大学共同利用機関法人 自然科学研究機構
核融合科学研究所
National Institute for Fusion Science

調和ある多様性の創造 国立研究開発法人
QST 量子科学技術研究開発機構
National Institutes for Quantum Science and Technology

# 新プラズマシミュレータシステムの概要

佐竹真介

核融合科学研究所　構造形成・持続性ユニット

研究データエコシステム「プラズマ・核融合クラウド」構築と関連技術の高度化　同時開催：クラウドを利用したQUEST装置実験のデータ解析環境構築

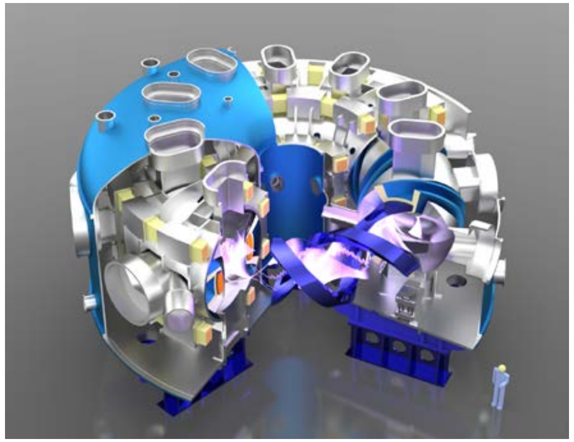
2025/2/25　@ NIFS

# OUTLINE

1. About NIFS and QST
2. Replacement of new supercomputer system in 2025
3. Prospects for future fusion research using the new supercomputer

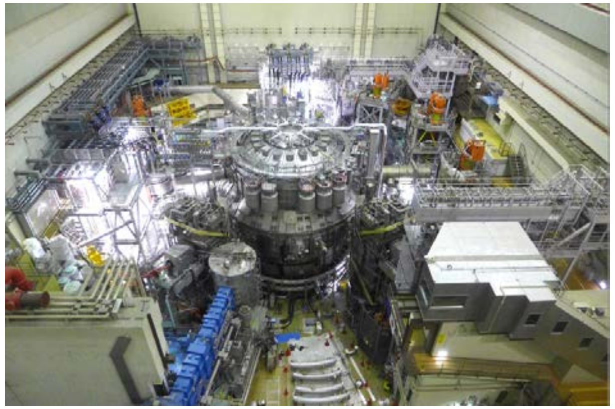
# 1. About NIFS and QST

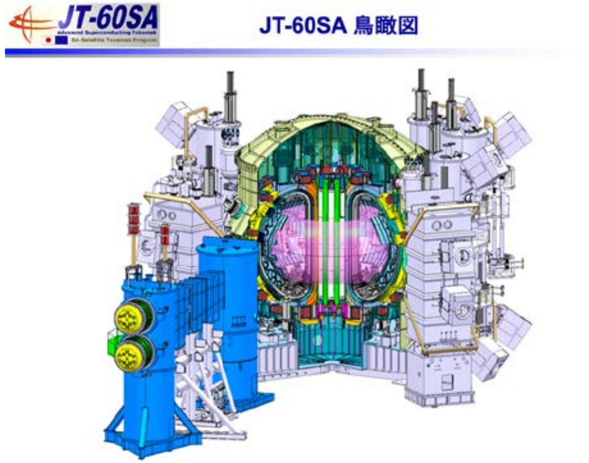
- NIFS (National Institute for Fusion Science)
  - One of the institutes in National Institutes of Natural Sciences, Japan
  - One of the international center of excellence for helical fusion reactor research
  - LHD (Large Helical Device) : One of the world-largest plasma confinement device ( shut down in 2025 )
  - Plasma Simulator : Supercomputer for simulation studies of fusion plasmas and related fields. Shared by collaborators in many domestic universities.



# 1. About NIFS and QST

- QST (National Institutes for Quantum Science and Technology)
  - Fusion research division (Formerly known as JAEA)
  - International scientific research & development center on fusion plasma, mainly tokamak-type.
  - JT-60SA (start operation 2023) : a joint Japanese-European project, carried out in parallel with the ITER program, for the early realization of fusion energy.
  - JFRS-1 : Supercomputer operated by International Fusion Energy Research Centre (IFERC), for fusion research community in Japan and EU.





# 2. Replacement of new supercomputer system

Precent system

NIFS (Plasma Simulator)
NEC SX-AURORA Tsubasa
540node x 8 Vector Engine
10.5 PF / 202TiB HBM2

QST (JFRS-1)
CRAY XC50
1370node x 2 Intel Xeon Gold 6148
4.2 PF / 256TB DDR4-2666

New Plasma Simulator (2025/7)

Joint operation by NIFS & QST
3 systems by NEC

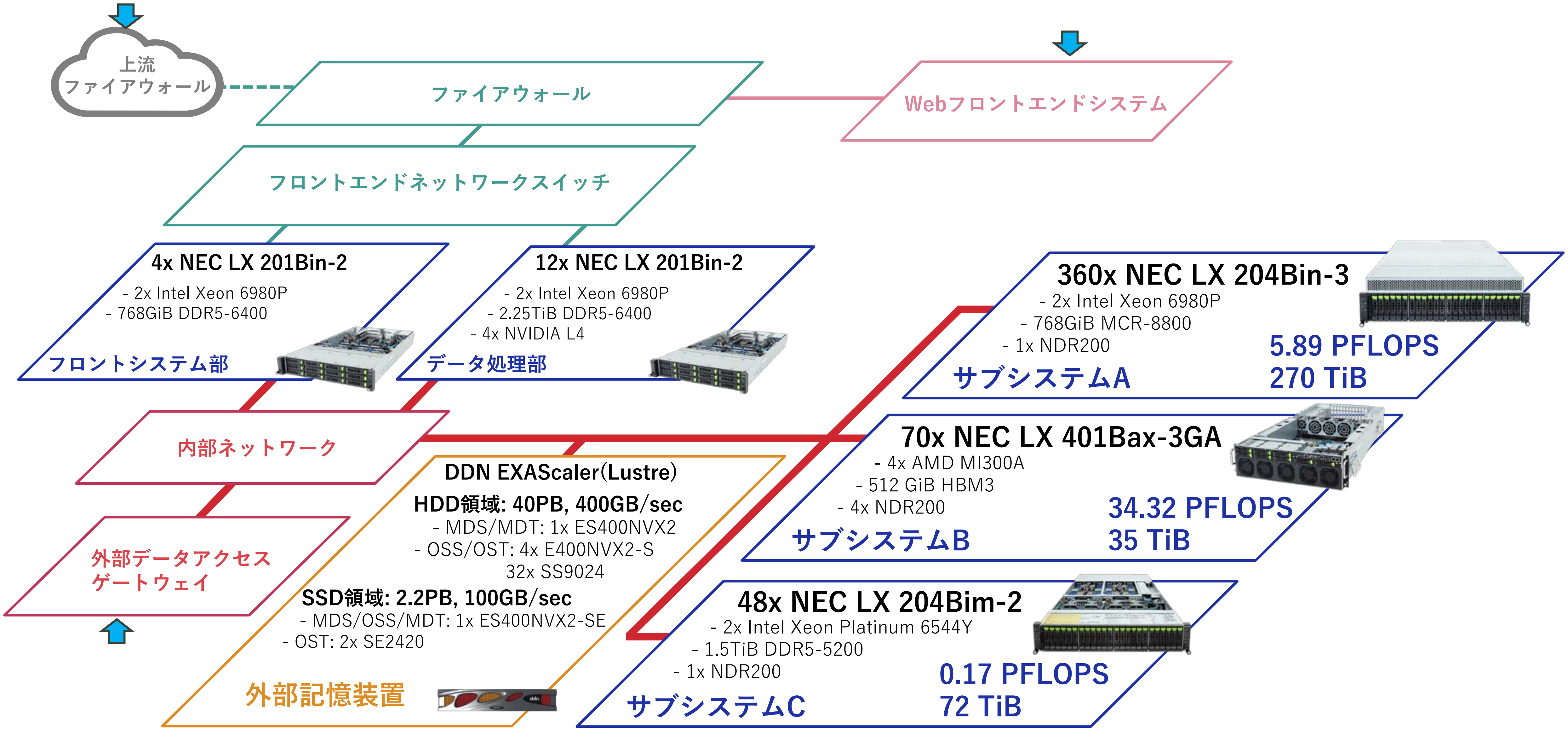
- Subsystem A
  360node x 2 Intel Xeon 6980P*
  5.90 PF / 270TiB MCR-8800 MRDIMM*
- Subsystem B
  70node x 4 AMD MI300A APU*
  34.3PF / 35TiB HBM3
- Subsystem C
  48node x 2 Intel Platinum 6544Y
  0.17PF / 72TiB DDR5-5200

Internal network : Infiniband NDR200

*First introduced in Japan

# システム全体概念図

⬇: 外部ネットワークからのアクセスポイント



上流
ファイアウォール

ファイアウォール

Webフロントエンドシステム

フロントエンドネットワークスイッチ

4x NEC LX 201Bin-2
- 2x Intel Xeon 6980P
- 768GiB DDR5-6400

フロントシステム部

12x NEC LX 201Bin-2
- 2x Intel Xeon 6980P
- 2.25TiB DDR5-6400
- 4x NVIDIA L4

データ処理部

360x NEC LX 204Bin-3
- 2x Intel Xeon 6980P
- 768GiB MCR-8800
- 1x NDR200

サブシステムA

5.89 PFLOPS
270 TiB

内部ネットワーク

70x NEC LX 401Bax-3GA
- 4x AMD MI300A
- 512 GiB HBM3
- 4x NDR200

サブシステムB

34.32 PFLOPS
35 TiB

外部データアクセス
ゲートウェイ

DDN EXAScaler(Lustre)

HDD領域: 40PB, 400GB/sec
- MDS/MDT: 1x ES400NVX2
- OSS/OST: 4x E400NVX2-S
32x SS9024

SSD領域: 2.2PB, 100GB/sec
- MDS/OSS/MDT: 1x ES400NVX2-SE
- OST: 2x SE2420

外部記憶装置

48x NEC LX 204Bim-2
- 2x Intel Xeon Platinum 6544Y
- 1.5TiB DDR5-5200
- 1x NDR200

サブシステムC

0.17 PFLOPS
72 TiB

## ➢ Subsystem A

per node :

- 2 x Intel Xeon 6980P (128core, 2.0GHz) 8.19TFLOPS, 504MB cache
- 24 x 32GiB=768GiB MCR-8800 MRDIMM
  - Memory band width per node : 1.7 TB/s
- Network : Infiniband NDR200 x 1port

System total : 5.90 PF / 270TiB

Full bisection fat-tree topology network, 25GB/s b/w 2 nodes

Comparison: SX AURORA Tsubasa (540 nodes)

- 8x Vector Engine (type 10AE) 2.43TFLOPS, 16MB cache
- 48GB HBM2, Memory band width 1.35TB/s
- 2 x Infiniband HDR200, 50GB/s b/w 2 nodes

## ➢ Subsystem B

per node :

4 x AMD MI300A APU

CPU : 24 core EPYC Zen4, 3.7GHz (no data, 0.7-1TFLOPS?)

GPU : FP64 123TFLOPS, 2.1GHz, 256 MB cache

4 x 128GiB =512GiB HBM3 memory (shared by CPU & GPU)

Memory band width per APU : 5.3 TB/s

Interconnect : Infinity Fabric 128 GB/s (x 2 x 3?)

Network : Infiniband NDR200 x 4port

System total : 34.32 PF / 35TiB

Full bisection fat-tree topology network, 25GB/s b/w arbitrary 2 nodes

## ➢ Subsystem C

per node :

- 2 x Intel Platinum 6544Y (16core, 3.6GHz) 3.69TFLOPS, 45MB cache
- 16 x 96GiB=1536GiB DDR5-5200 memory
  - Memory band width per node : 666GB/s
- Network : Infiniband NDR200 x 1port

System total : 0.17 PF / 72TiB

Full bisection fat-tree topology network, 25GB/s b/w 2 nodes

注：検討中の内容を含む

# システムの利用イメージ

## (1) 大規模並列演算部（サブシステムA/B/C）の利用

**利用シナリオ１：SSHでのバッチ利用**

**利用シナリオ２：SSHでのインタラクティブ利用**

**利用シナリオ３：Webブラウザでのバッチ利用**

**利用シナリオ４：Webブラウザでのインタラクティブ利用**
Webブラウザで、計算サーバ上で実行されているJupyterを操作

# A little analysis on the performance of new & old machines

| Computation unit | TFLOPS (A) | MBW[TB/s] (B) | B/A | Memory[GB] (C) | C/A | cache [MB] (D) | D/A | Network BW [GB/s] per unit | unit x node |
|---|---|---|---|---|---|---|---|---|---|
| Xeon 6980P (Subsystem A) | 8.19 | 0.845 | 0.103 | 384 | 46.9 | 504 | 61.5 | 12.5 | 2 x 360 |
| B: MI300A (Subsystem B) | 122.6 | 5.3 | 0.043 | 128 | 1.04 | 256 | 2.09 | 25.0 / 128(inside) | 4 x 70 |
| **VE Type 10AE** | 2.43 | 1.35 | 0.556 | 48 | 19.8 | 16 | 6.58 | 6.25 | 8 x 540 |
| Xeon Gold 6148 | 1.54 | 0.128 | 0.083 | 96 | 62.3 | 27 | 17.5 | 1.9~3.3 (?) | 2 x 1370 |

- **Vector Engine** has good MBW / FLOPS and cache / FLOPS for a machine built 4 years ago. High execution efficiency is achieved for well vector-tuned codes.

- However, because of small memory size, large MPI parallel numbers is required to run large-scale simulations, and the narrow network band width can be the rate-limiting factor.

# A little analysis on the performance of new & old machines

| Computation unit | TFLOPS (A) | MBW[TB/s] (B) | B/A | Memory[GB] (C) | C/A | cache [MB] (D) | D/A | Network BW [GB/s] per unit | unit x node |
|---|---|---|---|---|---|---|---|---|---|
| **Xeon 6980P (Subsystem A)** | 8.19 | 0.845 | 0.103 | 384 | 46.9 | 504 | 61.5 | 12.5 | 2 x 360 |
| B: MI300A (Subsystem B) | 122.6 | 5.3 | 0.043 | 128 | 1.04 | 256 | 2.09 | 25.0 / 128(inside) | 4 x 70 |
| VE Type 10AE | 2.43 | 1.35 | 0.556 | 48 | 19.8 | 16 | 6.58 | 6.25 | 8 x 540 |
| Xeon Gold 6148 | 1.54 | 0.128 | 0.083 | 96 | 62.3 | 27 | 17.5 | 1.9~3.3 (?) | 2 x 1370 |

- **Subsystem A** has large memory and MBW/FLOPS is better than subsystem B thanks to the adoption of rapid MRDIMM memory.

- The total cache size is large, cache / core ≒ 4MB is improved from 1.35MB in present QST supercomputer. Cache tuning is likely to be the key to high-performance.

- Large-scale simulation running on present VE system will efficiently run on Sub A, but code tuning for many-core (128C/node) architecture will be important.

# A little analysis on the performance of new & old machines

| Computation unit | TFLOPS (A) | MBW[TB/s] (B) | B/A | Memory[GB] (C) | C/A | cache [MB] (D) | D/A | Network BW [GB/s] per unit | unit x node |
|---|---|---|---|---|---|---|---|---|---|
| Xeon 6980P (Subsystem A) | 8.19 | 0.845 | 0.103 | 384 | 46.9 | 504 | 61.5 | 12.5 | 2 x 360 |
| **B: MI300A (Subsystem B)** | 122.6 | 5.3 | 0.043 | 128 | 1.04 | 256 | 2.09 | 25.0 / 128(inside) | 4 x 70 |
| VE Type 10AE | 2.43 | 1.35 | 0.556 | 48 | 19.8 | 16 | 6.58 | 6.25 | 8 x 540 |
| Xeon Gold 6148 | 1.54 | 0.128 | 0.083 | 96 | 62.3 | 27 | 17.5 | 1.9~3.3 (?) | 2 x 1370 |

- **Subsystem B** has large MBW and cache, but MBW/FLOPS and cache/FLOPS are the smallest among 4 systems. To achieve high execution speed, thorough cache tuning of the code will be essential.

- High network BW among APUs & nodes will help to efficiently run large simulations with multiple APUs in parallel (MPI will be the rate-limiting factor for multi-node calculation, though).

# 3. Prospects for future fusion research using the new supercomputer

- So far, the main users of supercomputers in fusion research community mainly uses FORTRAN90 with MPI and OpenMP (C / C++ codes also exists).

- In Japan, the large-scale simulations (mainly 5-dimensional gyrokinetic codes) have been developed and tuned for Japan-oriented computer architecture such as NEC SX (vector engine) and Fujitsu KEI and FUGAKU (Arm processor).

- While the adaptation of existing plasma simulation codes for GPU machines has been progressing in Europe and the United States, we has been relatively inactive in this field, because we have good supercomputer systems as above.

- Japanese fusion research community need to catch up with the global GPU computing trend.

# 3. Prospects for future fusion research using the new supercomputer

- The new NIFS & QST supercomputer provides the platform and opportunity to our research community to adapt, transplant, and develop simulation codes running on GPU machines.
  - For the first half year, NIFS research collaborators can use subsystem B (GPU nodes) without time limitation.
  - As a new attempt, a job class is planned to be set up to enable the execution of python code (including interactive execution) on large computation nodes.
  - Tutorial seminar for GPU computing and technical support to transplant simulation codes to new system will be provided.
- For the use of computers in data science, we prepare a direct connection to refer to experimental data base such as JT-60SA from the new supercomputers.

# GPU can be a game-changer of fusion plasma simulation research

**Example : GX code (GPU-native gyrokinetic code for tokamaks and stellarators)**
Developed by researchers in Princeton Plasma Physics Lab. and others.

Gyrokinetic simulation is most heavy simulation in fusion plasma research and is important to predict the heat and particle confinement ability of torus plasma.
So far, gyrokinetic simulation to predict nonlinear saturation level of turbulence requires O(100) CPU cores x O(100) hours.

Newly developed GPU-native code GX demonstrates that such calculation can be done only on O(1) GPU x O(1) hours. Efficient algorithms for GPU (ex. single-precision FFT) are chosen to achieve the high efficiency.

NIFS researchers are contacting with the GX team to introduce the code to new system.
(Now code is written with Nvidia CUDA, but transplant to AMD system is ongoing)