Energy-efficiency on the NVIDIA A100
from lattice QCD to large language models

Antonin Portelli — 03/08/2023

Lattice 2023, FermiLab, IL, USA

THE UNIVERSITY *of* EDINBURGH

Science and Technology Facilities Council

DiRAC

Goals

- Understand the energy **footprint of HPC calculations**
- Understand how to improve energy efficiency to help **reaching net zero computing targets**
- Understand how to mitigate the **impact of surging energy prices on scientific outputs**
- **Bottom-up approach**: start from **domain-specific studies**. Energy-efficiency is **domain-dependent**

Summer 2022 DiRAC study

Report and data

- Report commissioned by UK STFC DiRAC

  https://doi.org/10.5281/zenodo.7057318
- Report data and running environment

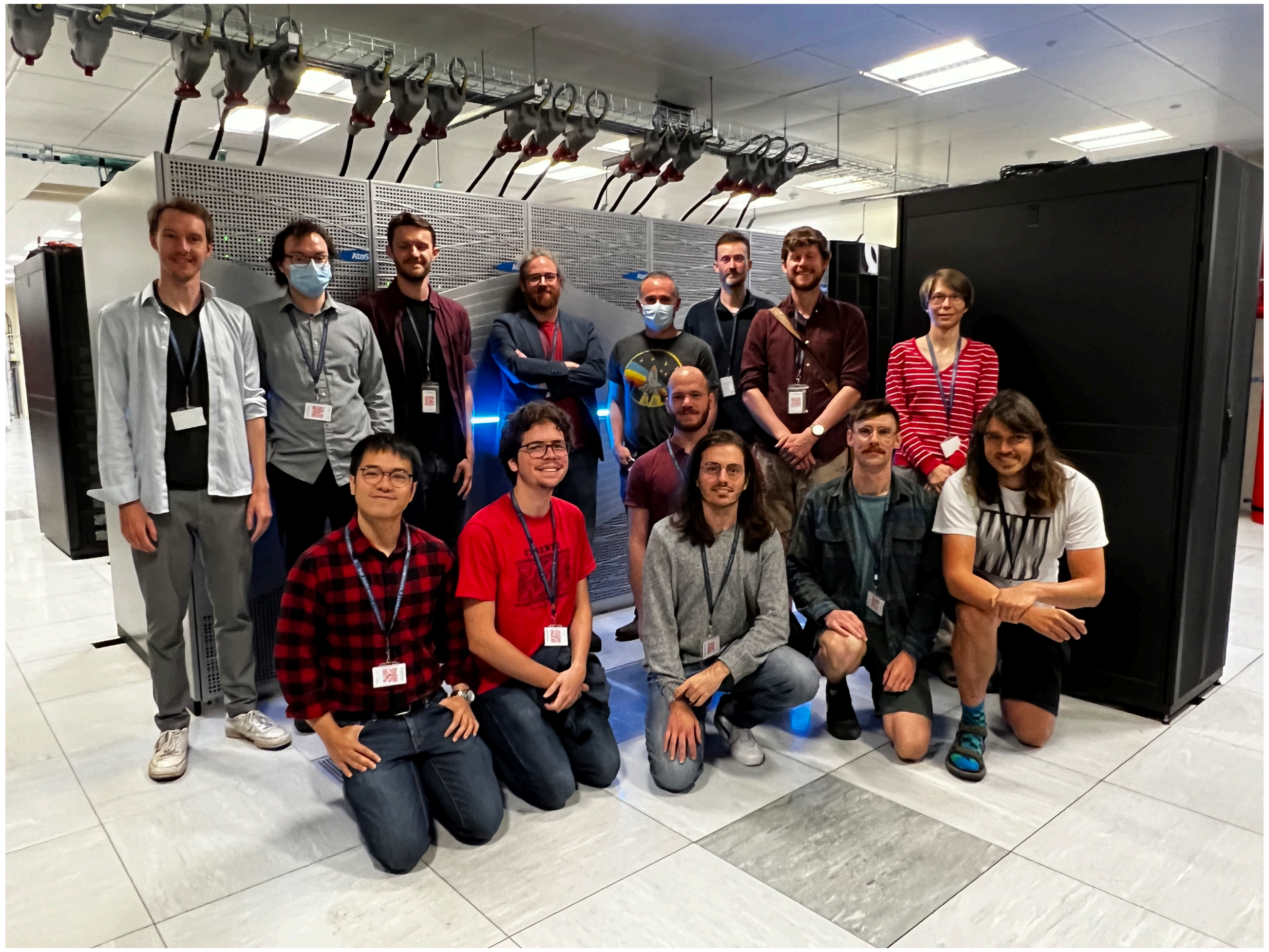
  https://doi.org/10.5281/zenodo.7057644

The Grid library

- C++14 data parallel C++ mathematical object library, **targeted at lattice QCD**
- **Cross-platform** with architecture-specific optimisations (x86, ARM, NVIDIA & AMD GPUs, …)
- Optimally use MPI, OpenMP and SIMD/SIMT parallelism under the hood
- Free and open-source (GPLv2)

  https://github.com/paboyle/Grid — https://doi.org/10.22323/1.251.0023

STFC DiRAC Tursa supercomputer



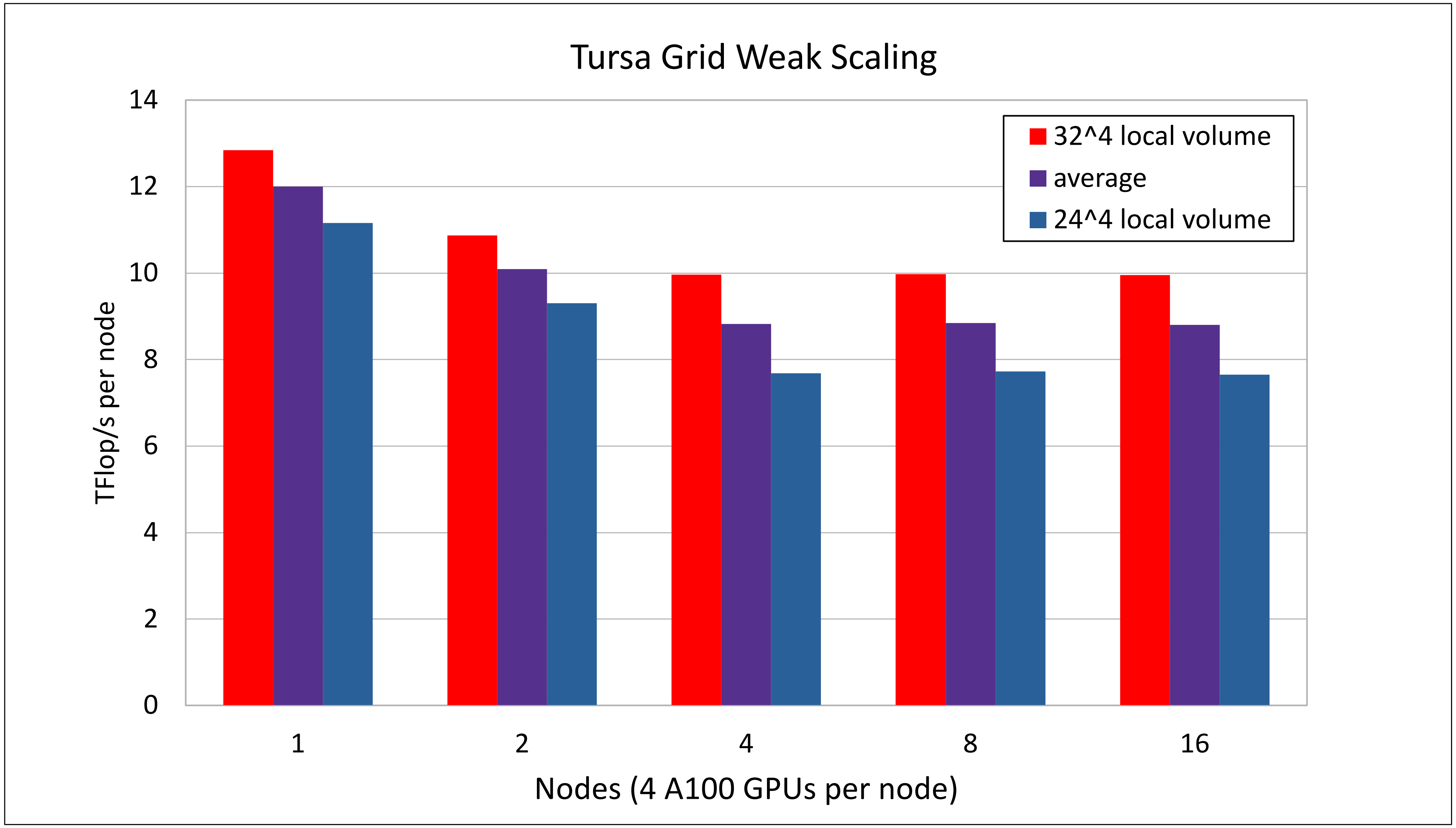
Edinburgh lattice team & Tursa, July 2022

UKRI Science and Technology Facilities Council DiRAC

epcc

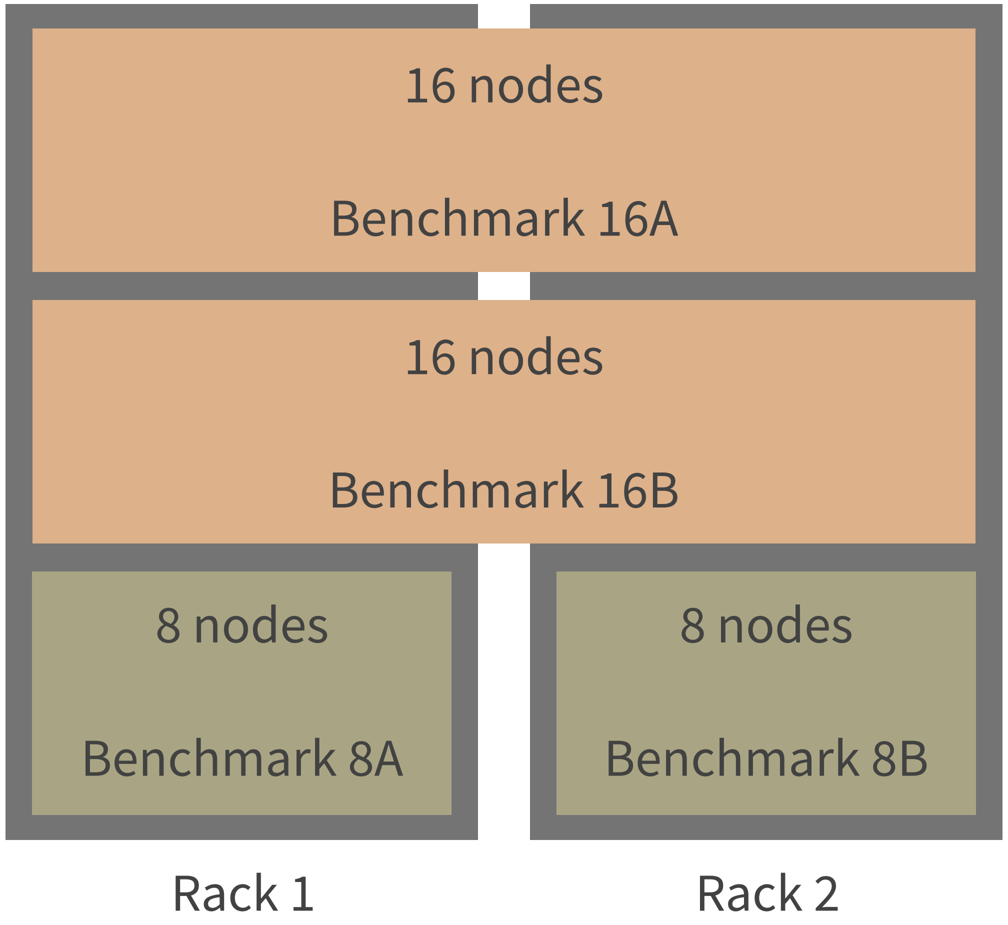
- Eviden BullSequana XH2000
- 468 NVIDIA A100 (+256 soon!)
- 4 x HDR200 NICs / node

Grid performances on Tursa

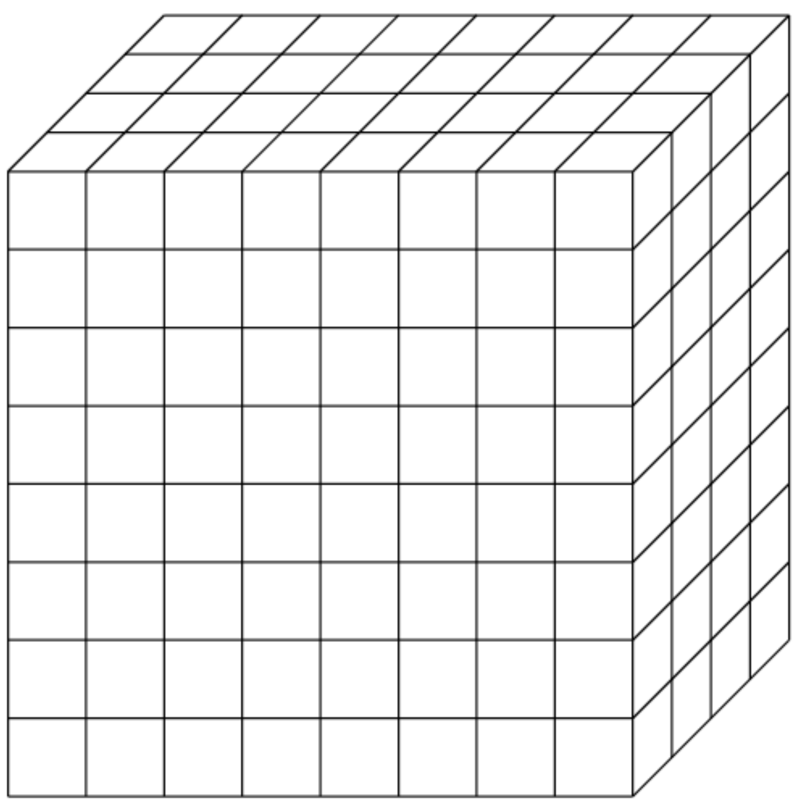


Benchmark setup

- Grid benchmark `Benchmark_dwf_fp32`, based on the single-precision domain-wall fermion sparse matrix
- 2 full XH2000 racks (48 nodes, 192 A100 GPUs)
- 2x16 nodes + 2x8 nodes
- Layout based on optimal communication topology
- Constant local problem size



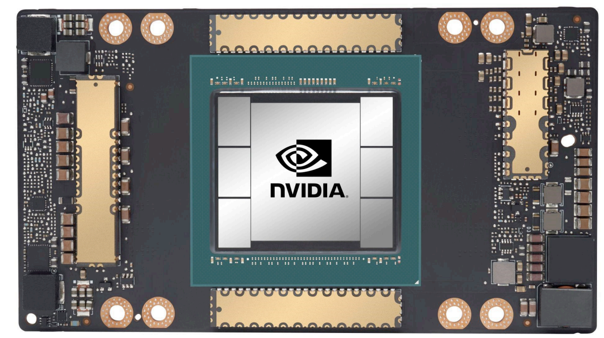
Problem size(s)



Two problem sizes

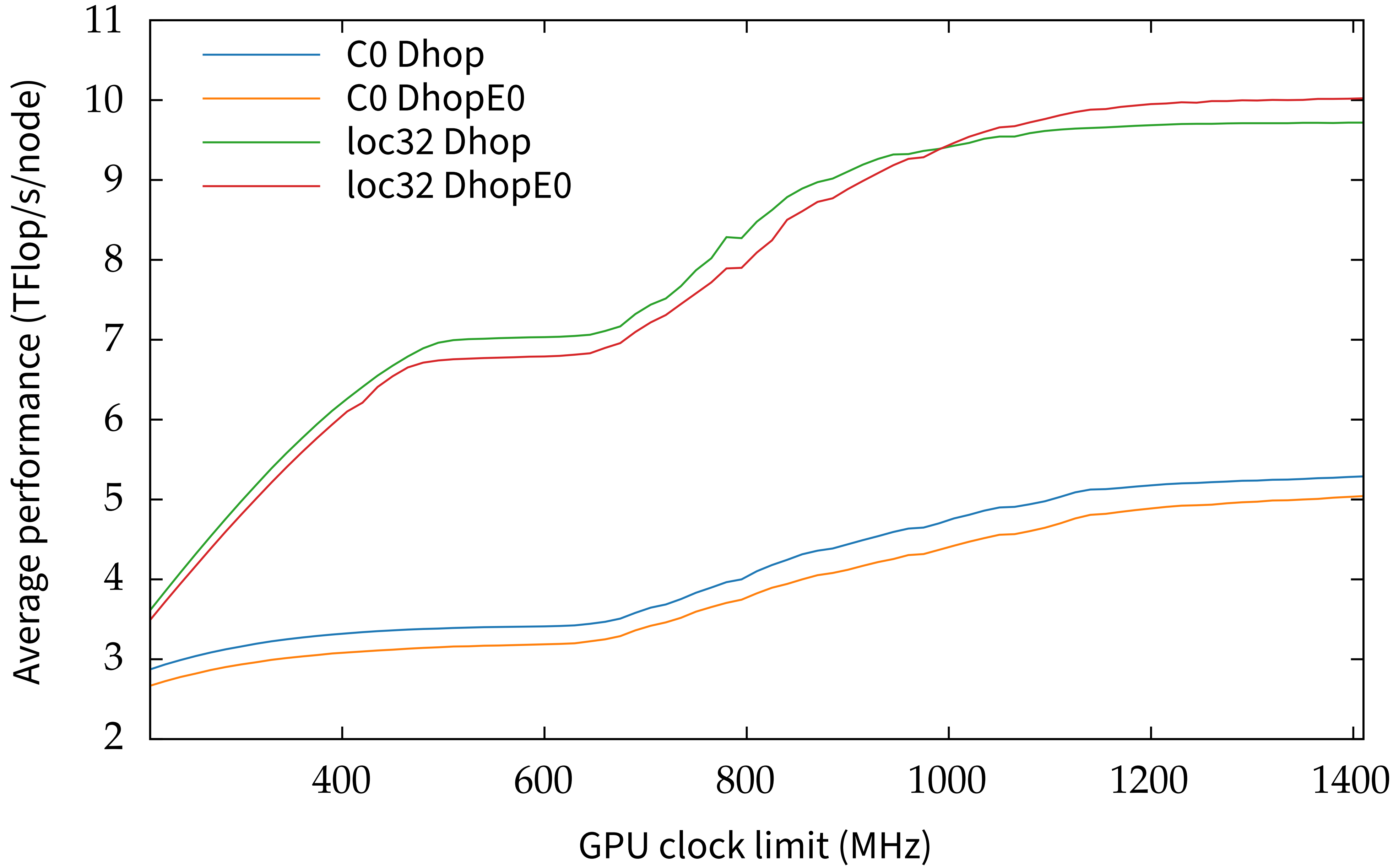
- “C0”: 24^3x12 local lattice
  used in production
- “loc32”: 32^4 local lattice
  realistic short-term future size

Power control and monitoring

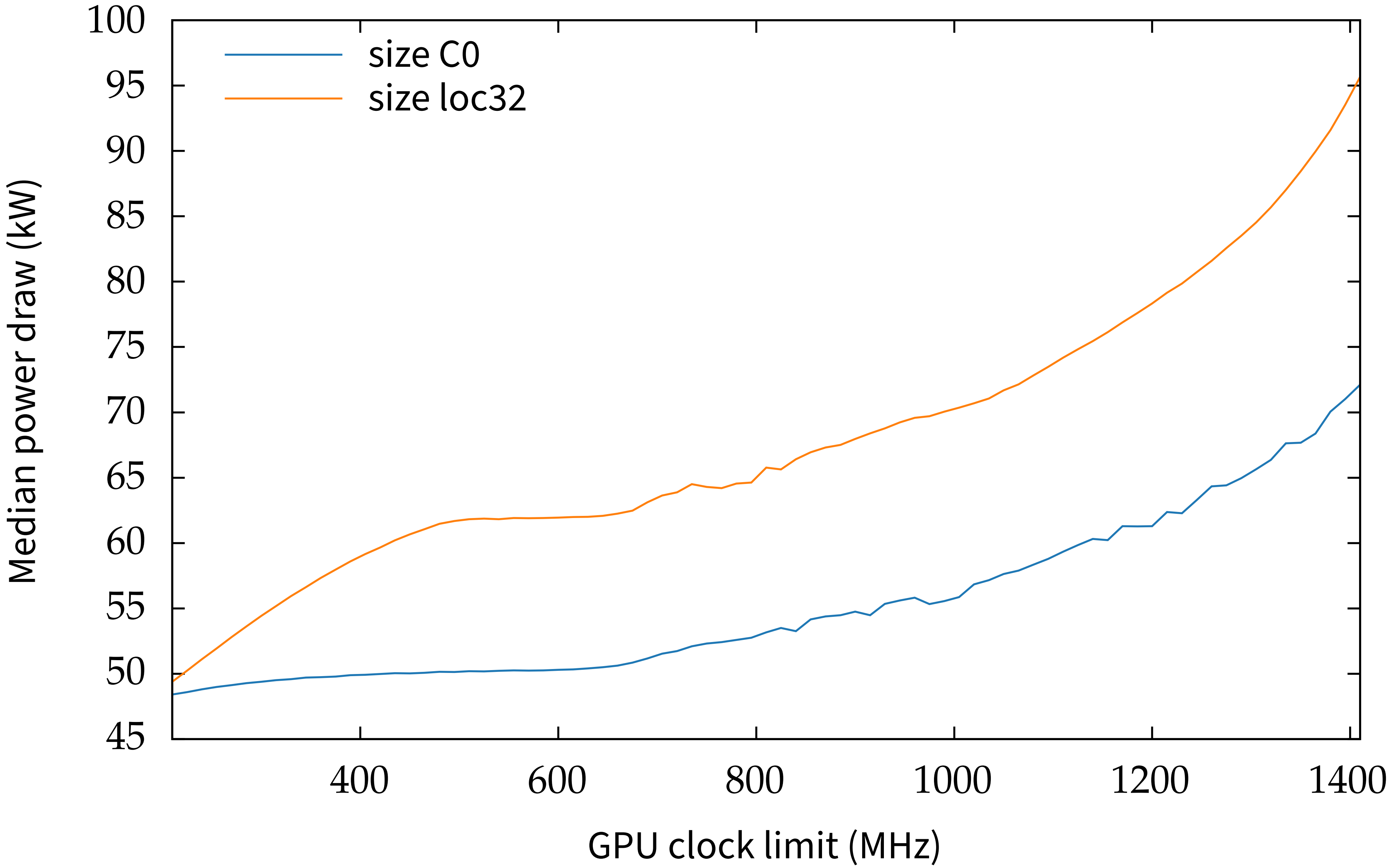


- Power controlled through **under-clocking of GPUs**
- Clock limit from **210 MHz to 1410 MHz** (increment 15 MHz)
- Default setting: maximum frequency 1410 MHz
- **Power monitoring**
  1) per GPU (NVIDIA SMI) 2) per rack (PDU through SNMP)

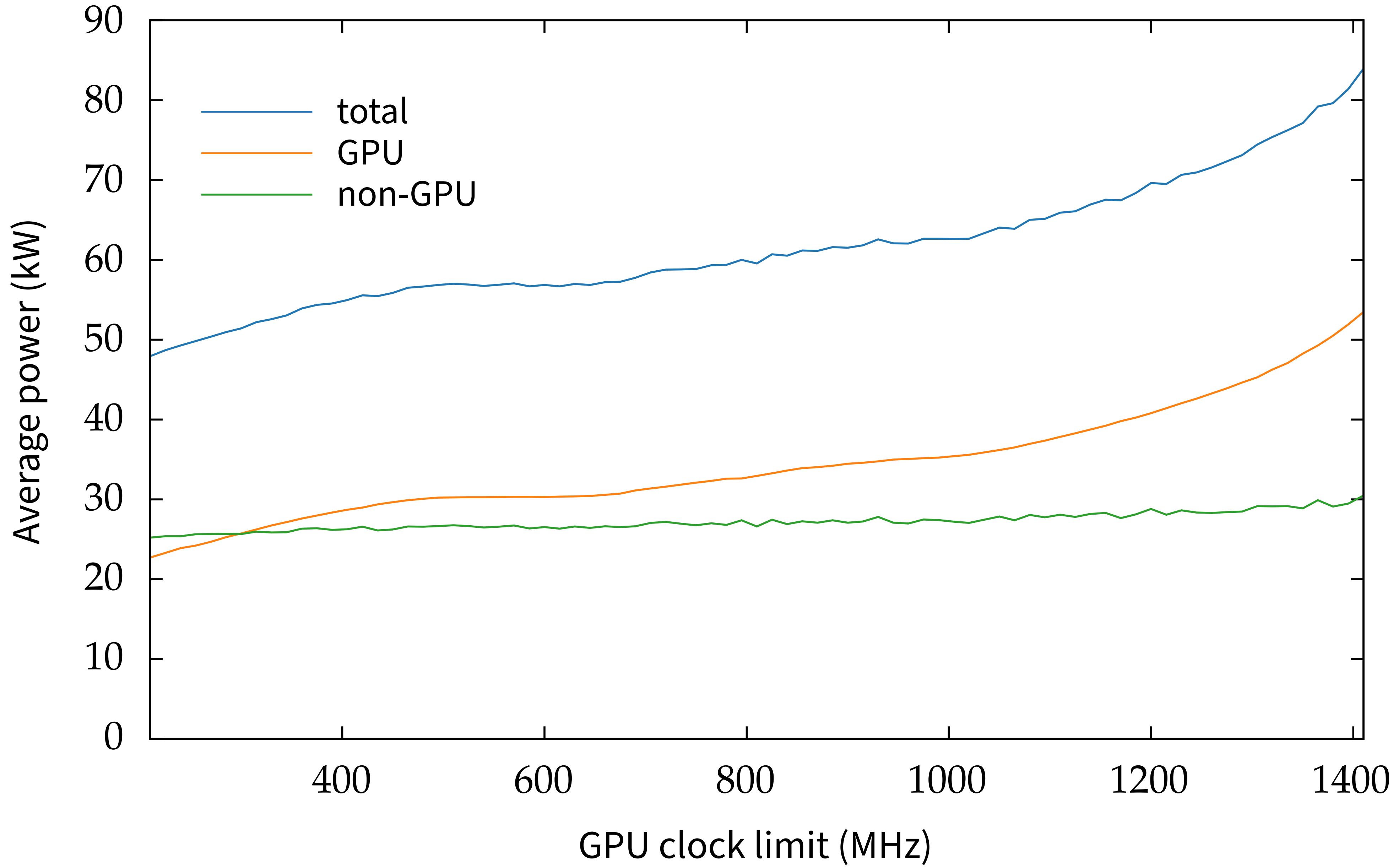
Performances vs GPU clock limit



Power draw vs clock limit

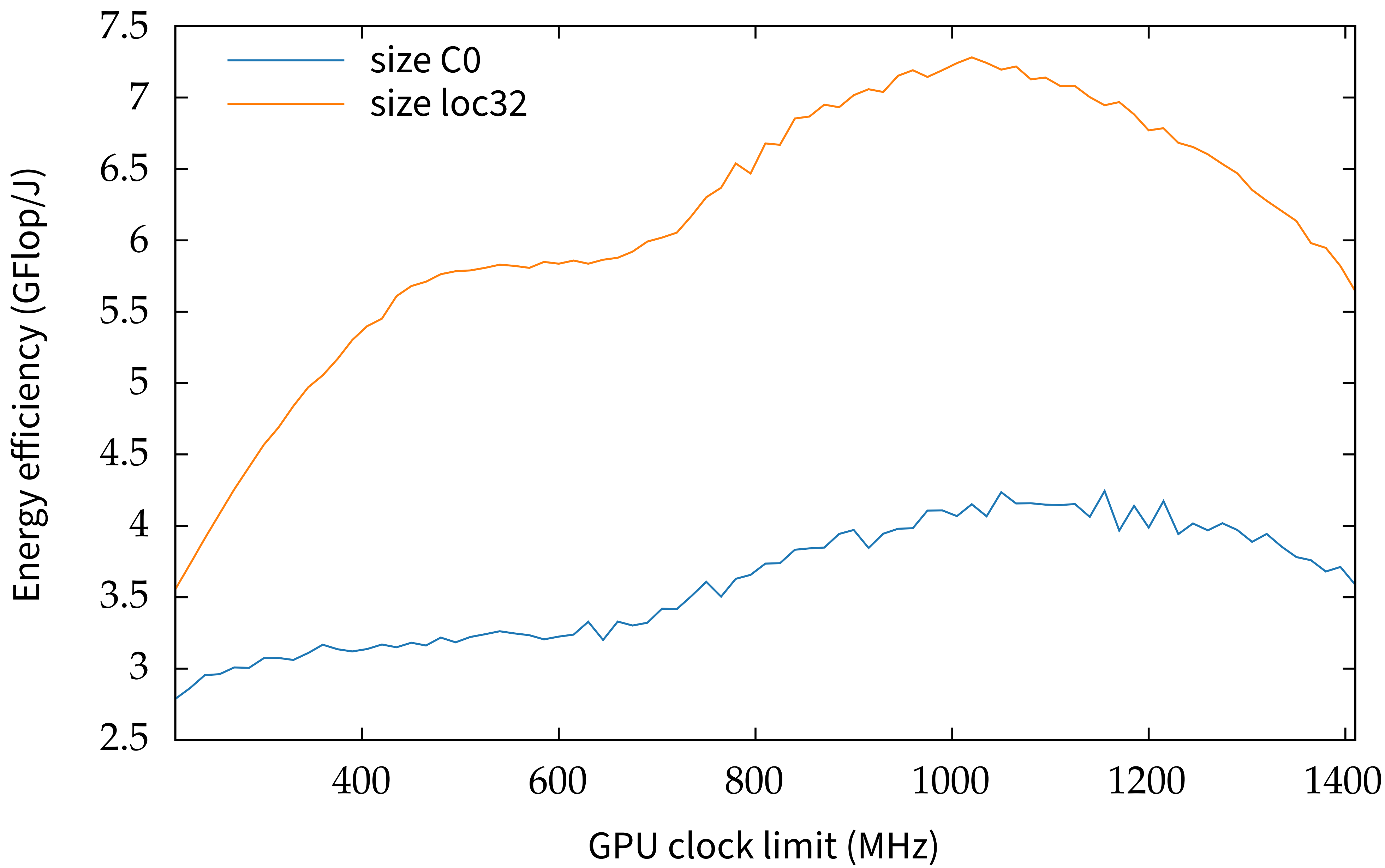


Power draw breakdown



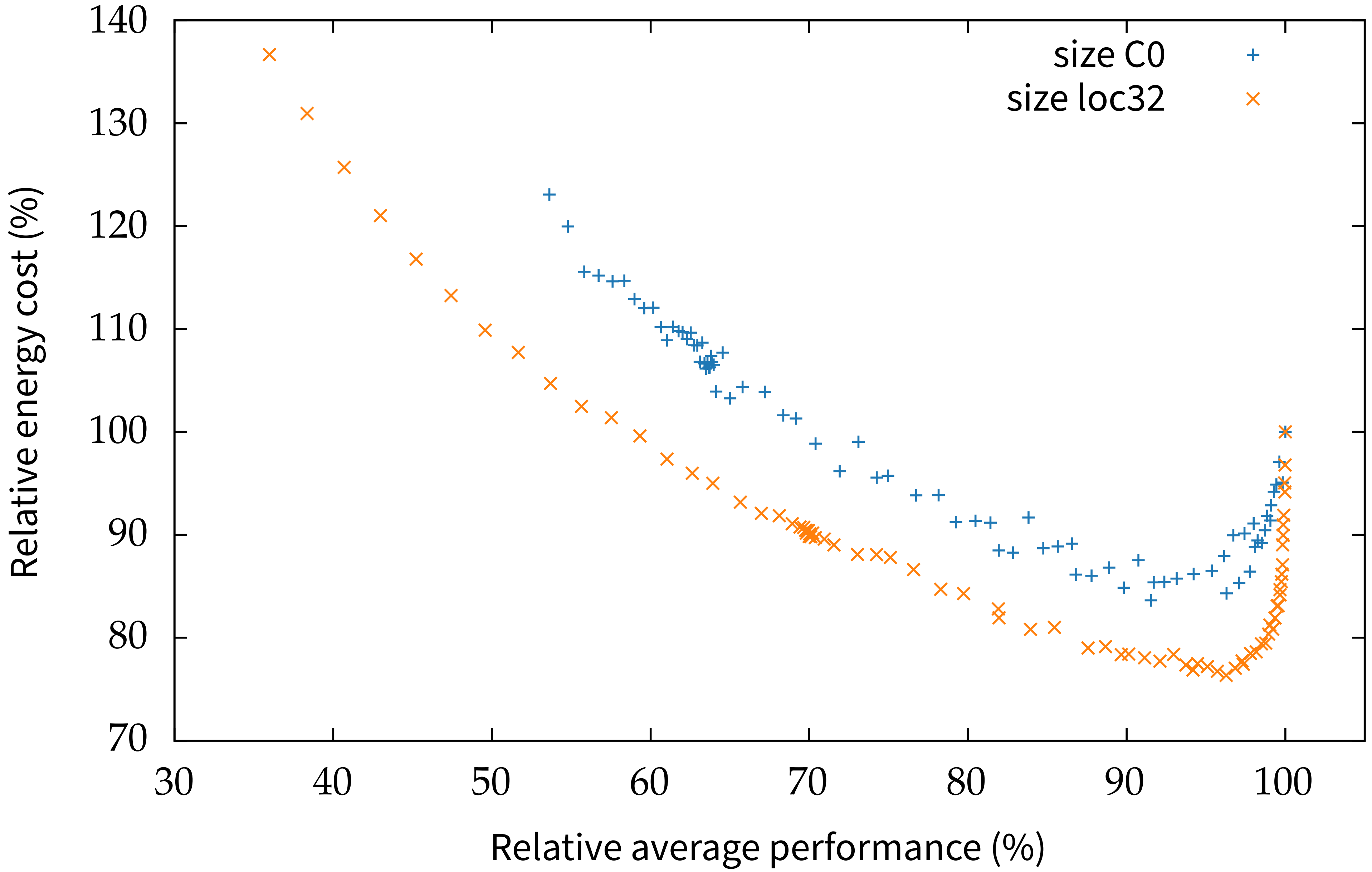
Non-GPU almost constant & consistent with idle power draw

Energy efficiency vs GPU clock



Default setting not energy-optimal!

Energy vs performance landscape



Outcome

- Tursa GPUs set to **1050 MHz by default** since Dec 22
- Monitoring show a 11% decrease in energy consumption
- Users reported no significant changes in throughput
- **Estimated energy savings are ~72,000 kWh** (today)

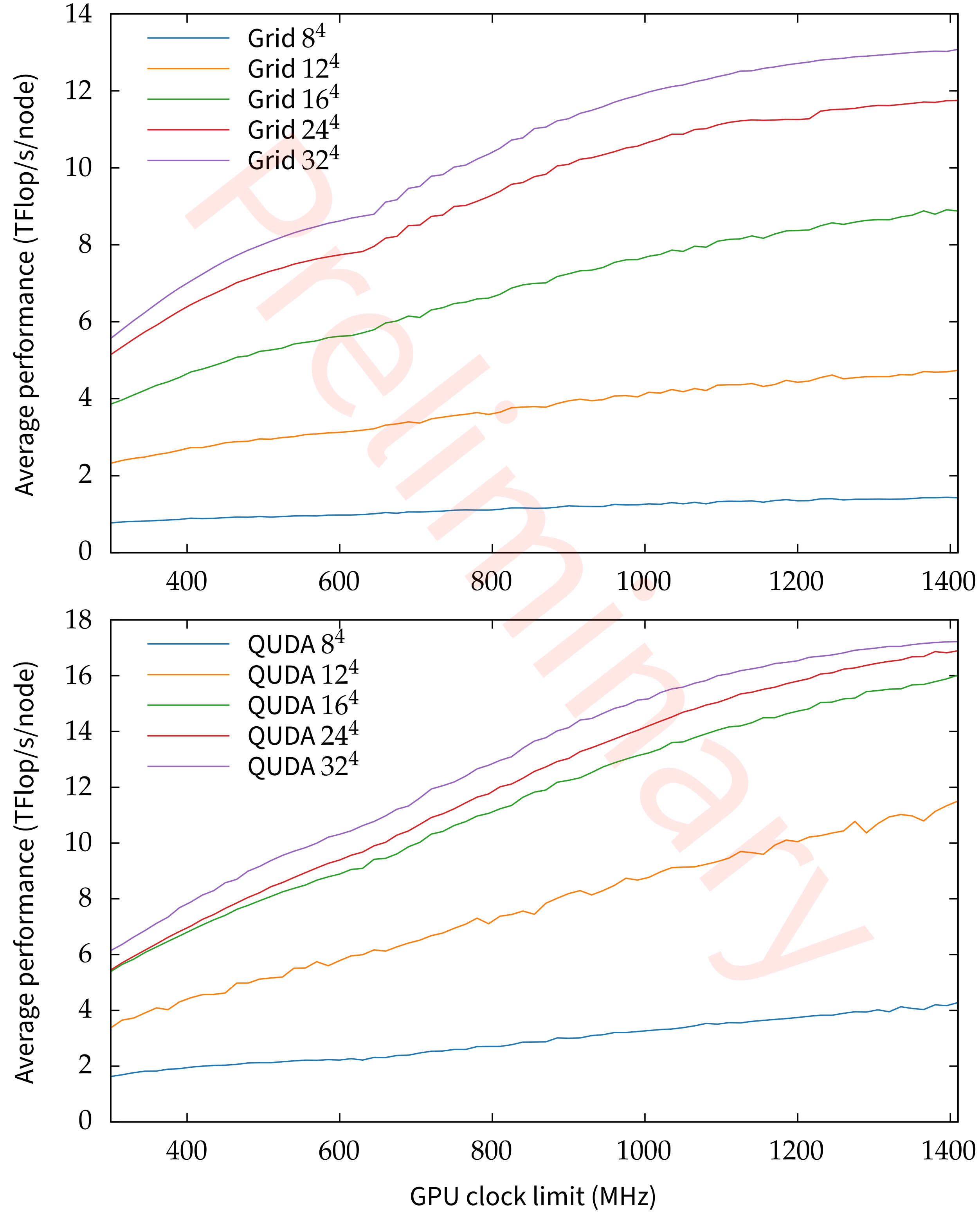
Beyond the Grid library

In collaboration with Simon Bürger (Edinburgh RSE)

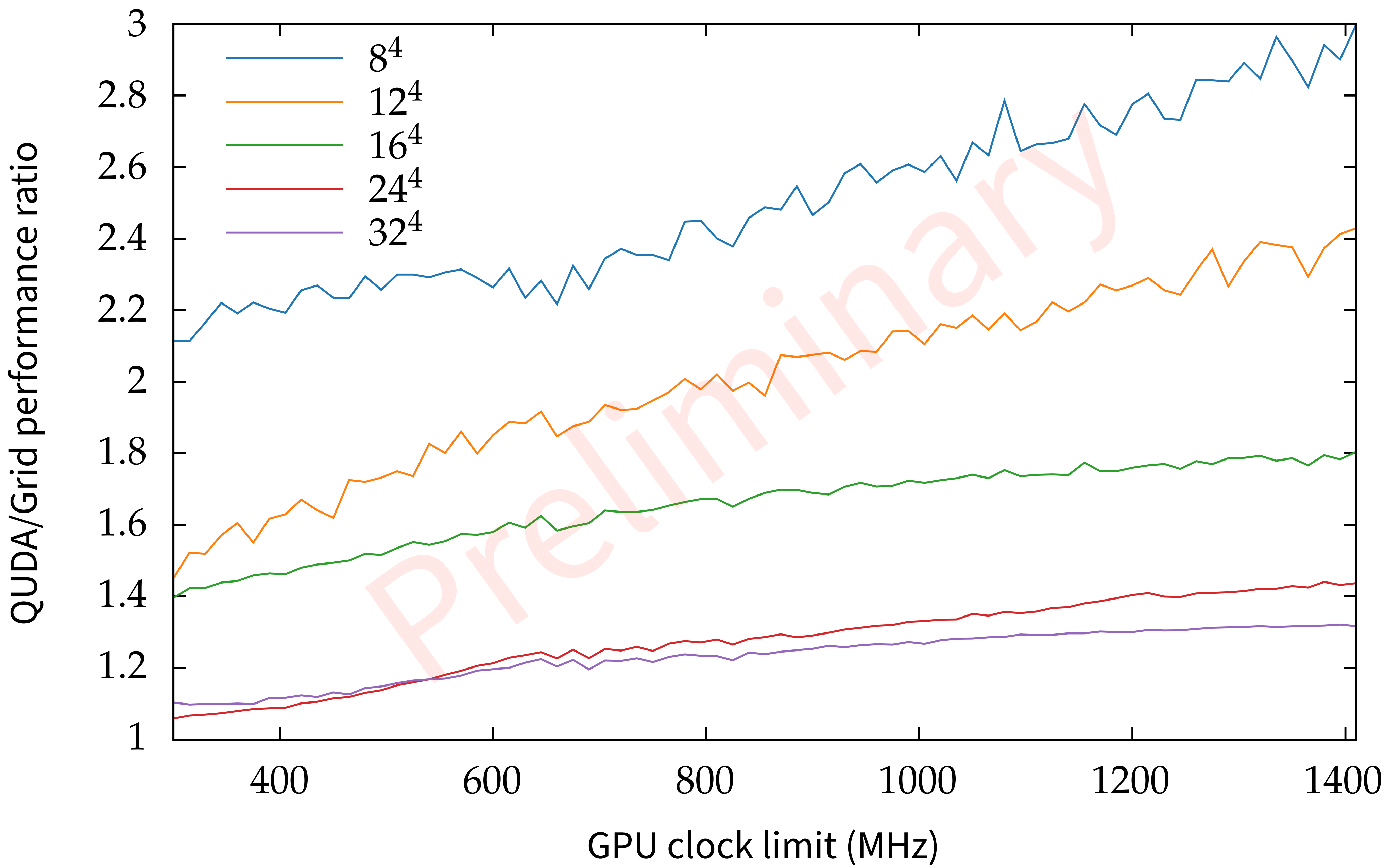
QUDA benchmark

- QUDA is one of the main library for lattice QCD on GPUs
- Open-source, developed and supported by NVIDIA
  https://github.com/lattice/quda
- Here: **custom QUDA benchmark**, matching Grid benchmark flop count and problem sizes
- Still using A100 GPUs on Tursa
- **For the moment single node, GPU power only**
  (work in progress)

Performance vs clock limit

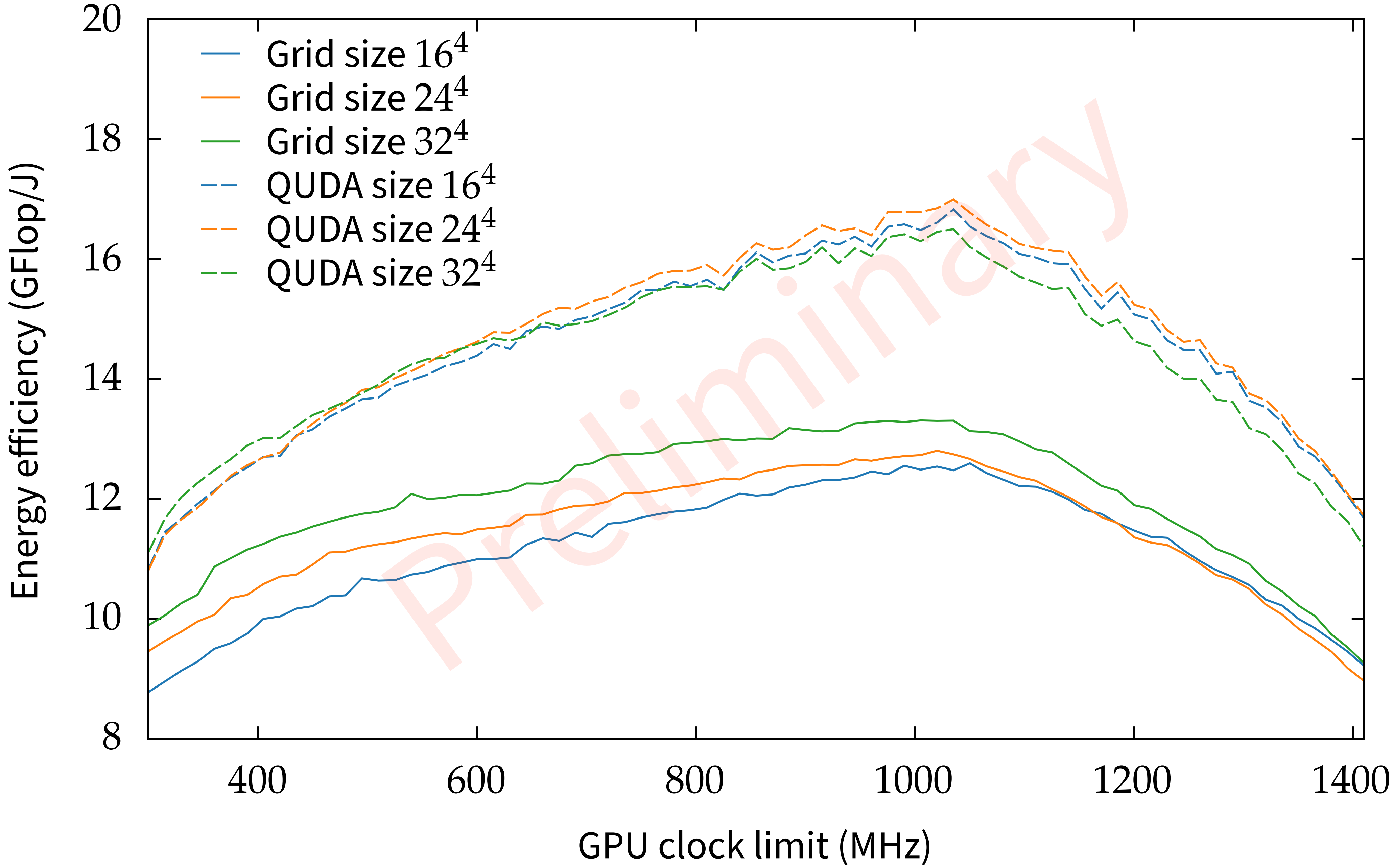


Performance vs clock limit, QUDA vs Grid

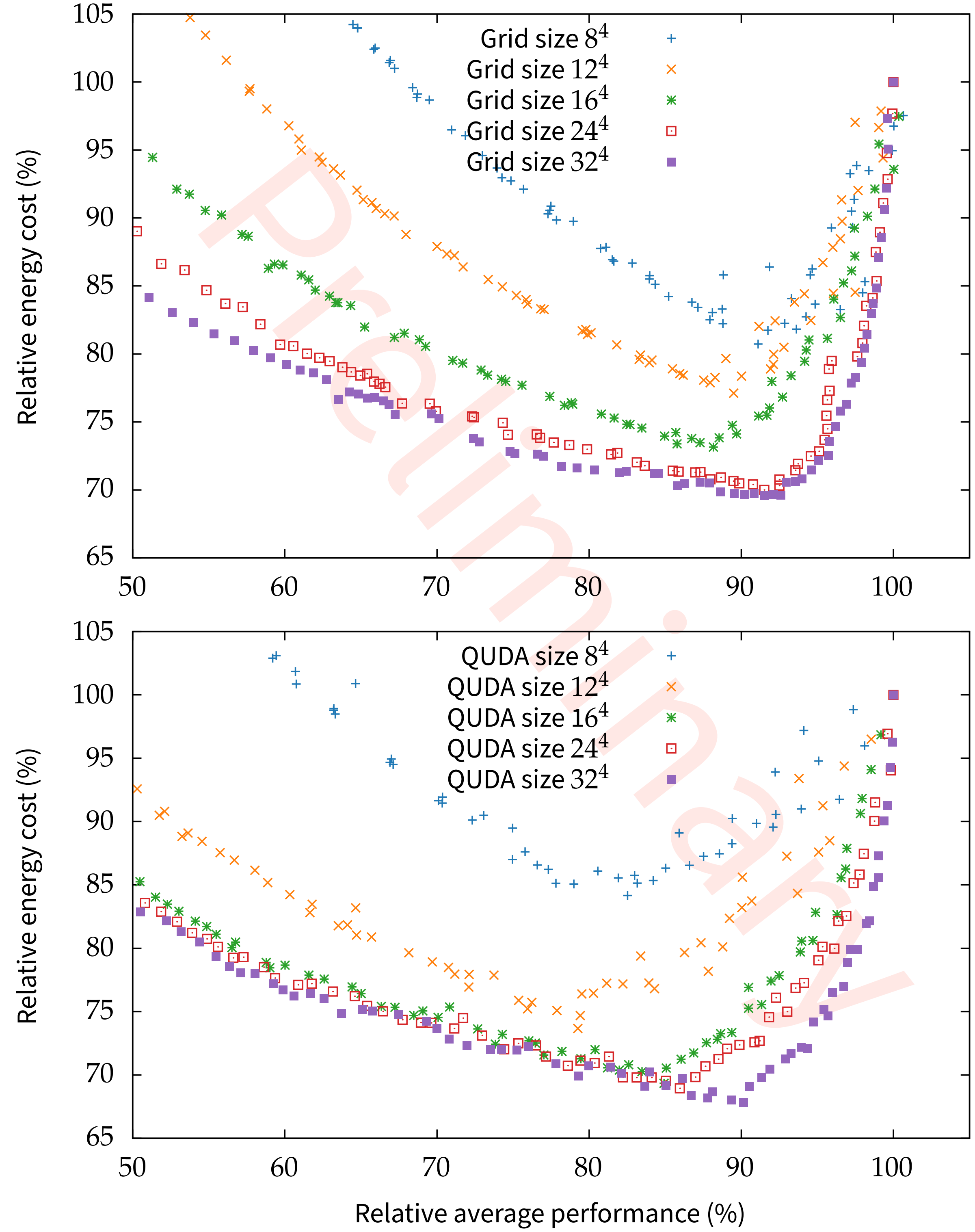


Remember: this is a single-node benchmark

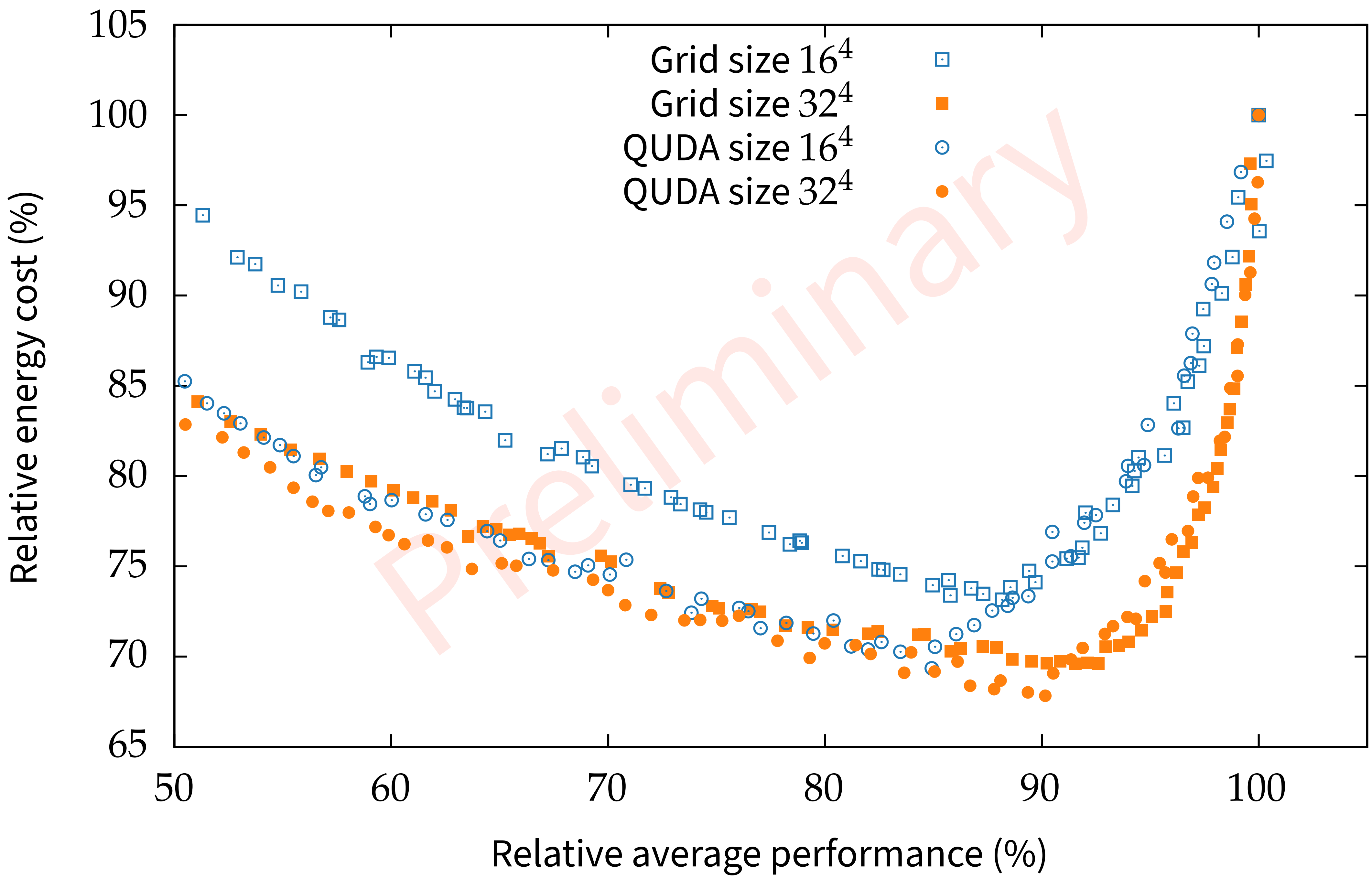
Energy efficiency, QUDA vs Grid



Energy vs performance landscape



Energy vs performance landscape, QUDA vs Grid



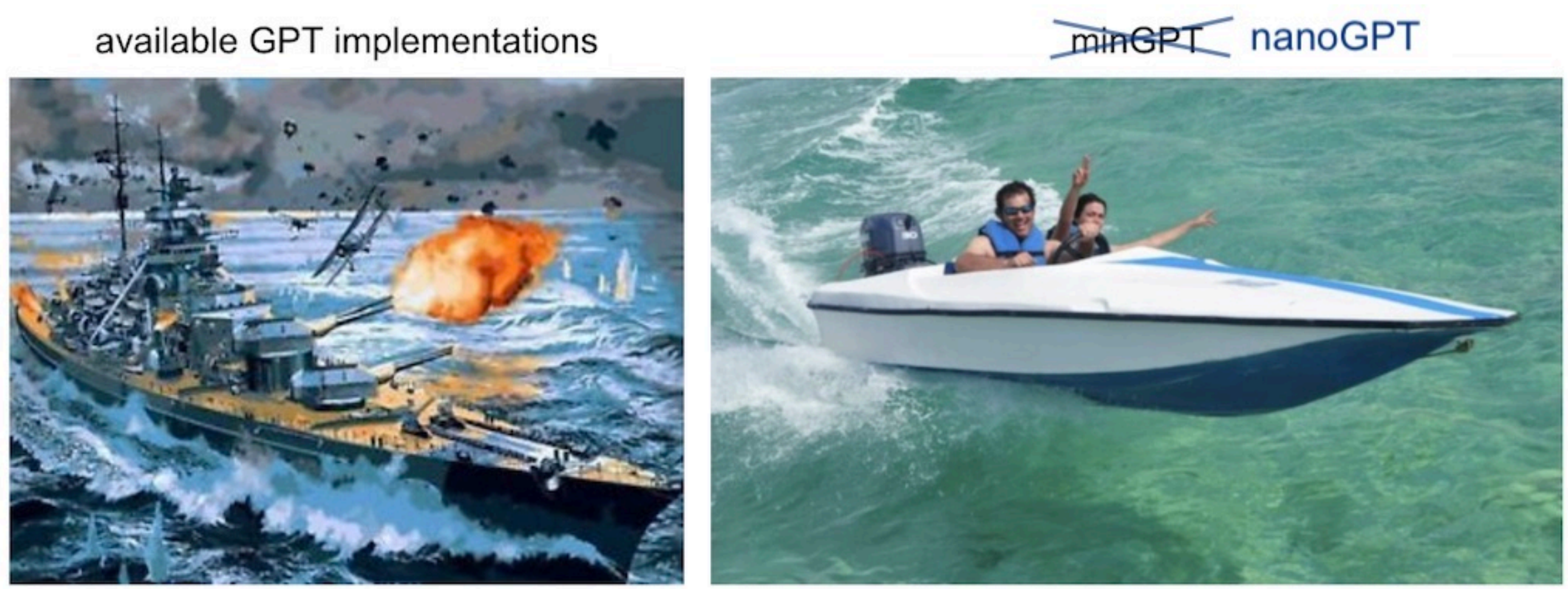
Conclusion

- QUDA and Grid share an **energy-optimal point at 1 GHz**
- QUDA significantly faster than Grid for small sizes, more similar for large sizes
- Different energy profiles for small sizes, almost identical at large sizes
- To be extended on multiple nodes!

GPT language model training

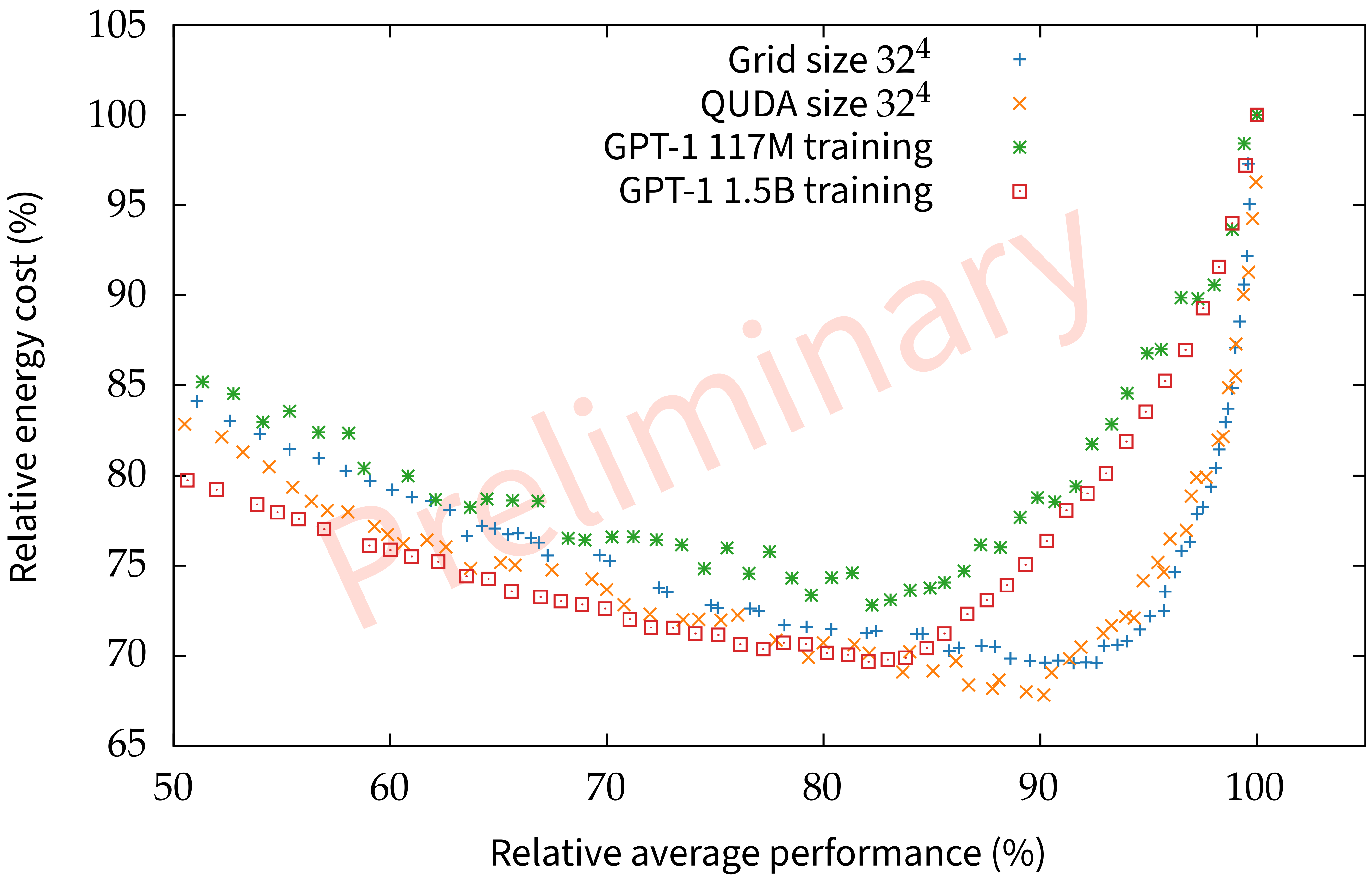
In collaboration with Fabian Joswig (DeepL, formerly Edinburgh)

Setup



- nanoGPT: open-source reproduction of GPT-2
- OpenWebText2 training set (whole of Reddit 2005-2020)
- Setup to reproduce GPT-1 (117 M) and GPT-2 (1.5 B)
- Single node 4x GPUs, ~700 TFlop/s for GPT-2 🤯

Results

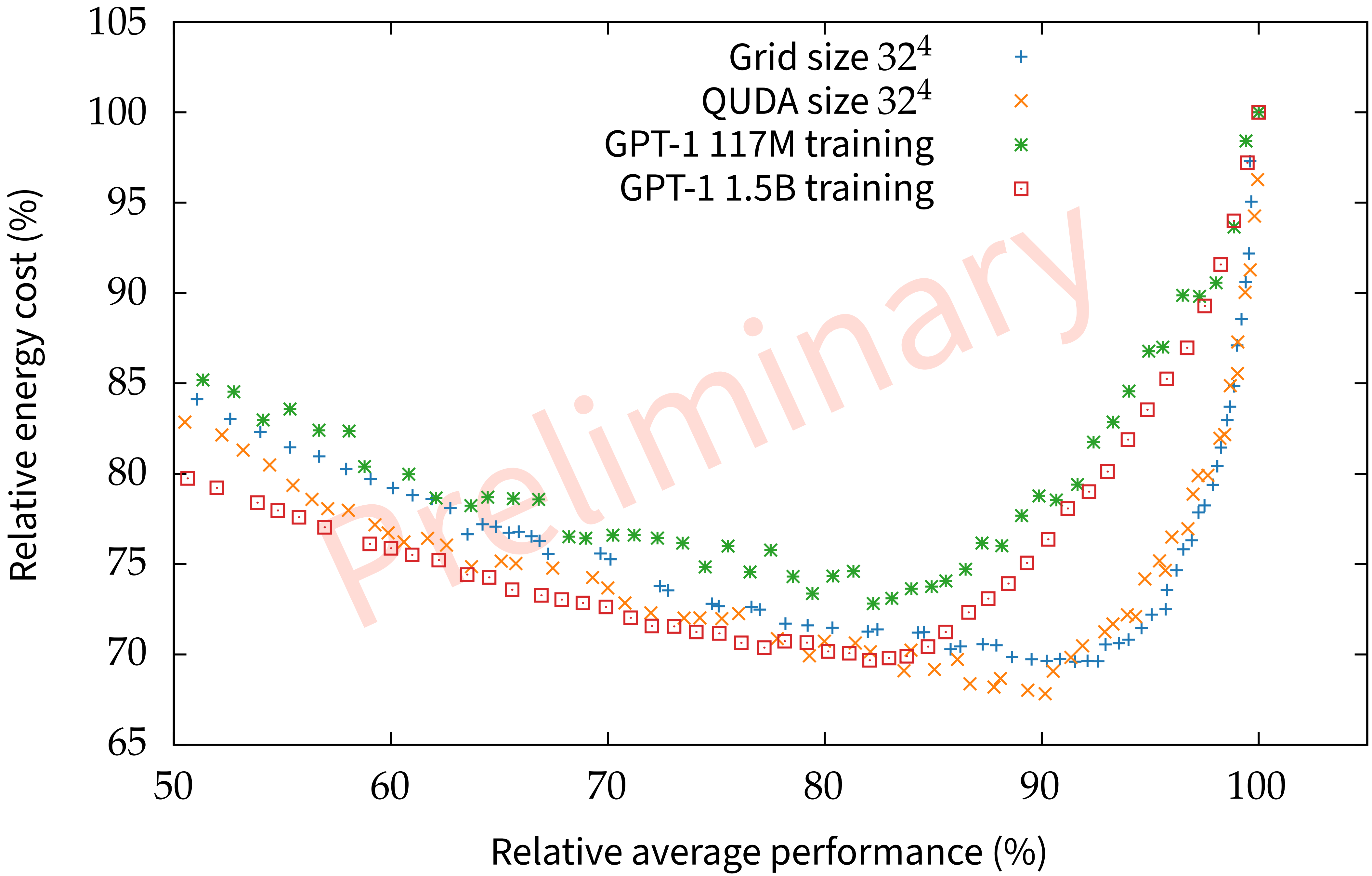

Grid size 32^4
QUDA size 32^4
GPT-1 117M training
GPT-1 1.5B training
Preliminary
Relative energy cost (%)
Relative average performance (%)
105
100
95
90
85
80
75
70
65
50
60
70
80
90
100

Conclusions

- A100 frequencies around **1 GHz** generally lead to 20-30% more energy efficient computations (GPUs only)
- Energy saving potentially reduced by non-GPU elements
- Impact on floating-point performances within
  10% (lattice) & 20% (LLM training)
- **Lower default frequencies** recommended on GPU clusters

Perspectives

- **Larger scale tests** with more accurate power monitoring
- **Multi-node LLM training** with model parallelism
- Extension to **other architectures & domains**
- **Multi-objective optimisation** across domains
- Toward policy changes: **should energy-efficiency become a standard performance figure?**



Thank you!

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreements No 757646 & 813942.