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
量子科学技術研究開発機構
National Institutes for Quantum Science and Technology

Deployment of a New Supercomputer System for Fusion Research at NIFS and QST

SHINSUKE SATAKE
NATIONAL INSTITUTE FOR FUSION SCIENCE

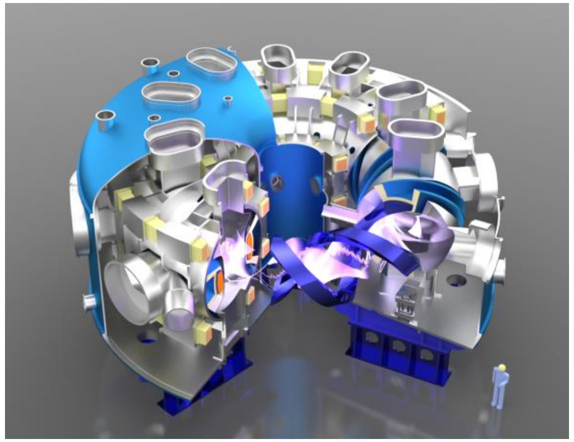
NEC USER GROUP MEETING(NUG) XXXVI, SENRI LIFE SCIENCE CENTER, 2025/5/13

OUTLINE

1. About NIFS and QST
2. Computer simulation research in fusion science
3. Replacement of new supercomputer system in 2025
4. Simulation code porting to AMD MI300A APUs

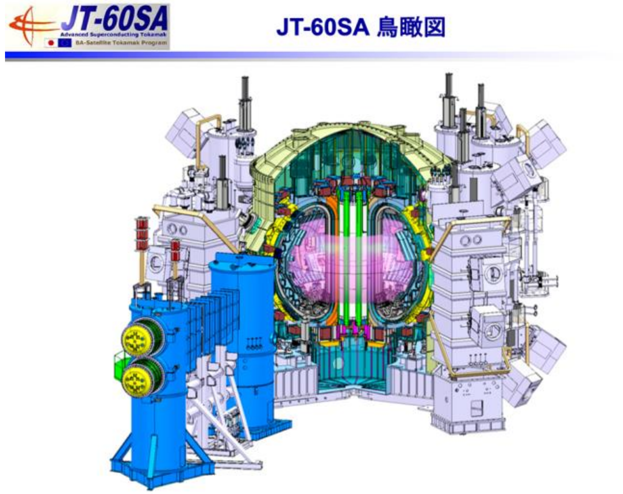
1. About NIFS and QST

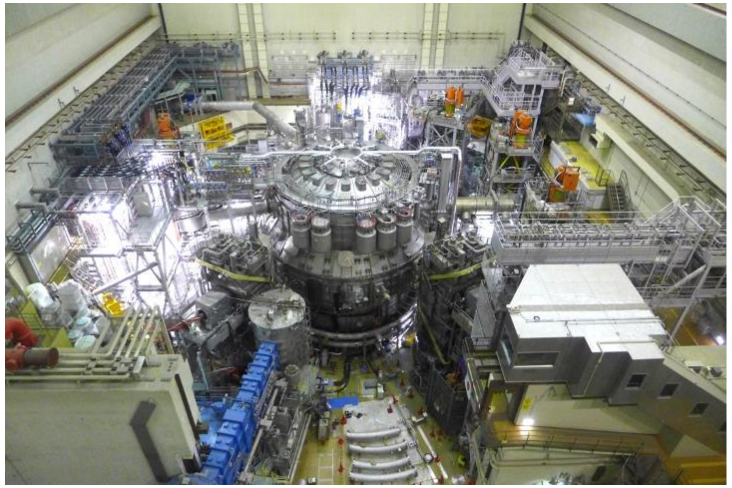
- **NIFS (National Institute for Fusion Science)**
  - One of the institutes in National Institutes of Natural Sciences (NINS), Japan
  - International center of excellence for plasma and fusion research
  - LHD (Large Helical Device) : One of the world-largest helical plasma confinement device (shut 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)**
  - Naka & Rokkasho institutes for fusion research (Formerly known as JAEA)
  - International scientific research & development center on fusion plasma, mainly tokamak-type.
  - JT-60SA (Naka, start operation 2023) : Tokamak device by a joint Japanese-European project. Carried out in parallel with the ITER program for the early realization of fusion energy.
  - JFRS-1 (Rokkasho) :
    Supercomputer in International Fusion Energy Research Centre (IFERC), for fusion research community in Japan and EU.





2. Simulation research in fusion science

- ◆Fusion research is one of the areas where large-scale numerical simulations have been active from a relatively early stage.
- ◆Applications of computer simulation in fusion science are wide-ranging.
  - ➢ Optimized design of confinement magnetic field
  - ➢ Desing of magnetic coils and vessel
  - ➢ Stability analysis and dynamics of magnetically-confined plasma (Magneto-Hydro-Dynamics, MHD)
  - ➢ Heat and particle transport processes in plasma (turbulence / collisional diffusion)
  - ➢ Interactions of plasma and walls (sputtering, erosion, etc.)
  - ➢ Simulation of plasma heating by microwave & neutral beam injection
  - ➢ Prediction and control of steady-state operation

Multi-scale, Multi-dimension, Multi-physics simulations are required

3. Replacement of new supercomputer system

Present 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**

➢ Subsystem A

per node :

- 2 x Intel Xeon 6980P (128core, 2.0GHz) 8.19TFLOPS, 504MB cache
- 24 x 32=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

128 x 4=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(25GB/s) x 4port

System total : 34.3PF / 35TiB

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

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
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

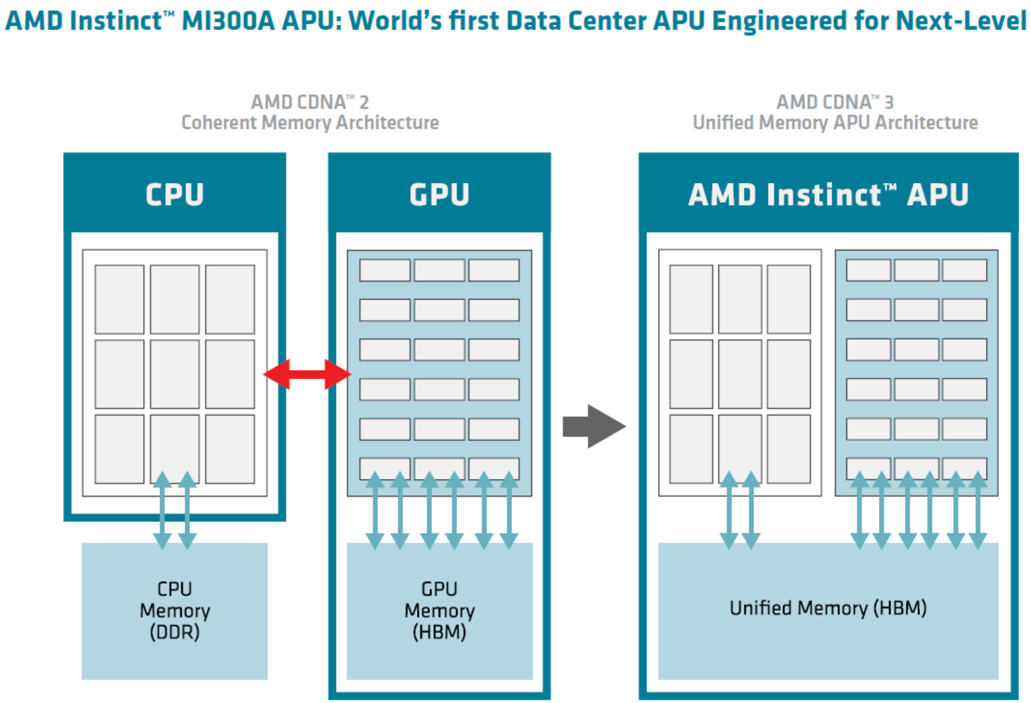
- Subsystem A has large memory and MBW/FLOPS is better than subsystem B thanks to the adoption of rapid MRDIMM memory.
- The total cache size is large, cache / core ≒ 4MB is improved from 1.35MB in present QST supercomputer. Cache tuning is likely to be the key to high-performance.
- Large-scale simulation running on present VE system will efficiently run on Sub A, but code tuning for many-core architecture (128C/CPU) will be important.

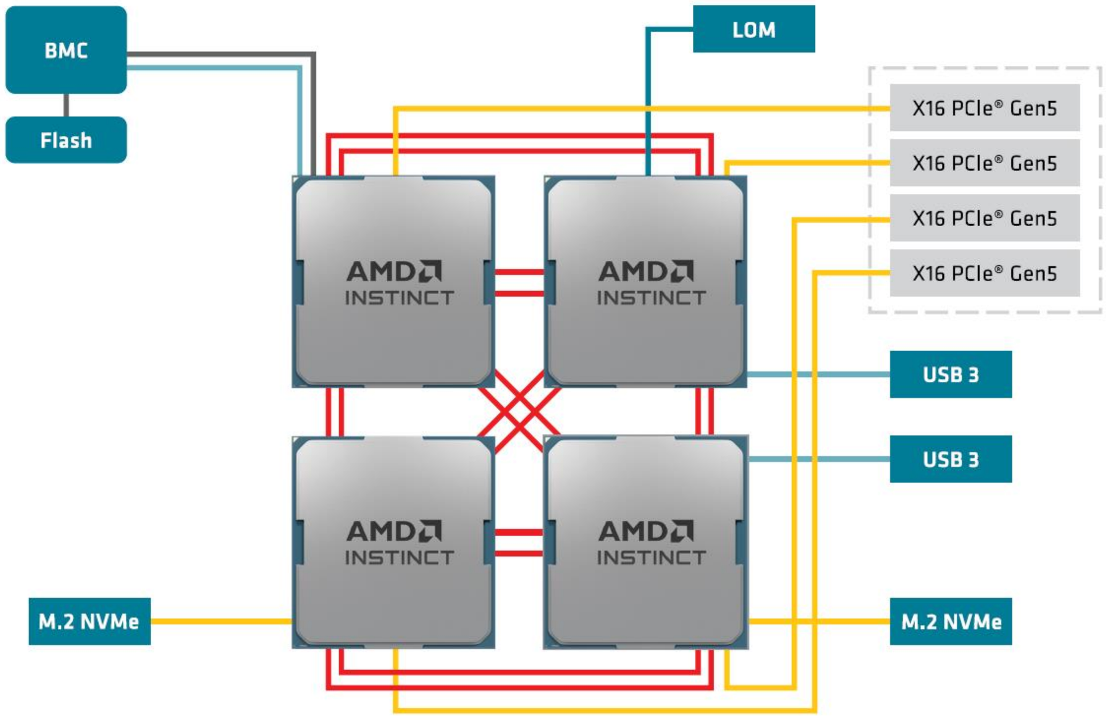
Analysis on the performance of new & old machines

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
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 the 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).





- On MI300A APU, both CPU and GPU share the HBM3 unified memory.
- Programmers do not need to care the location of the arrays on the memory when offloading a calculation to GPU.

- Four MI300A chips in one node are connected by the fast network (AMD infinity fabric).
- Each APU is equipped with an Infiniband NDR port.
- Coding taking account of the hierarchical network structure is important for high parallel execution performance.

(figures from AMD CDNA3 white paper)

4. Simulation code porting to AMD MI300A APUs

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

- In Japan, the large-scale simulations 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.

- We need to catch up with the global GPU computing trend.

- Our new supercomputer, AMD MI300A APU, will be the game changer of the simulation research in this field.

- **Transplant and Tuning of the plasma simulation codes**
  - Example : MEGA (hybrid fluid + particle simulation)

- Offload to GPU is done by OpenMP target instruction
- Some modifications are required :
  - Arrays cannot be the target of reduction
  - → reduction to temporal scalar variable
  - → or, transform the loop structure so that reduction procedure is not required

```
!$omp parallel do private(n1,n2) reduction(+:work) !2024-12-16, correction sugge
do k = lphistart, lphiend ※ k方向でreduction(縮約)
  do n = 0, lphi_n
    n1 = 2*n + 1
    n2 = 2*n + 2
    do i = 1, lrz
      do j = 1, imulti
        work(j,i,1,n1) = work(j,i,1,n1) + aaa(j,i,1,k)*cos_phi(n,k)
        work(j,i,1,n2) = work(j,i,1,n2) + aaa(j,i,1,k)*sin_phi(n,k)
```

↓

```
!$omp target teams distribute parallel do collapse(3) private(n,i,j,n1,n2,k)
do n = 0, lphi_n
  do i = 1, lrz
    do j = 1, imulti
      n1 = 2*n + 1 ※ 密なループにするためjループ内に移動
      n2 = 2*n + 2
      do k = lphistart, lphiend ※ k方向はreduction、並列化していない
        work(j,i,1,n1) = work(j,i,1,n1) + aaa(j,i,1,k)*cos_phi(n,k)
        work(j,i,1,n2) = work(j,i,1,n2) + aaa(j,i,1,k)*sin_phi(n,k)
```

- Presently tested only on 4 APUs.
- Compared with SX Aurora 4VEs, about x 2.1 speedup is achieved.
- After GPU tuning, MPI communication becomes the main bottleneck.

	Aurora 4 VE	MI300A 1/28版	MI300A 2/19版	MI300A 3/14版	MI300A 3/31版	MI300A 最新版
実行時間[s]	58.098	126.918	101.049	54.015	36.867	27.582
Aurora性能比[%]	1.00	0.46	0.57	1.08	1.58	2.10

- **Transplant and Tuning of the plasma simulation codes**
  - **Present status and known problems**
    - NEC is porting and tuning for MI300A in advance for three large simulation codes.
    - All three codes are written in Fortran90.
    - In some cases, do-loops are executed about 2 – 100 times faster on GPU than CPU in a MI300A APU only by offloading a loop with OpenMP target instruction.
    - However, in other cases, modification of the loop structure is required to compile and offload a do-loop.
    - Since there are about 15,000 stream cores, a long enough do-loop length is required to get the full performance out of the GPU (SX-Aurora VE : vector pipeline length is 256).
    - AMD's Fortran90 & OpenMP environment are immature. Some basic features have not yet been implemented.
      - ✓ For example, reduction and firstprivate instructions in OMP do-loop, mod function, array syntax, some complex functions (exp etc.), …
      - ✓ Since Fortran90 is still the major language in fusion research, we sincerely hope that AMD's Fortran compiler will be improved as soon as possible.

Thank you!

Simulation research activities in NIFS

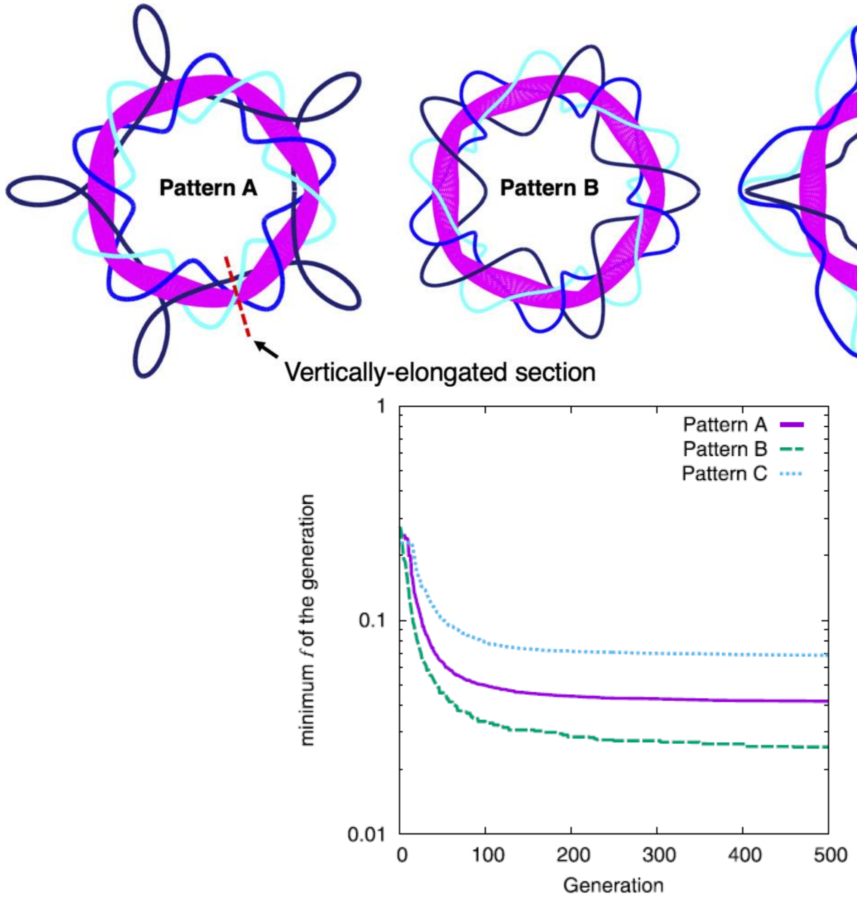
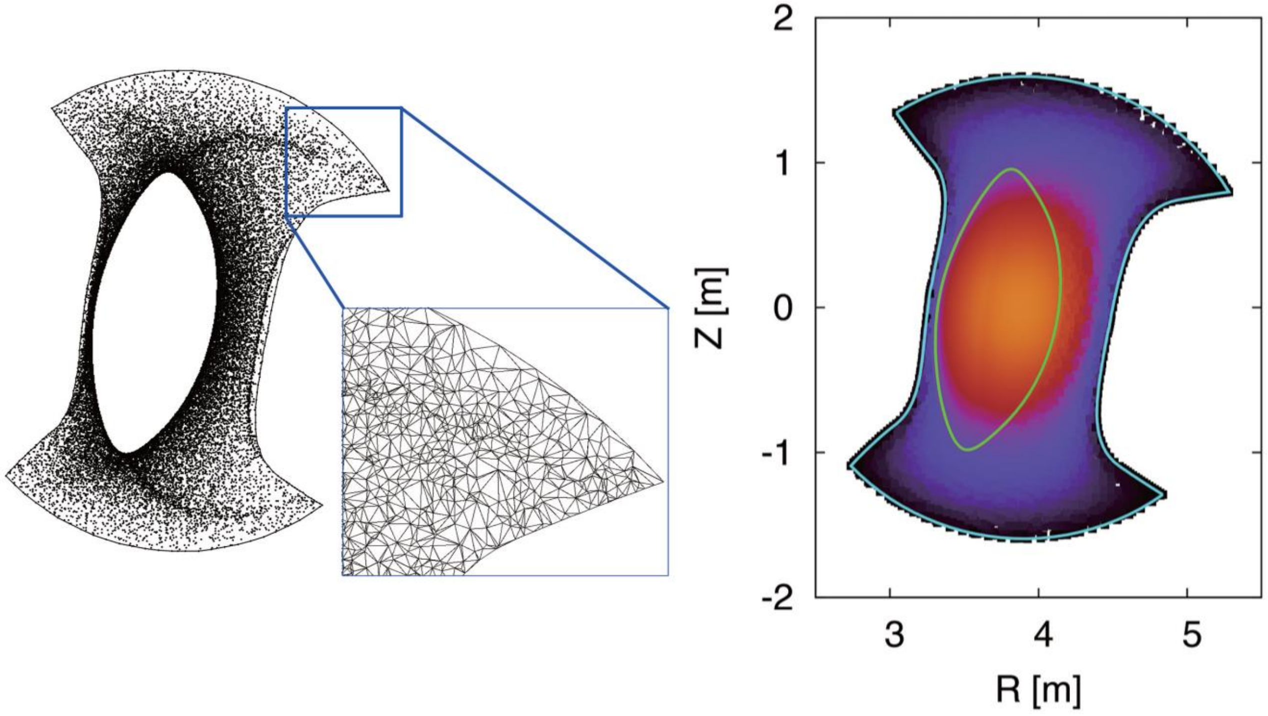
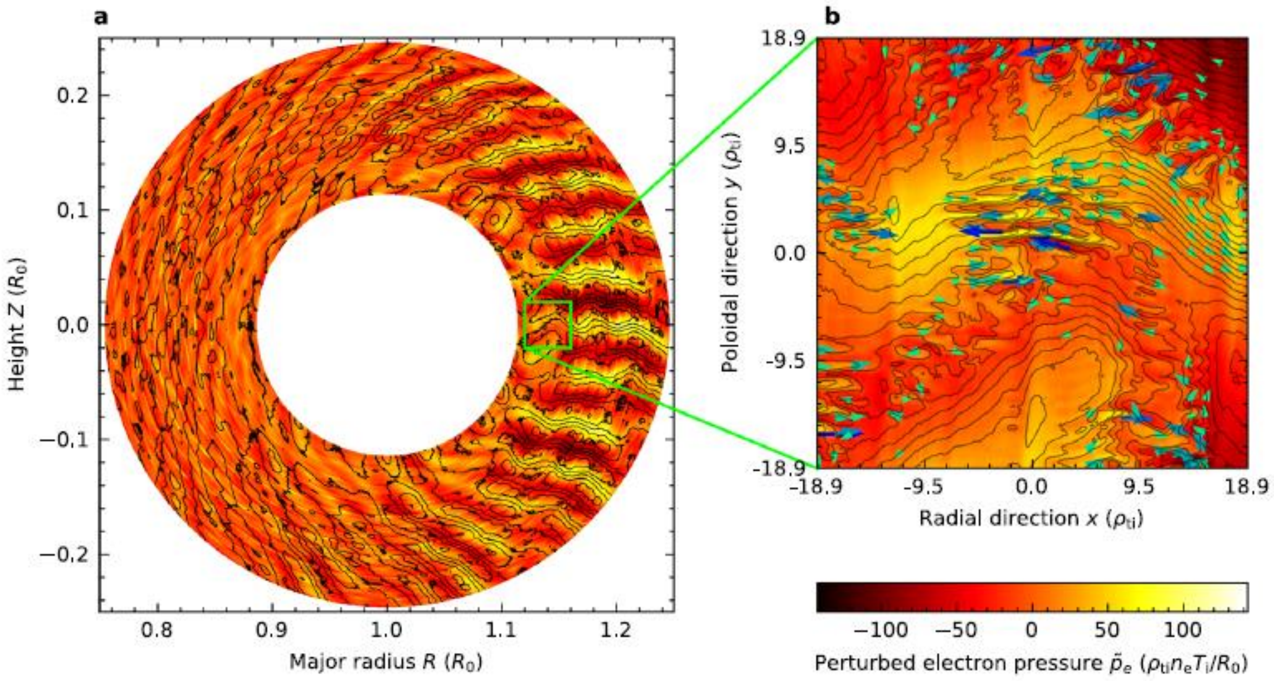


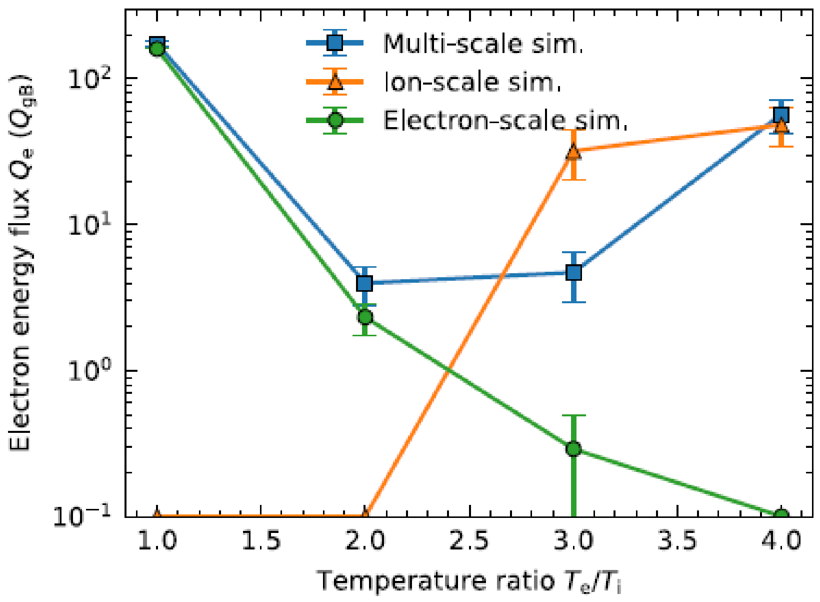
Figure 19. Convergence of f for different patterns of helical coil arrangement for W7-X.

Seeking magnet coil configurations by machine-learning (genetic algorithm) (Yamaguchi et al, Nuclear Fusion 2021)



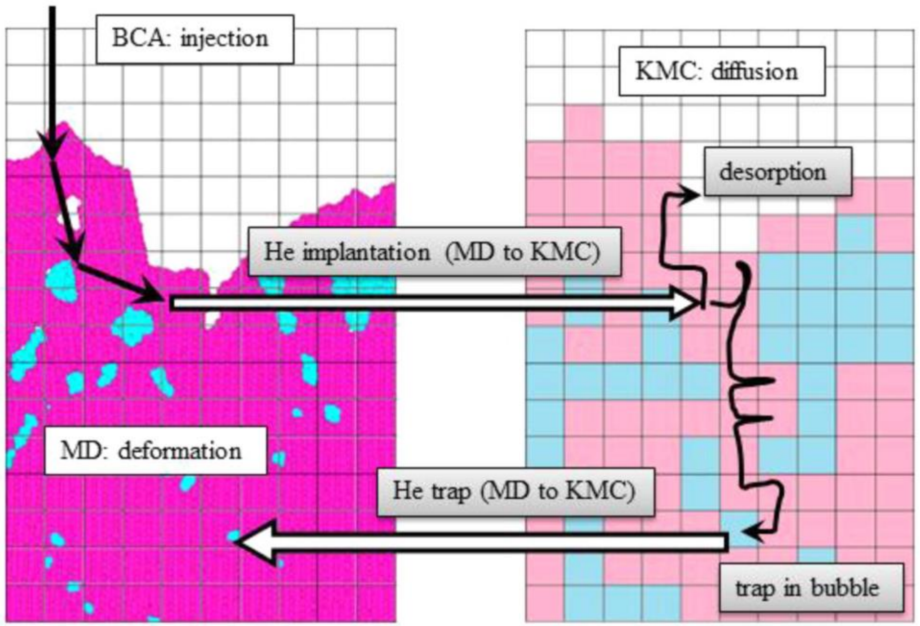
Poisson solver for complicated machine and magnetic field structure adopted in a global plasma turbulence simulation (Moritaka, MDPI Plasma 2019)

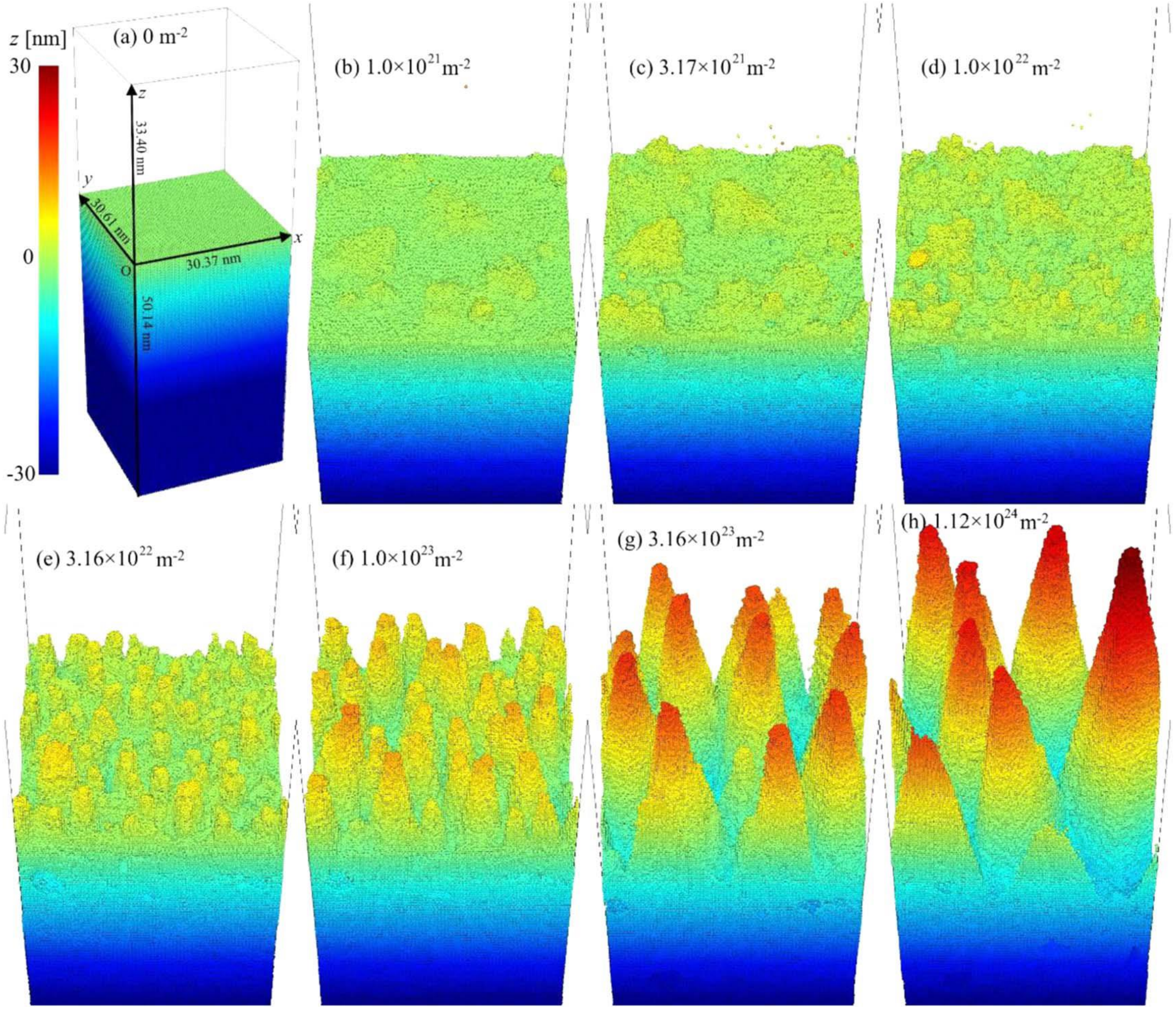




A multi-scale simulation (from electron Larmor radius – ion radius – machine size, about 100 times scale difference) of gyrokinetic plasma turbulence simulation (Maeyama, Nature comm 2022)

Cross-scale interaction of eddies affects both the electron- and ion-scale turbulent heat transport, which will be important for future burning plasma with large T_e/T_i ratio.





Simulation of plasma-wall interaction using hybrid simulation model for different physical processes in tungsten wall (A. Ito, Mater. Res. Express 2023)

BCA (Binary Collision Approximation)

KMC (Kinetic Monte Carlo)

MD (Molecular Dynamics)

The development of characteristic nanostructures on the surface of the tungsten wall, which has also been observed in experiments, is reproduced.