
UKRI
Science and Technology Facilities Council
Scientific Computing
Welcome

Science and Technology Facilities Council

Scientific Computing

GPU Benchmarking for Performance and NetZero at IRIS

Bryce Shirley
Line Manager: Jacob Ward
Technical Contacts: Deniza Chekrygina, Sam Tygier, Juri Papay, and Cloud Operations Group.

Agenda

1 The Case for a GPU Benchmarking Suite at IRIS: Why It's Essential

2 Project Aims

3 Collecting Results

4 Benchmark Results

5 Future work

Science and Technology Facilities Council

Scientific Computing

The Case for a GPU Benchmarking Suite at IRIS: Why It's Essential

- Most existing benchmarks are not **open-source or tailored to IRIS**
- **Understand Resource Usage**
- **Evaluate GPU performance** for IRIS workloads
- **To track CO2 emissions**
- **Informed Purchasing**

Project Aims

- **GPU benchmarking suite** tailored to IRIS workloads.
- **To assess different GPU performance and CO2 emissions**.

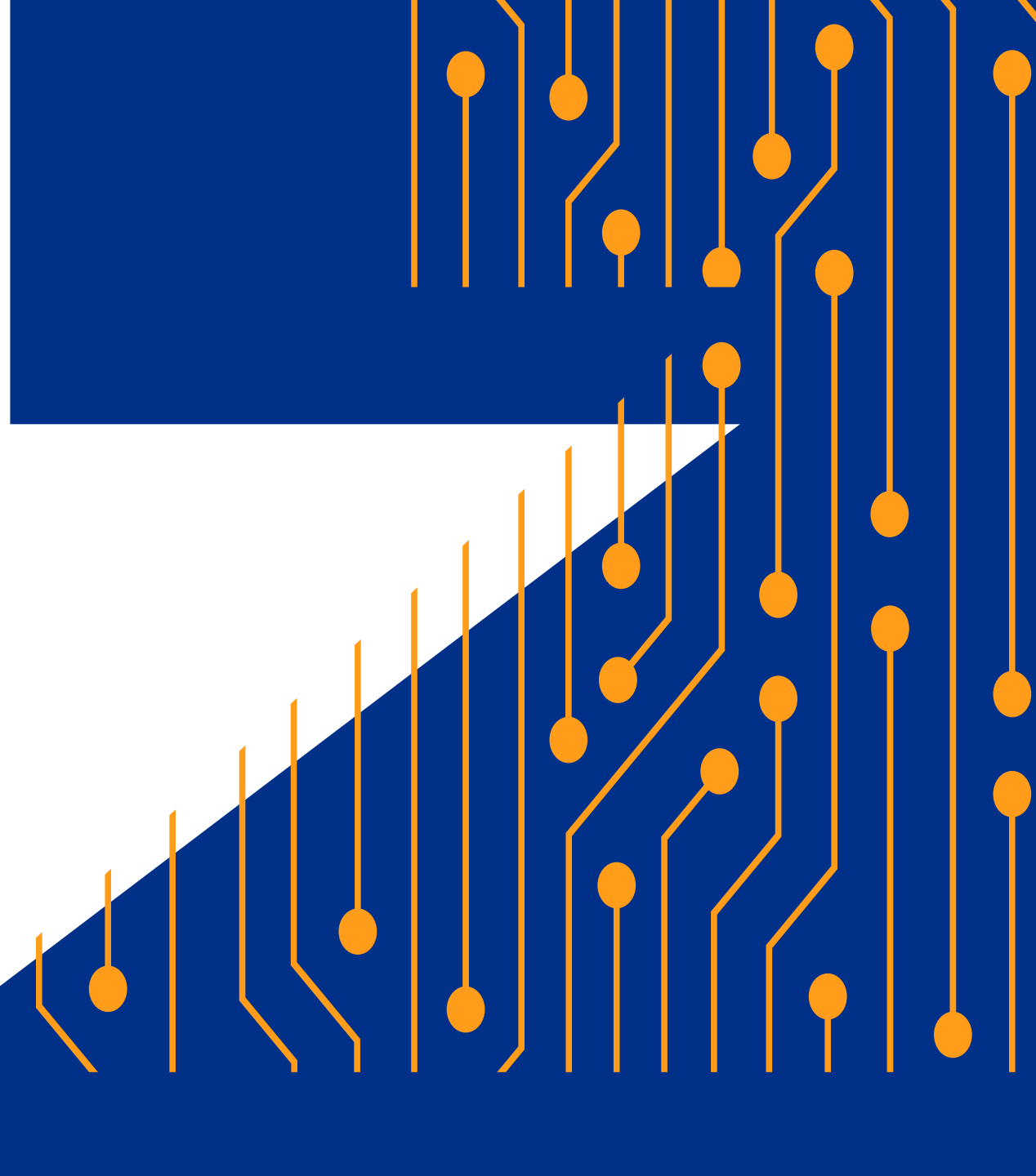
Accomplishments And Progress

- **Prototype GPU Benchmarking Suite** with integrated Mantid, SciML benchmarks.
- **Performance and CO2 Emission Tracking**: The suite assesses GPU performance (nvidia-smi) and tracks CO2 emissions (NESO API).
- **Grafana Dashboard**: To visualize and analysising benchmark performance data.

GitHub Repo and Documentation

IRIS Bench Repository: Access the benchmark suite.

Documentation: Includes installation, Docker files/Habor, command-line interface, and more.



Science and Technology Facilities Council

Scientific Computing

Collecting Results

Science and Technology Facilities Council

Scientific Computing

Running a Benchmark on IRIS Bench

GPU and Carbon Performance Results

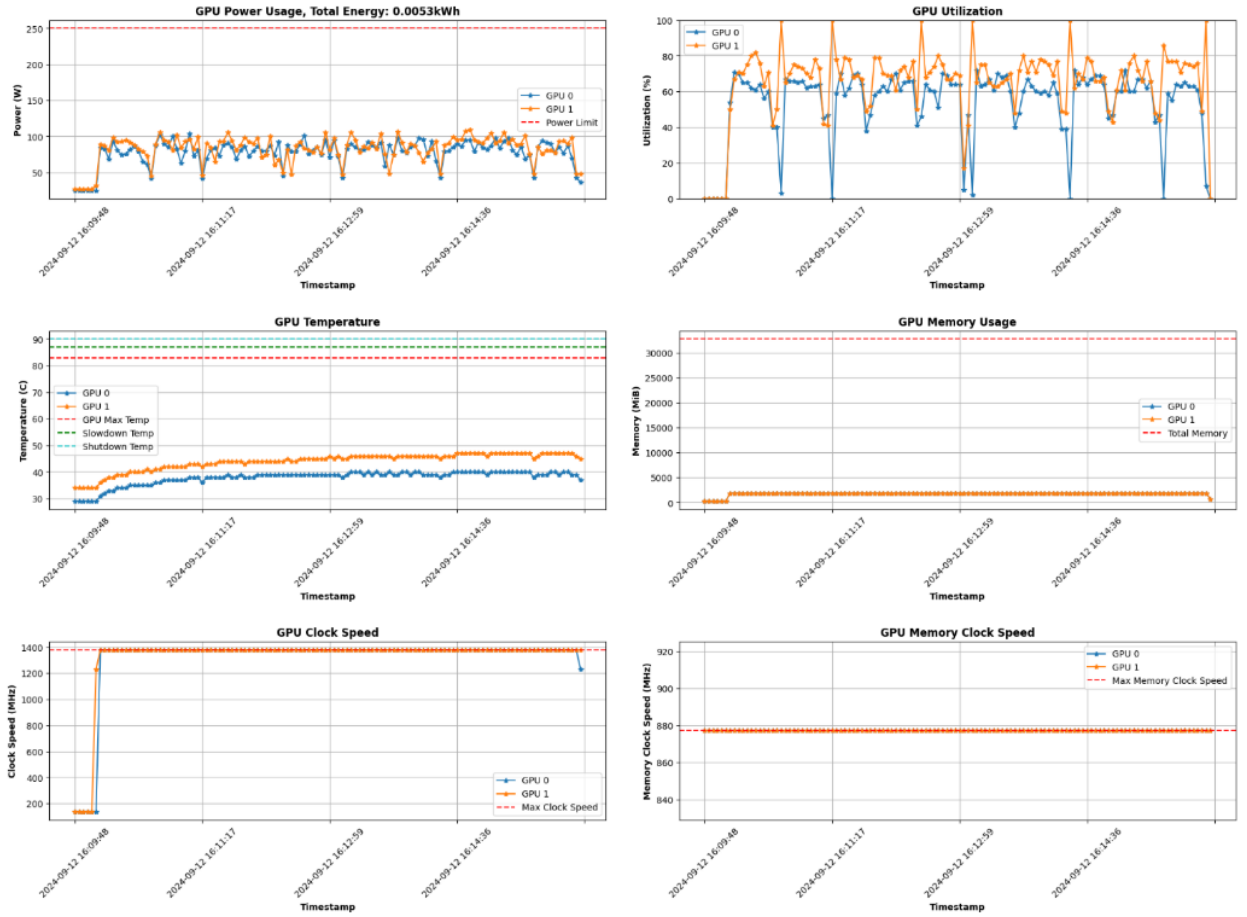
Metric	Value
Benchmark:	stemdl_classification
Benchmark Score (s)	1053.29596
Elapsed Monitor Time (s)	1202.78311
Total GPU Energy Consumed (kWh)	0.00561
Total GPU Carbon Emissions (gCO2)	0.54695

Carbon Information

Metric	Value
Average Carbon Forecast (gCO2/kWh)	97.5
Carbon Forecast Start Time	2024-09-11 14:25:55
Carbon Forecast End Time	2024-09-11 14:45:58

GPU Information

Metric	Value
GPU Type	Tesla V100-PCIE-32GB
No. of GPUs	2
Average GPU Utilization (for >0.00% GPU Util.) (%)	43.03604
Average GPU Power (for >0.00% GPU Util.) (W)	65.31081 (Power Limit: 250)
Average GPU Temperature (for >0.00% GPU Util.) (°C)	38.58559
Temperature Threshold - Slowdown (°C)	87.00
Average GPU Memory (for >0.00% GPU Util.) (MiB)	1877.58896 (Total Memory: 32768.0)
Average Clock Speed (MHz)	1340.87 (Max: 1380.00)
Average Memory Clock Speed (MHz)	877.00 (Max: 877.00)



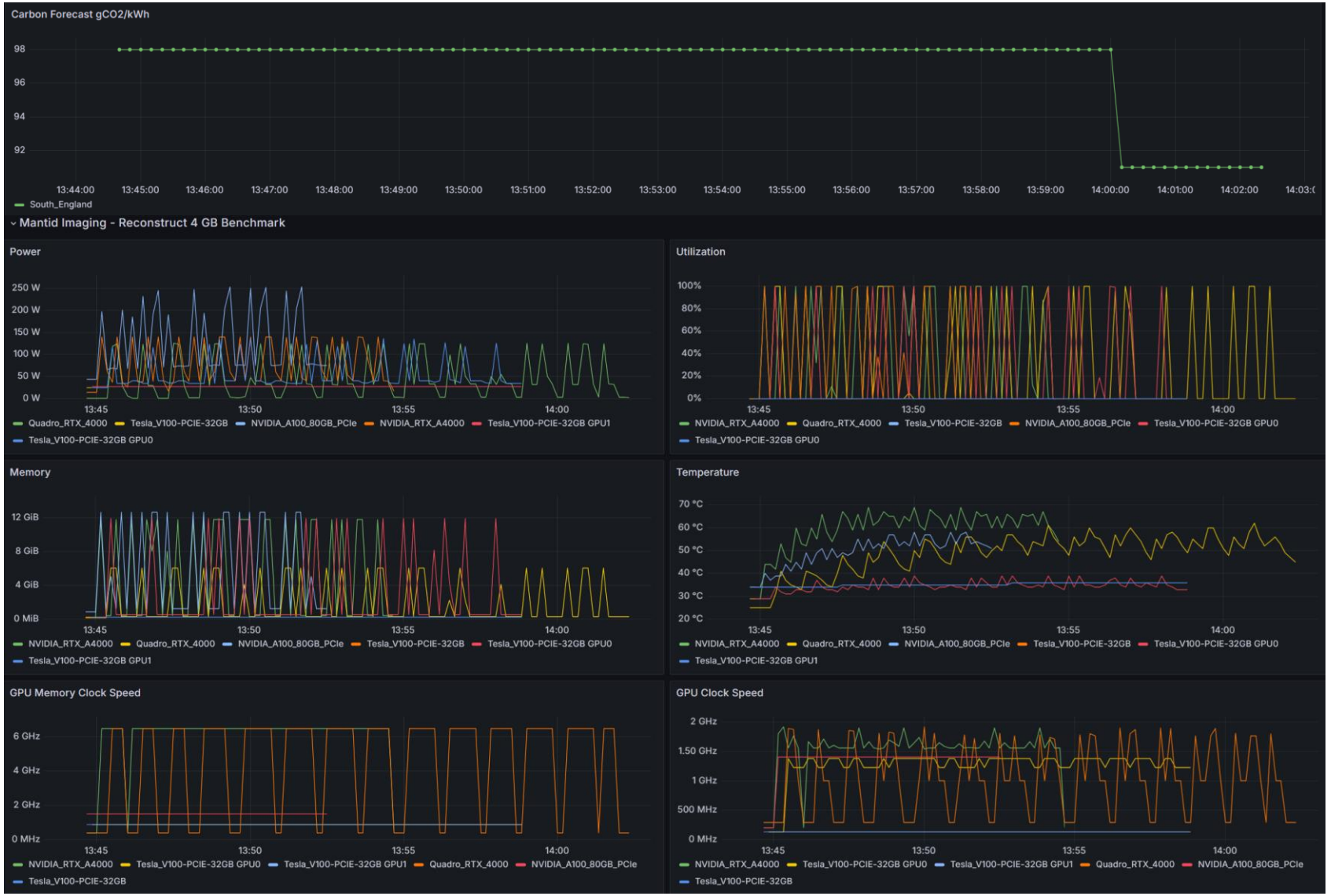
Science and Technology Facilities Council

Scientific Computing

Grafana Dashboard for Performance Analysis and Comparison

Benchmark Results

Time	GPU Type	Number of GPUs	Benchmark Score
2024-09-11 14:02:30	Quadro_RTX_4000	1	1037.06005
2024-09-11 13:45:00	Tesla_V100-PCIE-32GB	1	
2024-09-11 13:59:00	Tesla_V100-PCIE-32GB	2	812.68094
2024-09-11 13:52:30	NVIDIA_A100_80GB_PCIe	1	444.46629
2024-09-11 13:54:50	NVIDIA_RTX_A4000	1	576.91544



Science and Technology Facilities Council

Scientific Computing

Science and Technology Facilities Council

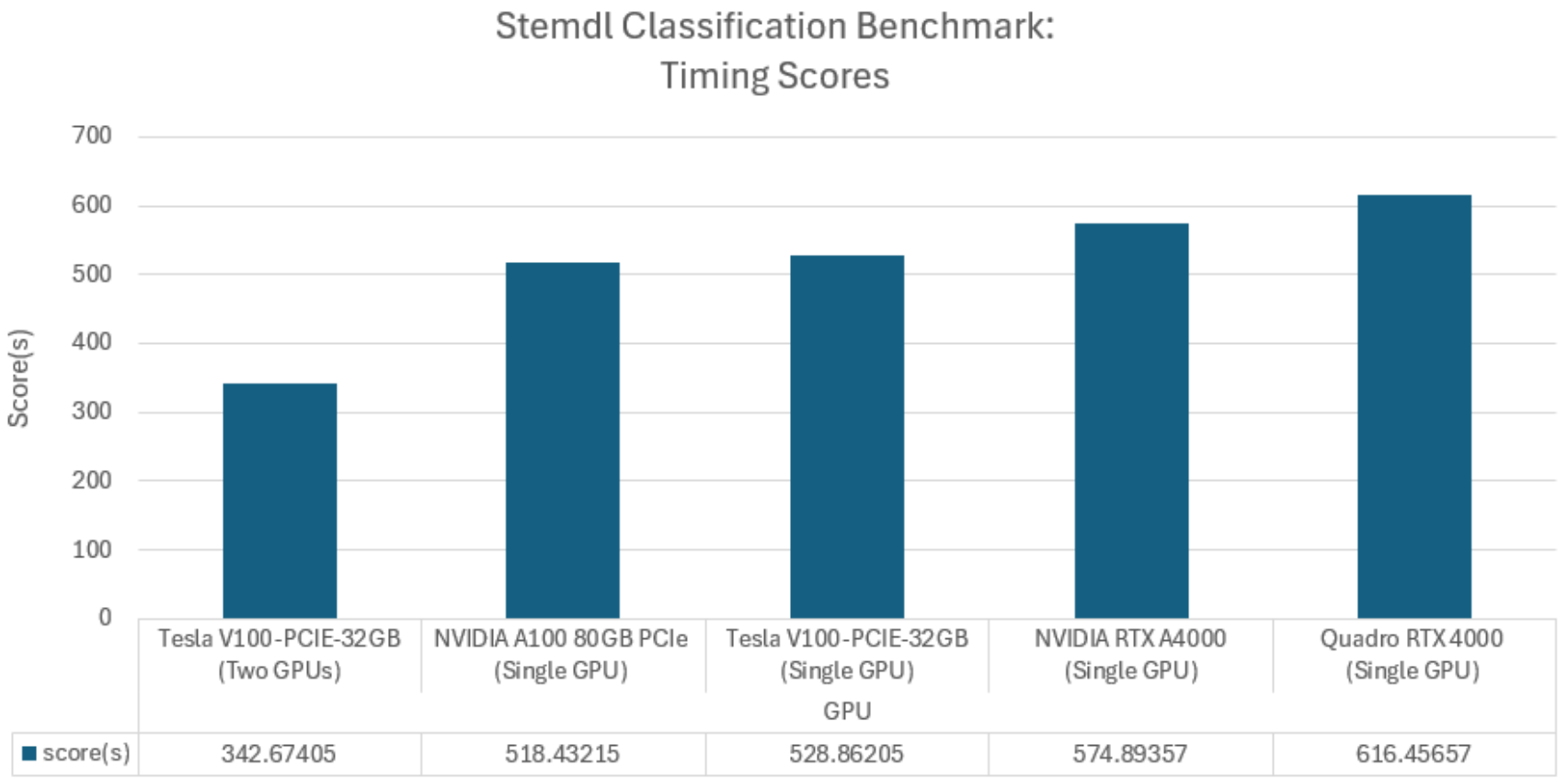
Scientific Computing

SciML Bench and MANTID Benchmark Results Examples

Science and Technology Facilities Council

Scientific Computing

SciML Bench - StemdlClassification

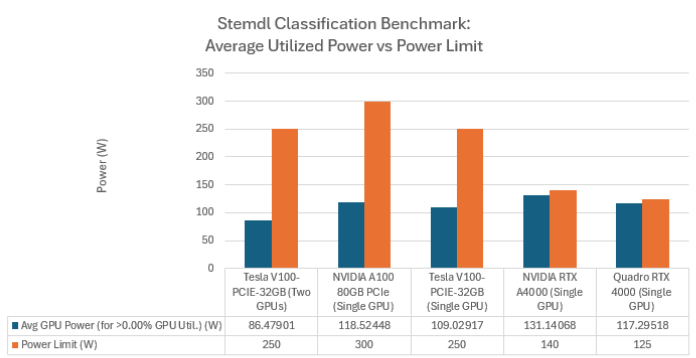


[SciML Benchmark Repo]

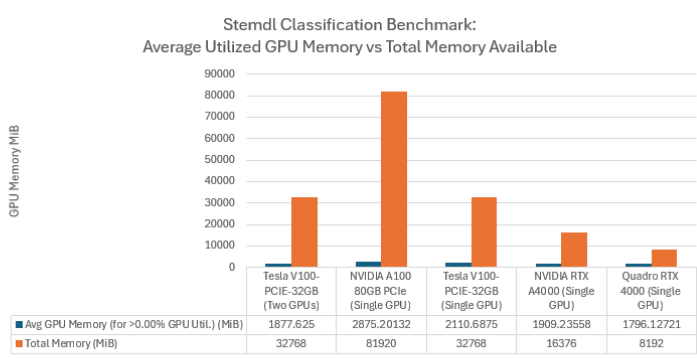
Science and Technology Facilities Council
Scientific Computing

SciML Benchmarks

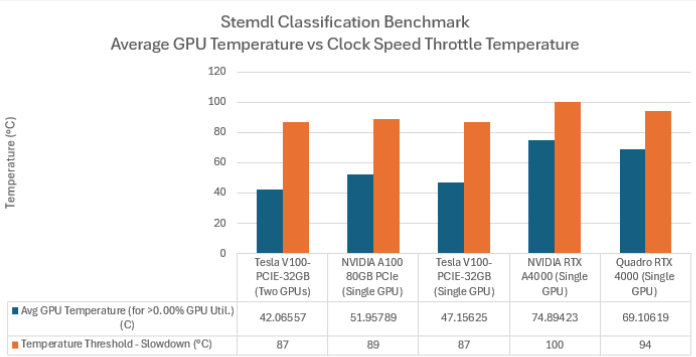
SciML Bench - StemdlClassification



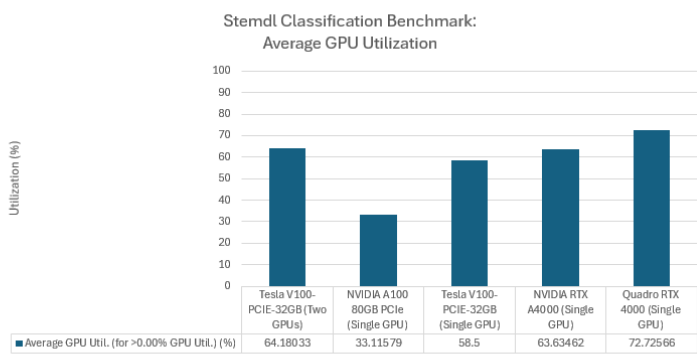
	Tesla V100-PCIE-32GB (Two GPUs)	NVIDIA A100 80GB PCIe (Single GPU)	Tesla V100-PCIE-32GB (Single GPU)	NVIDIA RTX A4000 (Single GPU)	Quadro RTX 4000 (Single GPU)
Avg GPU Power (for >0.00% GPU Util.) (W)	86.47901	118.52448	109.02917	131.14068	117.29518
Power Limit (W)	250	300	250	140	125



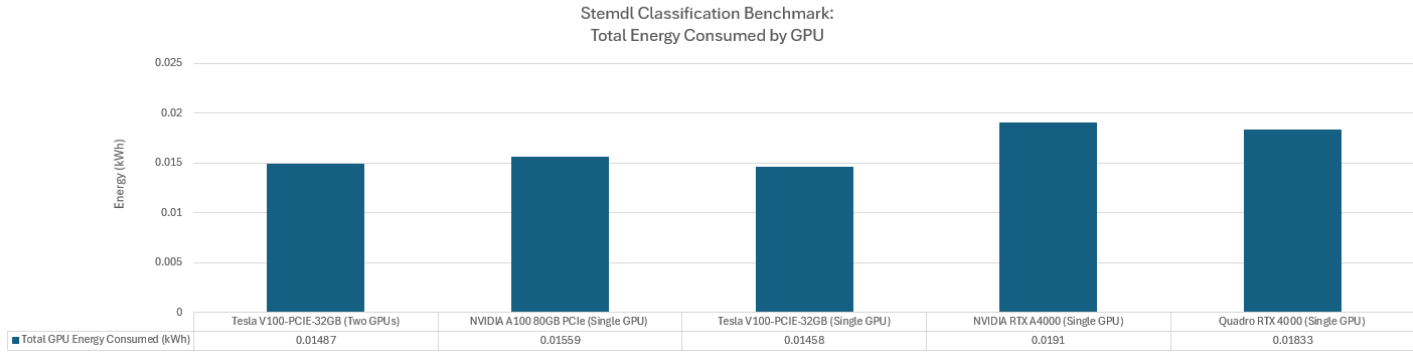
	Tesla V100-PCIE-32GB (Two GPUs)	NVIDIA A100 80GB PCIe (Single GPU)	Tesla V100-PCIE-32GB (Single GPU)	NVIDIA RTX A4000 (Single GPU)	Quadro RTX 4000 (Single GPU)
Avg GPU Memory (for >0.00% GPU Util.) (MiB)	1877.625	2875.20132	2110.6875	1909.23558	1796.12721
Total Memory (MiB)	32768	81920	32768	16376	8192



	Tesla V100-PCIE-32GB (Two GPUs)	NVIDIA A100 80GB PCIe (Single GPU)	Tesla V100-PCIE-32GB (Single GPU)	NVIDIA RTX A4000 (Single GPU)	Quadro RTX 4000 (Single GPU)
Avg GPU Temperature (for >0.00% GPU Util.) (C)	42.06557	51.95789	47.15625	74.89423	69.10619
Temperature Threshold - Slowdown (°C)	87	89	87	100	94



	Tesla V100-PCIE-32GB (Two GPUs)	NVIDIA A100 80GB PCIe (Single GPU)	Tesla V100-PCIE-32GB (Single GPU)	NVIDIA RTX A4000 (Single GPU)	Quadro RTX 4000 (Single GPU)
Average GPU Util. (for >0.00% GPU Util.) (%)	64.18033	33.11579	58.5	63.63462	72.72566



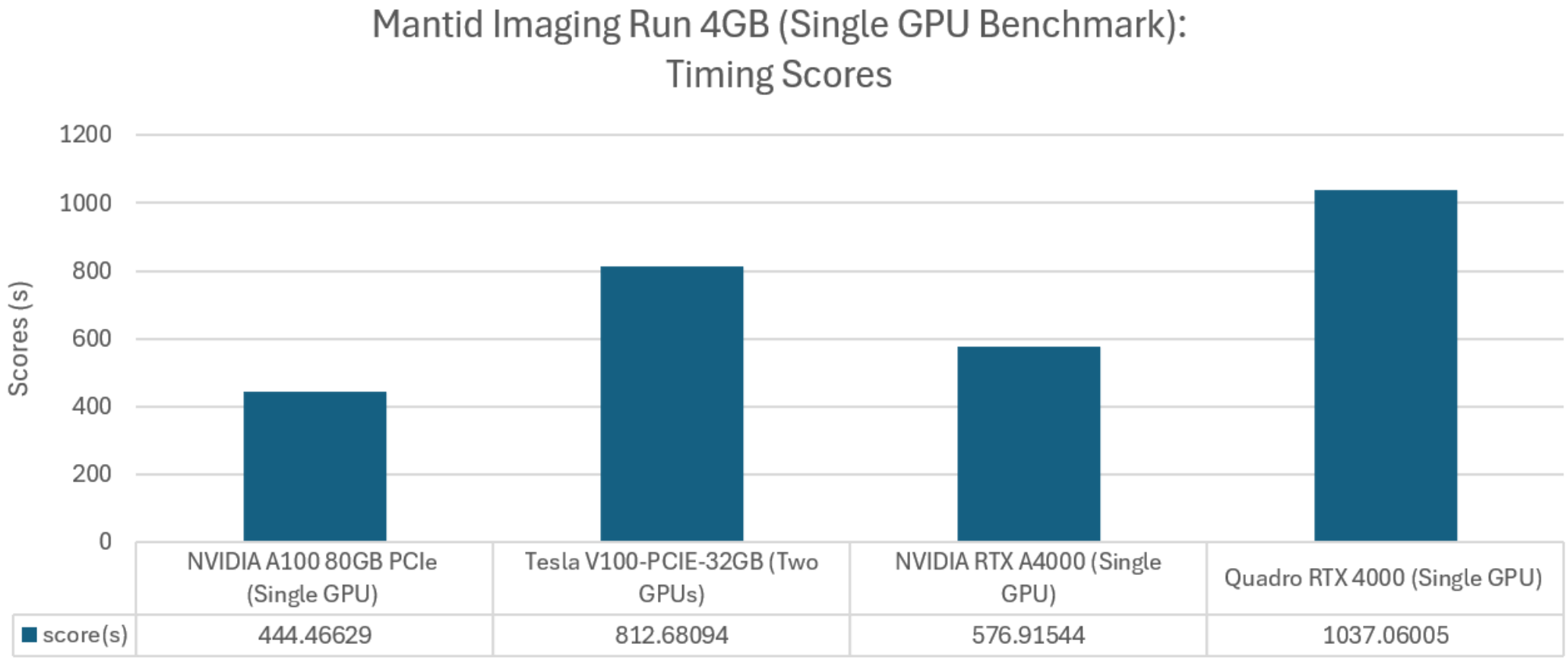
	Tesla V100-PCIE-32GB (Two GPUs)	NVIDIA A100 80GB PCIe (Single GPU)	Tesla V100-PCIE-32GB (Single GPU)	NVIDIA RTX A4000 (Single GPU)	Quadro RTX 4000 (Single GPU)
Total GPU Energy Consumed (kWh)	0.01487	0.01559	0.01458	0.0191	0.01833

Science and Technology Facilities Council
Scientific Computing

SciML Benchmark Repo

SciML Benchmarks

MANTID Imaging Benchmark 4GB Dataset

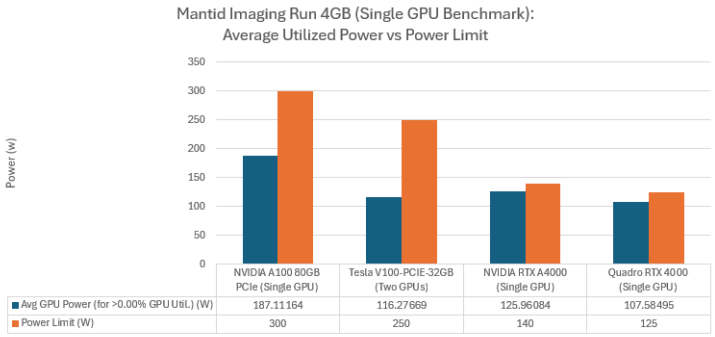


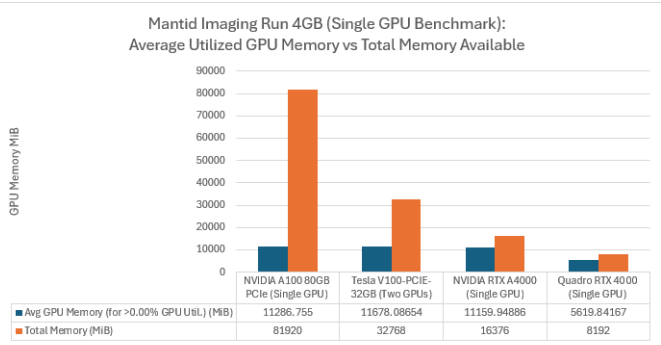
Science and Technology Facilities Council

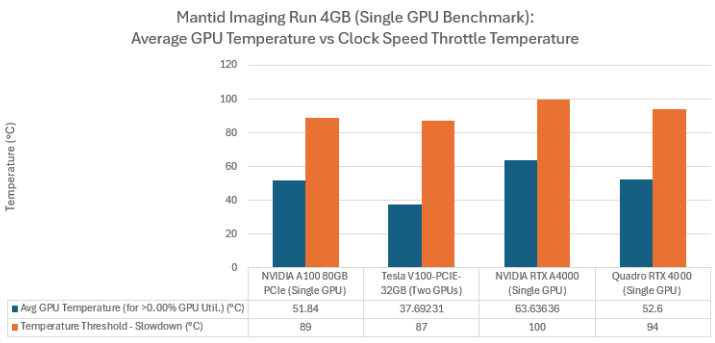
Scientific Computing

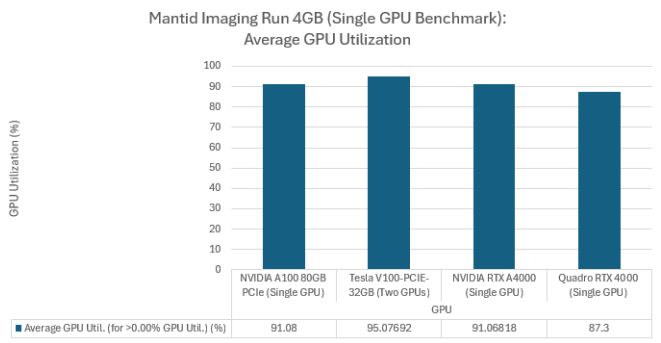
[MANTID Imaging Benchmarks Repo]

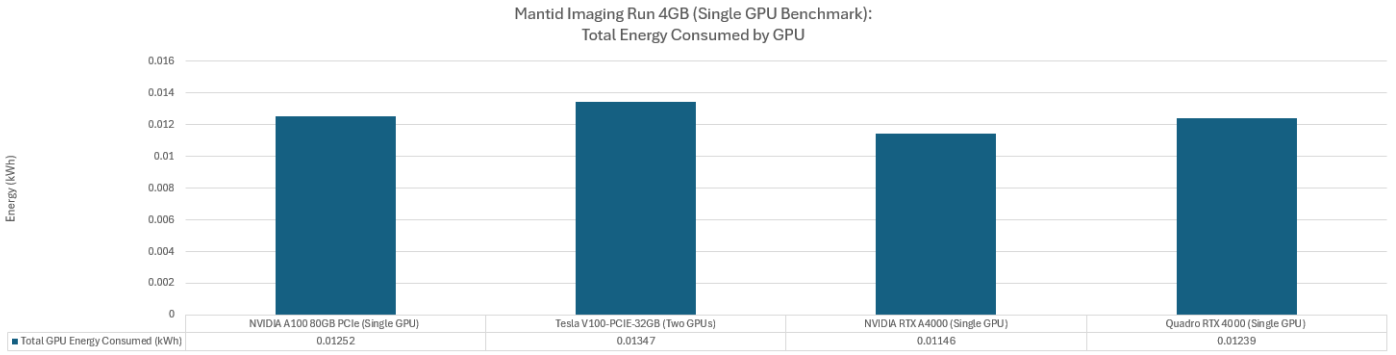
MANTID Imaging Benchmark 4GB Dataset











Science and Technology Facilities Council

Scientific Computing

[MANTID Imaging Benchmarks Repo]

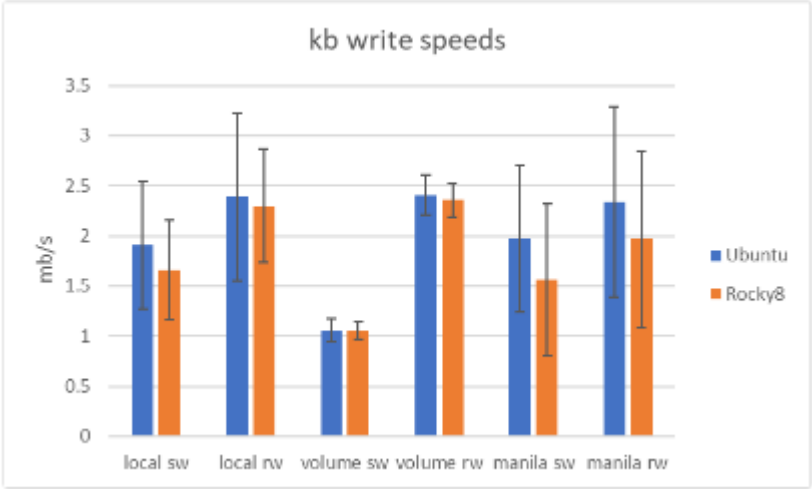
Scientific Computing

Future Work: Improvements to the GPU Benchmarking Suite

Scientific Computing

Improvements to the GPU Benchmarking Suite

- **Integrating with Meerkat**
- **Work with IRIS users to integrate more of their benchmarks**
- **Overall Score (Normalize)**
- **FLOP Estimates (+ Efficiency scores)**
- **Integrate CPU, Storage and Memory Benchmarks**
- **Carbon and Cost Forecasting (Next Section)**



Meerkat Storage Benchmark by Chris Green from Cloud Operations Group

Science and Technology Facilities Council

Scientific Computing

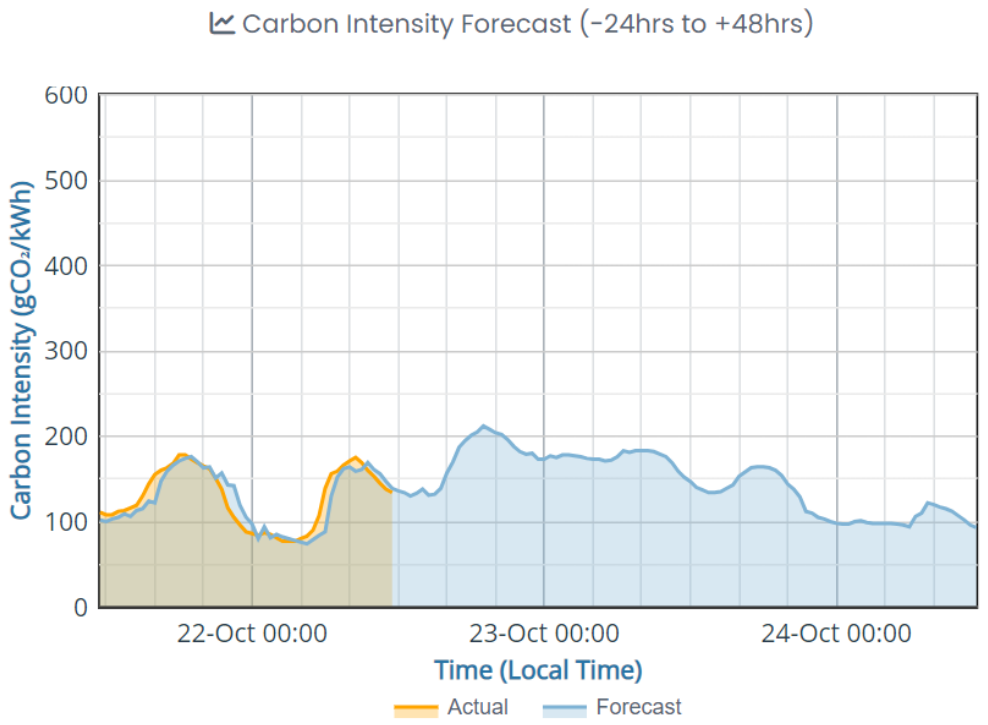
Future Work: Optimizing Energy-Intensive Tasks

Science and Technology Facilities Council

Scientific Computing

Carbon Emission per kWh Forecasting

- **NESO API**
- **Regional Optimization**: Select times with lower carbon intensity
- **Regional Scheduling**: Select regions with lower carbon intensity (e.g., Daresbury vs. RAL).



Carbon Intensity Forecast from NESO API

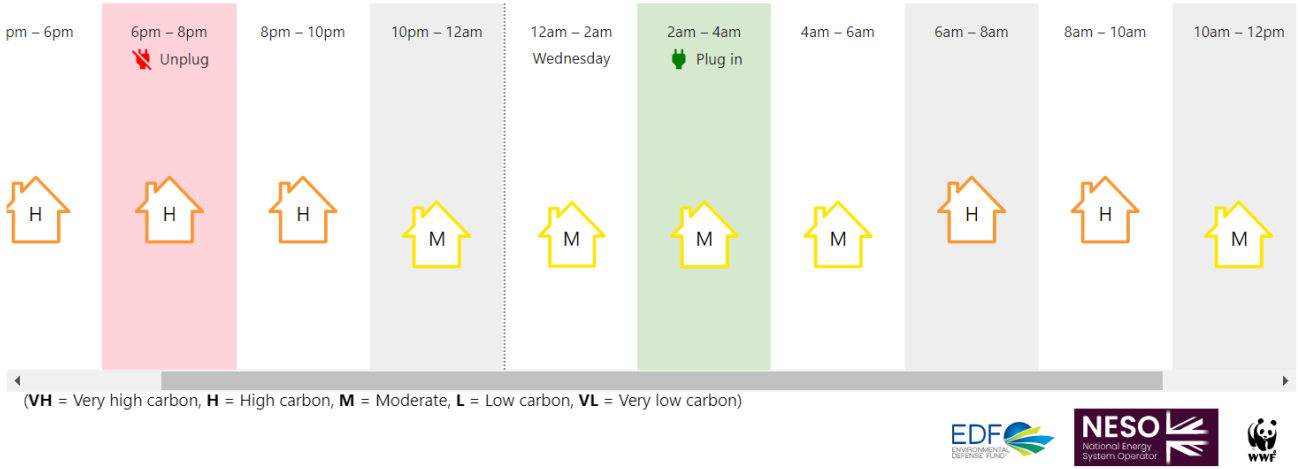
Science and Technology Facilities Council
Scientific Computing

Financial Cost Forecasting

- **Imperial College London API**: Forecast energy costs (pence/kWh)
- Use forecasts to schedule tasks during low-cost periods.

Combining Carbon and Cost Forecasting with IRIS Bench UI

- Displays **runtime with carbon emissions, costs forecasts** for various times and regions (if possible)
- **GPU options with performance** for similar workloads
- (option for **automated scheduling)**



Science and Technology Facilities Council

Scientific Computing

Additional Carbon, Financial / Ethical Costs (Beyond Power)

- **Fixed Costs**: Includes initial purchase, maintenance, and infrastructure costs, depreciation value
- **Idle Costs**: Value lost when GPUs remain unused but still incur depreciation and infrastructure costs
- **Loss of potential research opportunity** (Idle Cost).
- **Embodied Carbon**: Emissions from GPU manufacturing and delivery.

$$\left(\frac{\text{time used}}{\text{lifetime}} \times \text{total fixed cost of running}\right) - \left(\frac{\text{time idle}}{\text{lifetime}} \times \text{cost of idle time}\right)$$

Want More Information?

- Explore the repo and documentation for detailed guidance.
- Review the IRIS Bench Report for more insights into project research, accuracy considerations, carbon and benchmarking considerations.

Science and Technology Facilities Council

Scientific Computing

Questions?

Science and Technology Facilities Council

Scientific Computing

Thank you

scd.stfc.ac.uk

@SciComp_STFC