度からの稼働に向けた整備を着実に進める。

大型放射光施設「SPring-8」

9,518百万円[※1]（9,518百万円[※1]）
※1 SACLA分の利用促進交付金を含む
【令和3年度補正予算額 1,006百万円】

生命科学や地球・惑星科学等の基礎研究から新規材料開発や創薬等の産業利用に至るまで幅広い分野の研究者に世界最高性能の放射光利用環境を提供し、学術的にも社会的にもインパクトの高い成果の創出を促進。さらに、データ創出基盤の整備を行い、研究DXを推進。

スーパーコンピュータ「富岳」・HPCIの運営

18,117百万円（17,215百万円）

スーパーコンピュータ「富岳」を中核とし、多様な利用者のニーズに応える革新的な計算環境（HPCI：革新的ハイパフォーマンス・コンピューティング・インフラ）を構築し、その利用を推進することで、我が国の科学技術の発展、産業競争力の強化、安全・安心な社会の構築に貢献。また、次世代計算基盤の在り方について、国内外の周辺技術動向や利用側のニーズの調査、要素技術の研究開発など必要な調査研究を実施。

研究施設・設備の整備・共用

最先端大型研究施設

〔特定先端大型研究施設の共用の促進に関する法律に基づき指定〕

研究設備のプラットフォーム化

機関単位での共用システム構築

X線自由電子レーザー施設「SACLA」

6,916百万円[※2]（6,916百万円[※2]）
※2 SPring-8分の利用促進交付金を含む

国家基幹技術として整備されてきたＸ線自由電子レーザーの性能（超高輝度、極短パルス幅、高コヒーレンス）を最大限に活かし、原子レベルの超微細構造解析や化学反応の超高速動態・変化の瞬時計測・分析等の最先端研究を実施。

大強度陽子加速器施設「J-PARC」

10,923百万円（10,923百万円）

世界最高レベルの大強度陽子ビームから生成される中性子、ミュオン等の多彩な２次粒子ビームを利用し、素粒子・原子核物理、物質・生命科学、産業利用など広範な分野において先導的な研究成果を創出。

先端研究基盤共用促進事業

1,180百万円（1,185百万円）

○国内有数の研究基盤（産学官に共用可能な大型研究施設・設備）：プラットフォーム化により、ワンストップで全国に共用。

○各機関の研究設備・機器群：「統括部局」の機能を強化し、組織的な共用体制の構築（コアファシティ化）を推進。

63

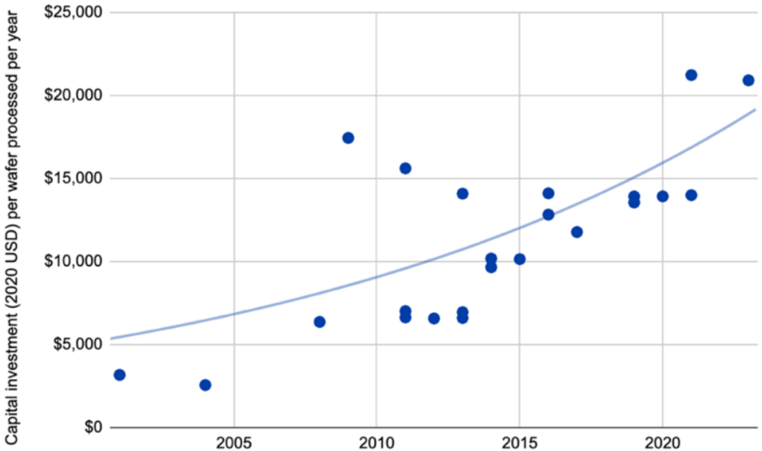
https://www.mext.go.jp/content/20211223-mxt_kouhou02-000017672_1.pdf

Emerging Technologies for Future System

Increasing Cost of Processes Technologies

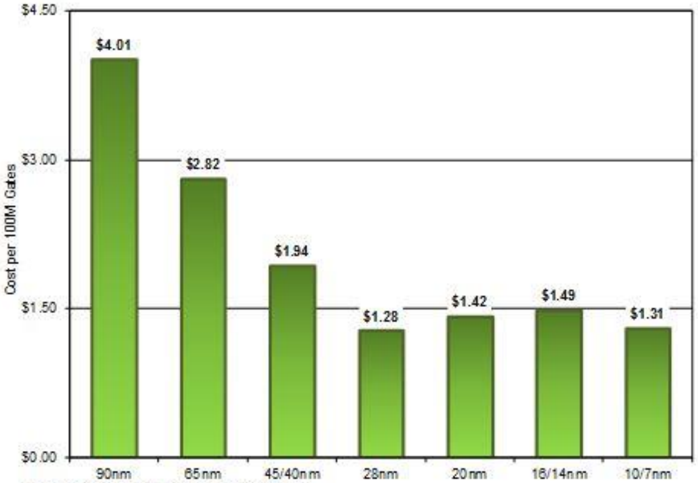
- Investment per wafer is rising exponentially
- No significant reduction in cost per gate after 28nm process, even though densities are increasing

Figure 13: Capital investment per 300 mm wafer processed per year[212]



https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/

Gate Cost

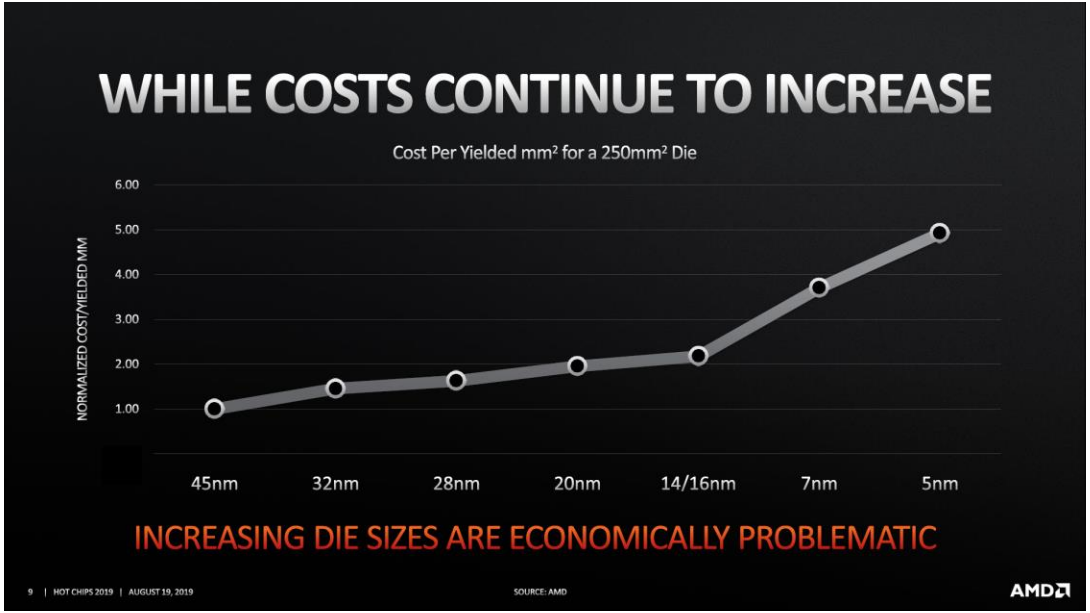


Source: International Business Strategies, Inc.

https://www.semi.org/en/semiconductor-industry-2015-2025

Chiplet to Counter Increasing Die Cost

- Manufacturing large die is no longer economical
- Increasing yield with “chiplets” becomes important



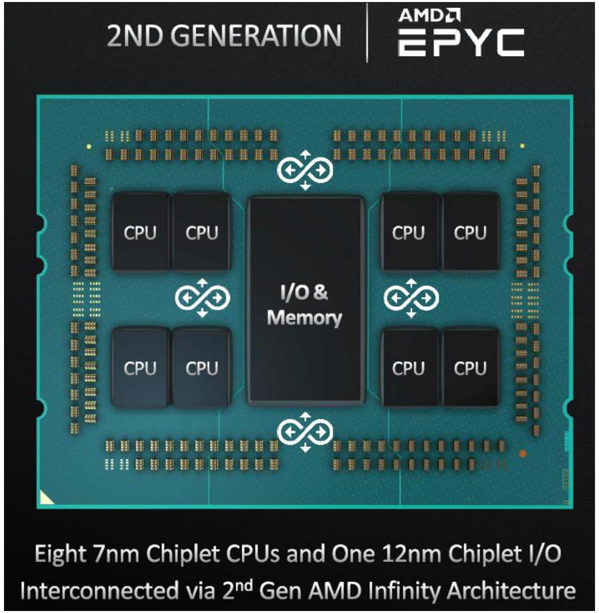
WHILE COSTS CONTINUE TO INCREASE

INCREASING DIE SIZES ARE ECONOMICALLY PROBLEMATIC

9 | HOT CHIPS 2019 | AUGUST 19, 2019

SOURCE: AMD

AMD



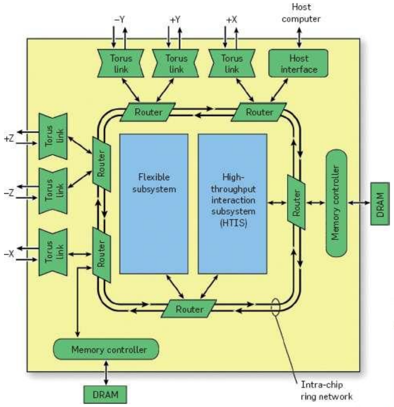
2ND GENERATION | AMD EPYC

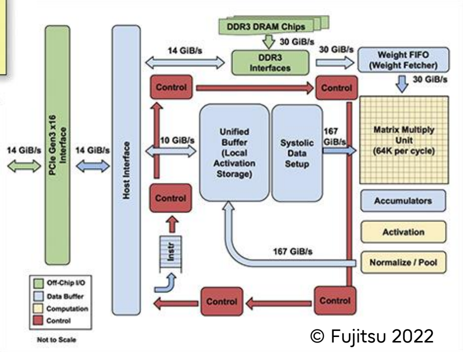
Eight 7nm Chiplet CPUs and One 12nm Chiplet I/O Interconnected via 2nd Gen AMD Infinity Architecture

https://old.hotchips.org/hc31/Hot_Chips_2019_DrLisaSu_AMD_0819.pdf

Domain-Specific Architecture

- Specialized circuits that execute specific workload at high throughput are becoming more important
  - Transistor density improvement has slowed down and small circuits are required
- Application-specific architecture
  - e.g., Anton for molecular dynamics
    - Ultra-low latency torus network
- Domain-specific architecture
  - e.g., Google TPU for Deep Learning
    - Low-precision, high-throughput systolic array

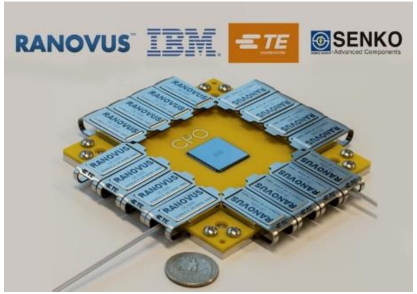


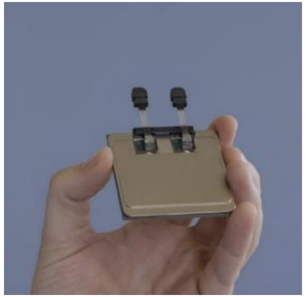


High-Density, High-Speed Transmission

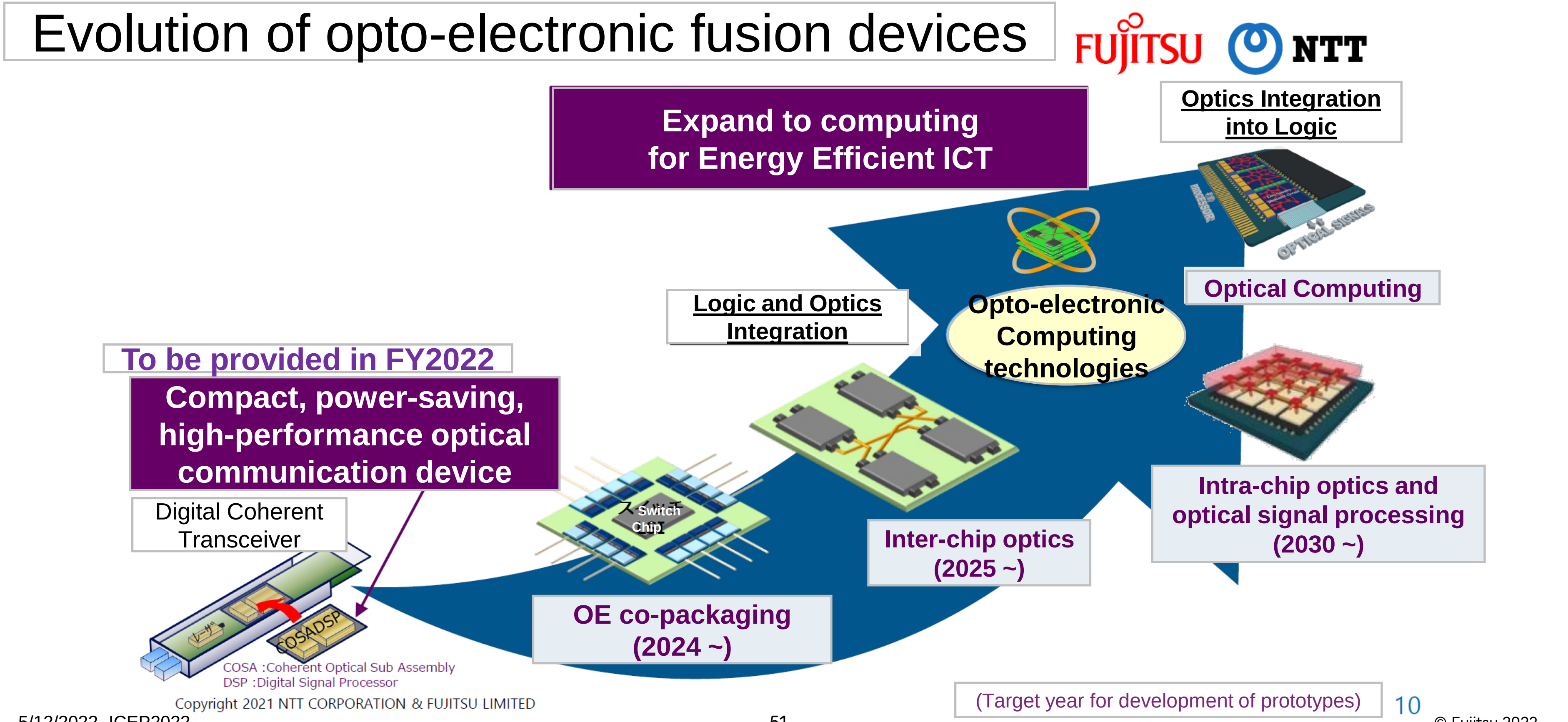
- Insertion loss on PCB is severe at >50 Gbps / lane
  - Retimers and active electrical cables
  - Fly-over cable connection

- Co-packaged modules and connectors at >100 Gbps / lane
  - CPO: Co-Packaged Optics, CPE: Co-Packaged Electronics





Roadmap for Future Optical Technologies



Evolution of opto-electronic fusion devices

Expand to computing for Energy Efficient ICT

Optics Integration into Logic

Optical Computing

Logic and Optics Integration

Opto-electronic Computing technologies

To be provided in FY2022

Compact, power-saving, high-performance optical communication device

Digital Coherent Transceiver

OE co-packaging (2024 ~)

Inter-chip optics (2025 ~)

Intra-chip optics and optical signal processing (2030 ~)

COSA :Coherent Optical Sub Assembly
DSP :Digital Signal Processor

(Target year for development of prototypes)

Summary

- Supercomputers are the key infrastructure for the SDGs
- System scale and density continue to increase
- Fugaku is the #1 system in four major benchmarks
- System architecture and structure of Fugaku
- Other top-level supercomputers in recent years
- Timeline of Fugaku development and NGACI
- Emerging technologies for future systems
  - Chiplet, Domain-Specific Architecture, Co-Packaged Optics

Thank you