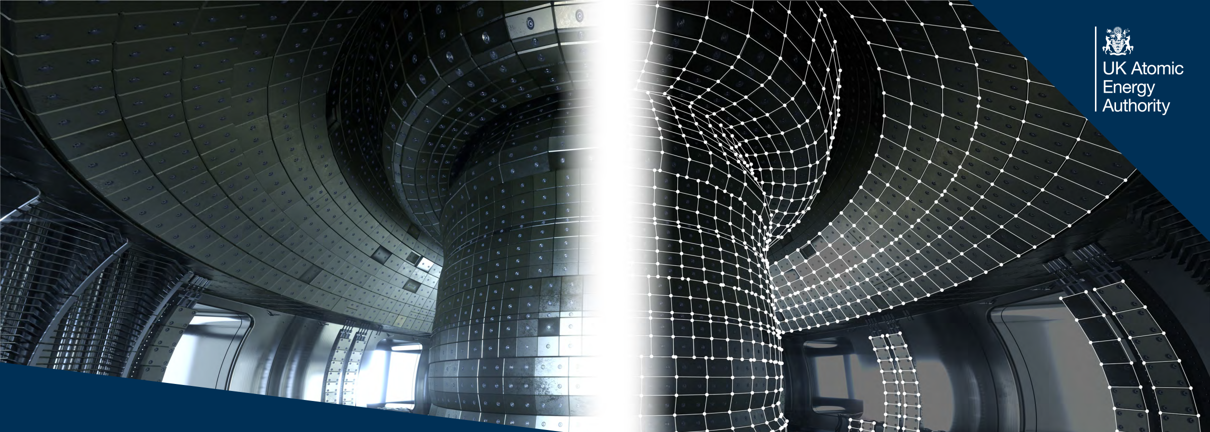


UK Atomic Energy Authority

# GPU Acceleration of OpenMC Neutronics for Fusion Applications

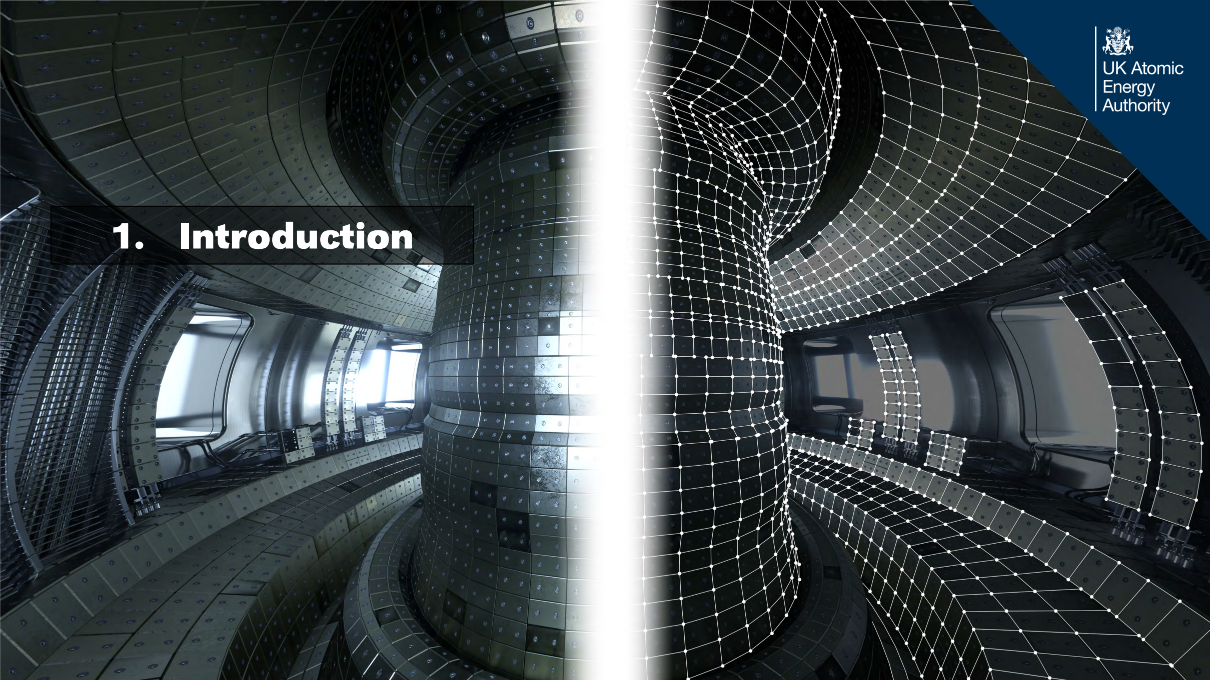
4th Fusion HPC Workshop | 30 November 2023

John Tramm[*], Helen Brooks[†], Paul Romano[*], Alex Valentine[†]
**UKAEA[†], ANL[*]**

# Overview

UK Atomic Energy Authority

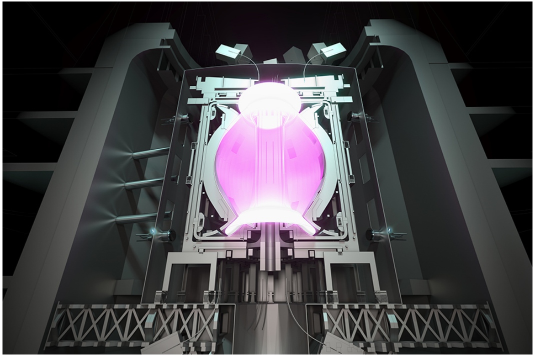
# 1. Introduction



UK Atomic Energy Authority

# Accelerating Fusion Pilot Plant Delivery

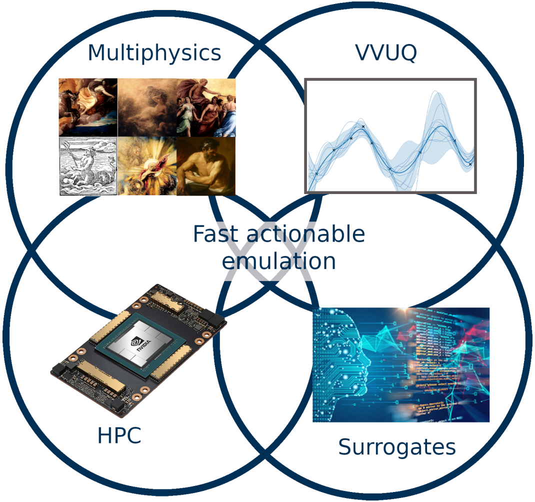
UK Atomic Energy Authority



*The UK STEP program aims to deliver a protoype spherical tokamak by 2040.*

The next two decades are projected to see emergence of first-of-a-kind fusion power-plants.

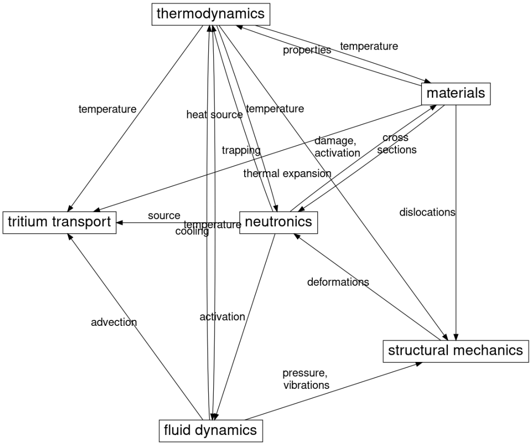
# Accelerating Fusion Pilot Plant Delivery




Delivering such ambitious programs requires **fast, actionable emulation**.

With this in place, we can aspire to automation, intelligent design and optimsation.

# Why Accelerate Monte Carlo Neutronics




We are modelling a highly complex system, with neutronics at its heart.

# Why Accelerate Monte Carlo Neutronics

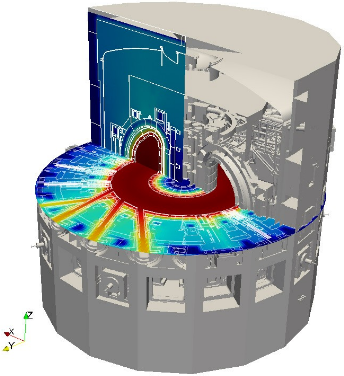



*Image credit: https://fusionforenergy.europa.eu, Juarez, R., Pedroche, G., Loughlin, M.J. et al. A full and heterogeneous model of the ITER tokamak for comprehensive nuclear analyses. Nat Energy 6, 150–157 (2021).*

Monte Carlo is widely considered the **gold standard** for modelling complex geometry.

Converging scores in hard-to-reach regions requires considerable computational resource, and may be a **bottleneck** in the context of multi-physics simulation and design.

# Why Accelerate Monte Carlo Neutronics

UK Atomic Energy Authority

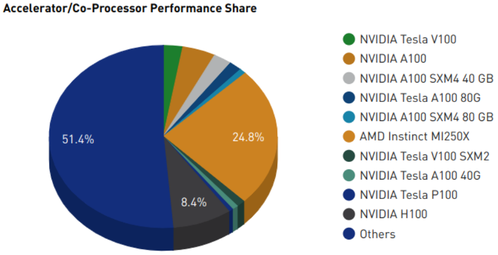



*Image credit: https://www.top500.org/statistics/list/*

GPU accelerators are now dominating the global performance share of supercomputers.

Expectation: Monte Carlo particle histories should be **embarassingly parallelisable** $\rightarrow$good candidate for GPU acceleration

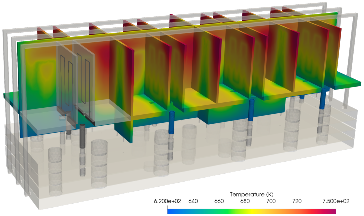
# Acceleration of OpenMC Neutronics: Mini-Review

UK Atomic Energy Authority

This talk: OpenMC [1] is code of choice.

- Open Source
- Significant documentation, training material and user support
- Available as a MOOSE-wrapped app (e.g. AURORA [2], Cardinal [3])

OpenMC




[1] github.com/openmc-dev/openmc
[2] github.com/aurora-multiphysics/aurora
[3] github.com/neams-th-coe/cardinal

# Acceleration of OpenMC Neutronics: Mini-Review

UK Atomic Energy Authority

Approaches to GPU Acceleration of OpenMC :

# Acceleration of OpenMC Neutronics: Mini-Review

UK Atomic Energy Authority

Approaches to GPU Acceleration of OpenMC :

- NVidia OptiX [4]

[4] **Exploiting Hardware-Accelerated Ray Tracing for Monte Carlo Particle Transport with OpenMC**, 2017 IEEE International Conference on Cluster Computing, *J. Salmon and S. Smith*

# Acceleration of OpenMC Neutronics: Mini-Review

UK Atomic Energy Authority

Approaches to GPU Acceleration of OpenMC :

- NVidia OptiX [4]
- CUDA [5]

[5] **Design and Optimization of GPU Capabilities in OpenMC**,Trans Am Nucl Soc, volume 125, p. 456–459 *G. Ridley and B. Forget* (2021)

# Acceleration of OpenMC Neutronics: Mini-Review



Approaches to GPU Acceleration of OpenMC :

- NVidia OptiX [4]
- CUDA [5]
- **OpenMP target offload*** [6]
  - Most mature
  - Benefit of performance portability
  - Available at github.com/exasmr/openmc
  - **Focus of this talk**

[6] **Toward Portable GPU Acceleration of the OpenMC Monte Carlo Particle Transport Code**, International Conference on Physics of Reactors (PHYSOR 2022), *J. Tramm et al*

# Why Portable Performance is Important

UK Atomic Energy Authority

Top 3 supercomputers are all GPU machines from different vendors.




## #1: Frontier

- Perf: 1.194 Exaflops/s
- CPUs: AMD EPYC
- GPUs: AMD MI250X

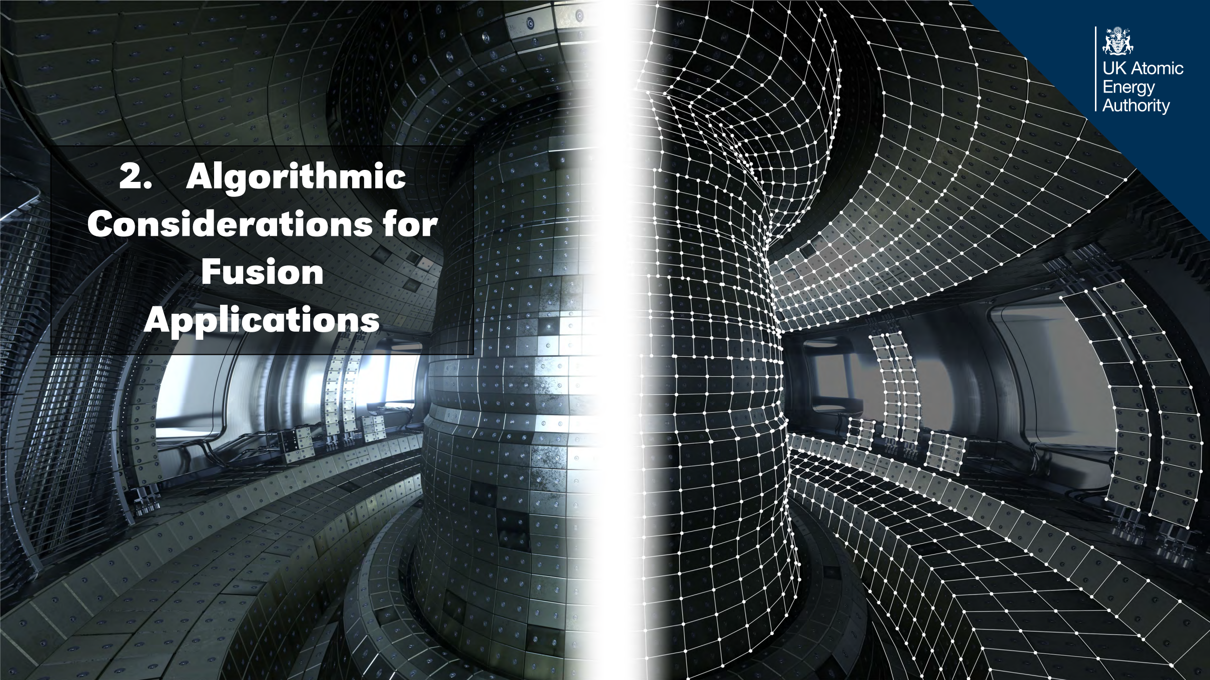
## #2: Aurora

- Perf: 0.585 Exaflops/s
- CPUs: Intel Xeon Max
- GPUs: Intel Ponte Vecchio

## #3: Eagle

- Perf: 0.561 Exaflops/s
- CPUs: Intel Xeon Platinum
- GPUs: NVidia H100

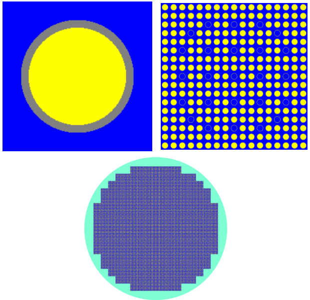
# 2. Algorithmic Considerations for Fusion Applications



UK Atomic Energy Authority

# Key Results From Previous Work

UK Atomic Energy Authority

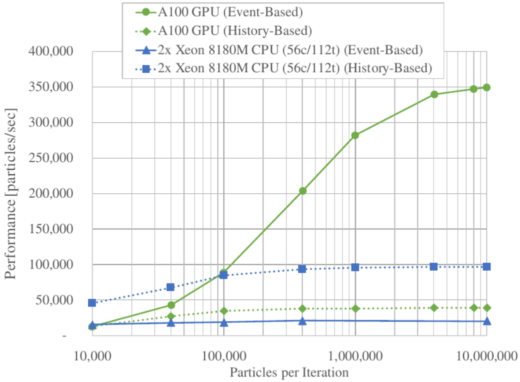


*Visualisations in the x-y plane of the Hoogenboom-Martin benchmark geometry, depicting a pin, pin-assembly, and reactor core.*

1. Prior to this contribution, results based upon **fission** applications.

# Key Results From Previous Work





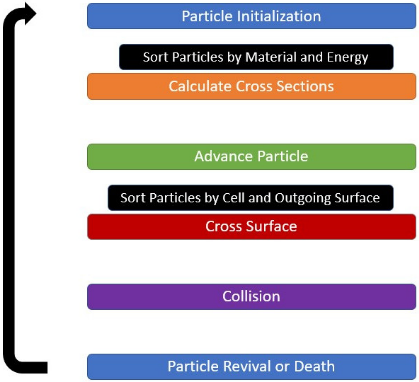

*Results presented by Tramm et al at PHYSOR 2022.*

1. Prior to this contribution, results based upon **fission** applications.
2. Critical step was to introduce **event-based** transport to attain speed-up, and **saturate GPU memory** through setting maximising particles/batch.

# Event-based Transport: A Closer Look

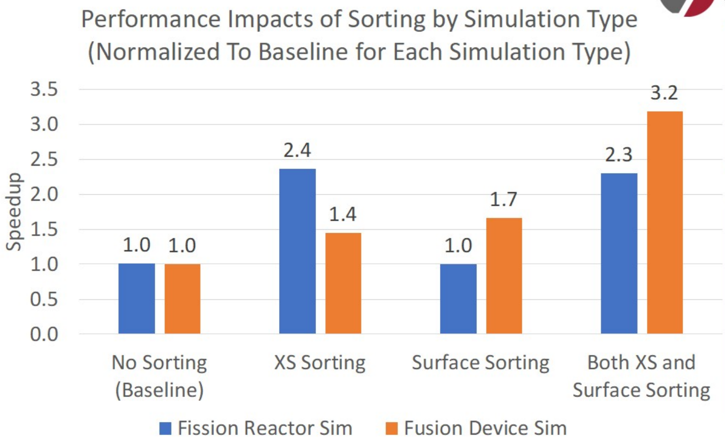
UK Atomic Energy Authority

- Traditional ***history-based*** transport: each thread processes a particle from life to death.
- ***Event-based*** transport: queue particles in buffer by event-type, then parallelize over all particles requiring that event.
- A major optimization is **sort particles** before event execution, to improve **efficiency** and **locality** of memory access.



# Comparison of Fission and Fusion Optimisations

UK Atomic Energy Authority



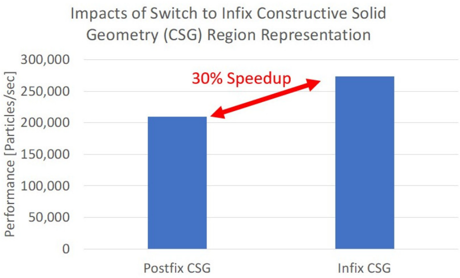

Fission simulations benefit only from XS sorting

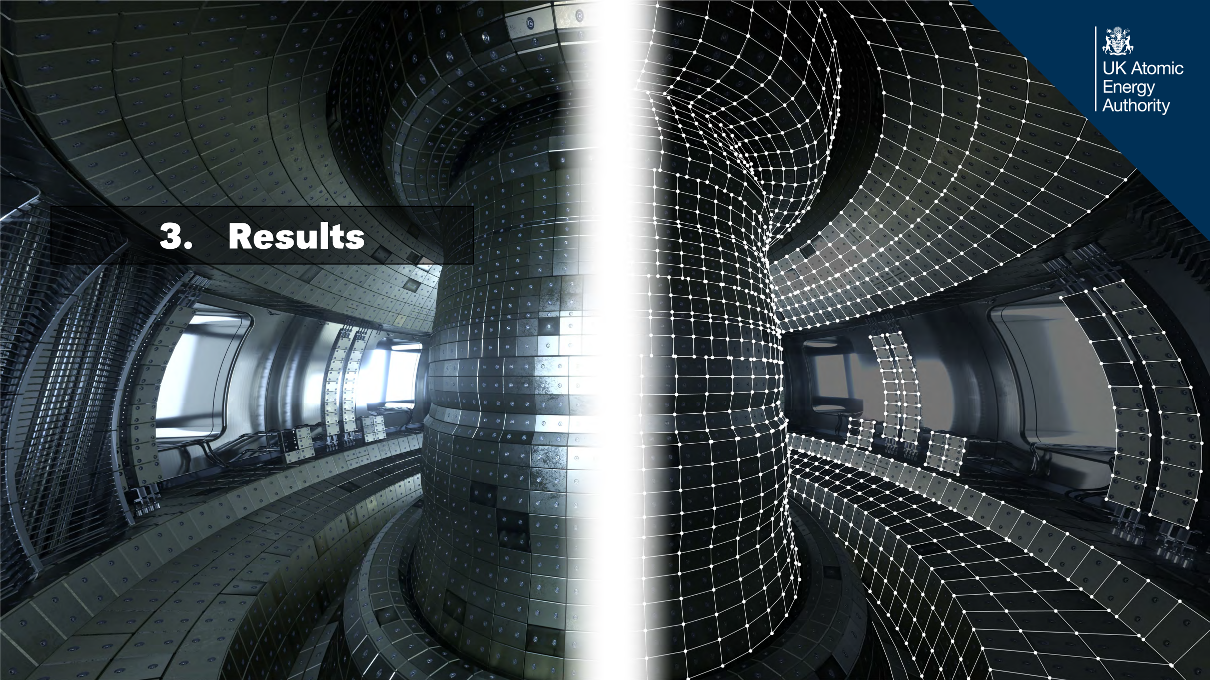
Fusion simulations benefit from both XS and surface sorting

# Additional Optimisation Needed For Fixed Source Mode

UK Atomic Energy Authority

- Evaluating **point-containment** for CSG requires querying an **implicit tree** of boolean operators.
- To improve **locality** of memory, the tree is expressed in a **flattened** structure, with several possible choices:
  - Infix: $A \cap B$ (operator in-between)
  - Postfix: $AB\cap$ (operator after)
  - Prefix: $\cap AB$ (operator before)
- Small speed-up from selecting infix notation over postfix.



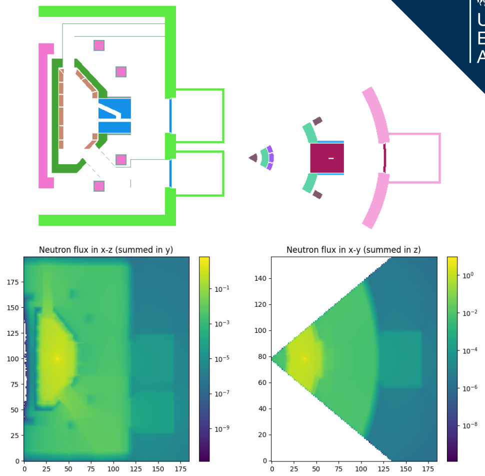


# 3. Results

UK Atomic Energy Authority

# Model Details

- Geometry: “simplified tokamak” with 168 cells / 344 surfaces / 92 nuclides
- Tallies: compute neutron flux on regular mesh with 3 million bins
- Problem size:
    - Fixed **total** number of particles: 10.24 billion
    - Fixed particles / batch / GPU: 10 million
    - Number of batches scaling inversely with number of GPU

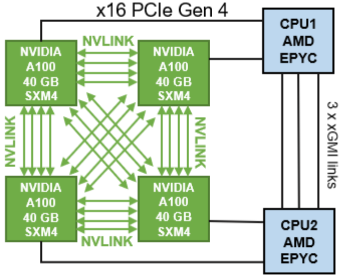



*Simplified tokamak geometry coloured by material (top), and summed flux (bottom) viewed in xz (left) / xy (right).*

# Hardware Details

UK Atomic Energy Authority

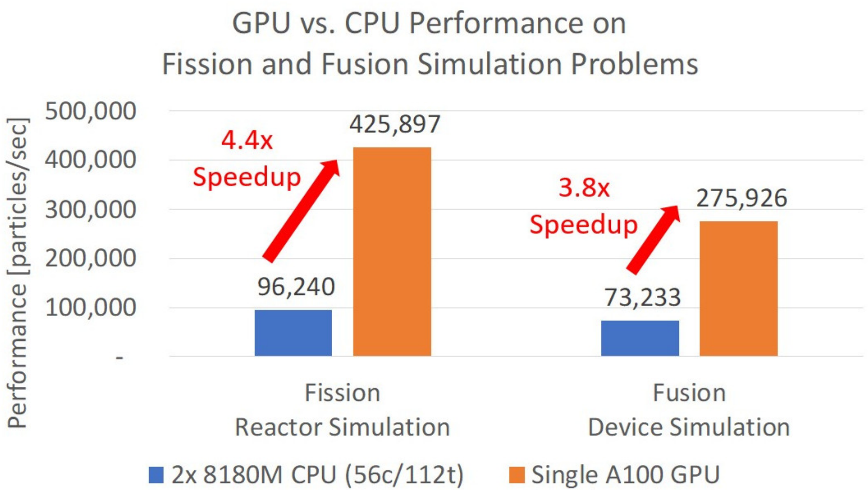
- CPU baseline: dual socket node with 2x Intel Xeon Platinum 8180M (56 cores total) for comparison with prior work.
- GPU strong scaling results obtained on Cambridge CSD3 supercomputer's Ampere partition.
- 1 Ampere node (Dell PowerEdge XE8545 server) has:
  - 2x AMD EPYC 64-core processor (128 cores total).
  - 4x NVIDIA A100 (80GB) GPUs.
  - Dual-rail Mellanox InfiniBand interconnect.



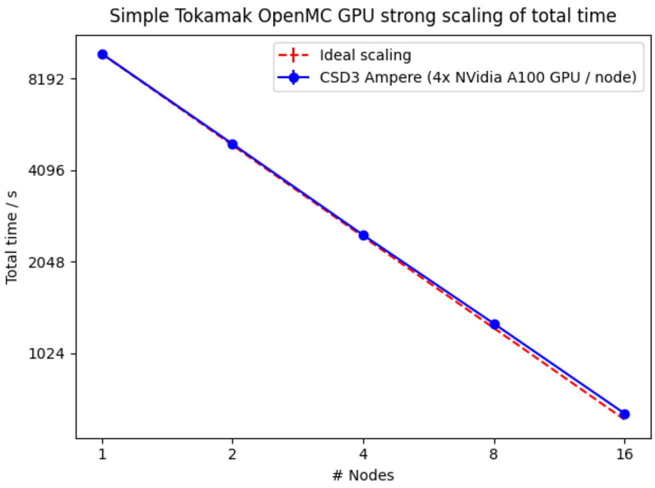

*Dell PowerEdge XE8545 CPU-GPU connectivity. Image source: https://infohub.delltechnologies.com/*

# Speed-up Results



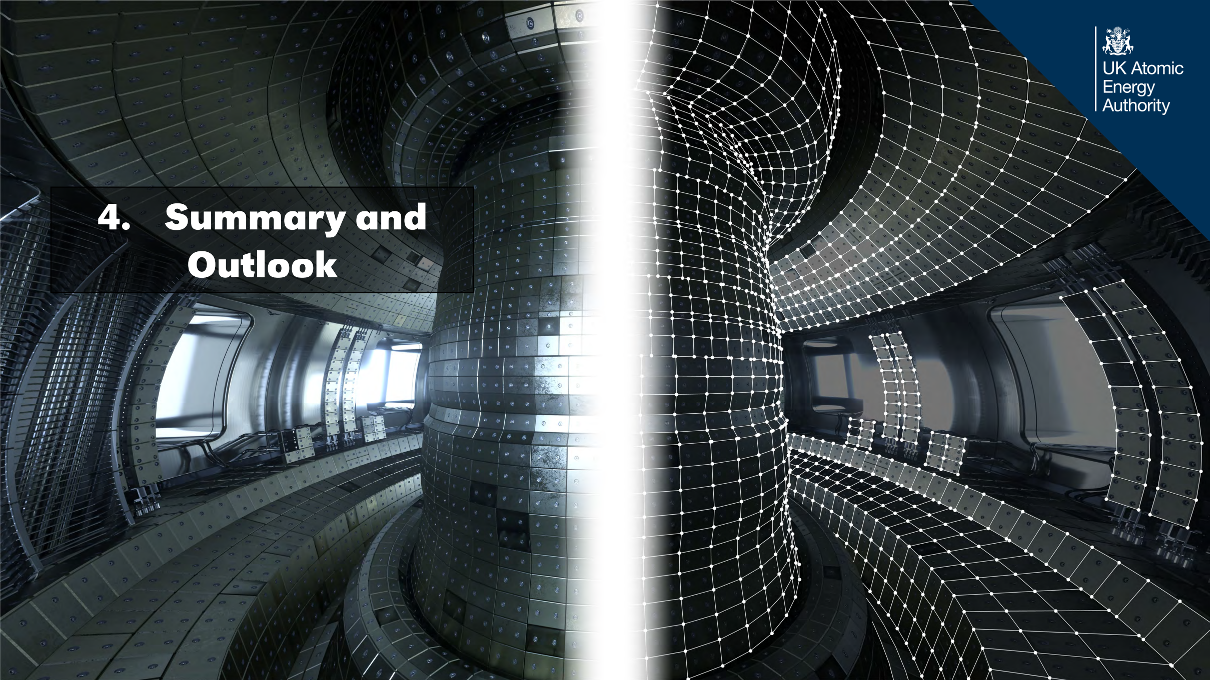


# Strong Scaling Results




95% strong scaling efficiency out to 64 GPUs.

# 4. Summary and Outlook



# Summary and Outlook

UK Atomic Energy Authority

- First demonstration of OpenMC on GPU with OpenMP target offload applied to a fusion model:
    - **3.8x speedup** for 1 Nvidia A100 GPU compared to dual-socket Intel Xeon Platinum CPU (56 cores total)
    - **95% strong scaling efficiency** attained out to 64 NVidia A100 GPU with fixed **total** particles and particles / batch / GPU.
- Next steps:
    - Apply to models of fusion-relevant devices having higher complexity.
    - Performance comparison against GPU from other vendors.
    - Assess integration into multi-physics workflows.

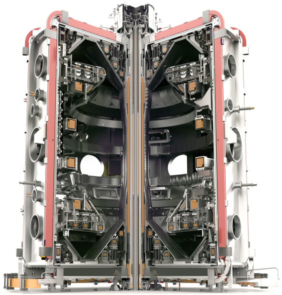


*Illustration of MAST-U tokamak.*
*Image Credit: CCFE*