Benchmarking Modern HPC Architectures with Lattice QCD

Ryan Hill

in collaboration with Antonin Portelli

20$^{\mathrm{th}}$ May 2026
Nordic Lattice Meeting 2026

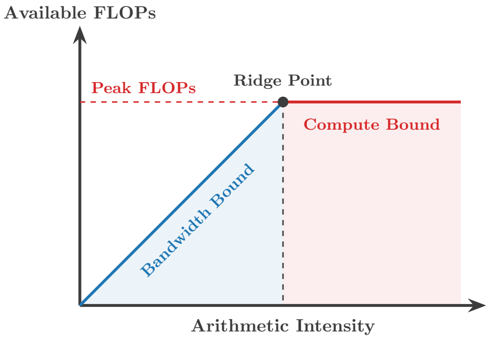
Motivation

- We want meaningful benchmarks of our codes against HPC systems.
  - Measuring a $2\times$ improvement in time-to-solution between two systems sounds good...
  - But if you *should* have achieved $4\times$, you've lost half your potential science output.
- Benchmark against the hardware: measure fraction of **peak theoretical performance**.
- Lattice QCD might achieve single-digit percentages of peak performance.
  $\rightarrow$ We will see that this can be **expected and ok**.
- We'll also see that Lattice QCD probes system networks more rigorously than the main HPC benchmarks.
  $\rightarrow$ Lattice QCD has a role as a **system benchmark**.

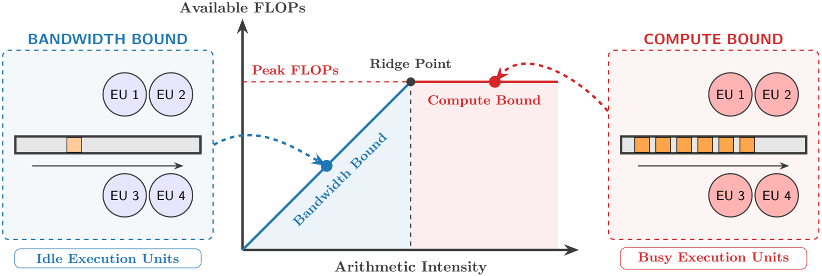
Measurement

- How much of a machine's theoretical capacity can we hit?
- Dirac operator multiplies...
  - ...involve **predictable**, **packed** data accesses.
    $\Rightarrow$ expect *throughput* to be the limit, not *latency*.
  - ...are **floating-point algorithms**, not control flow, I/O...
    $\Rightarrow$ *FLOPs* are a faithful, *portable* measure of work.

- These are exactly the workloads that GPUs are designed for.
- Such algorithms are well-captured by a *roofline model*.

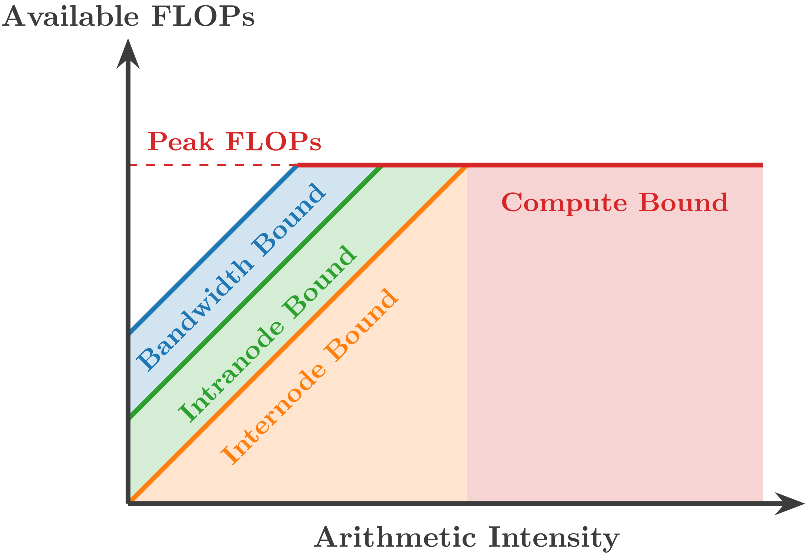


The Roofline Model

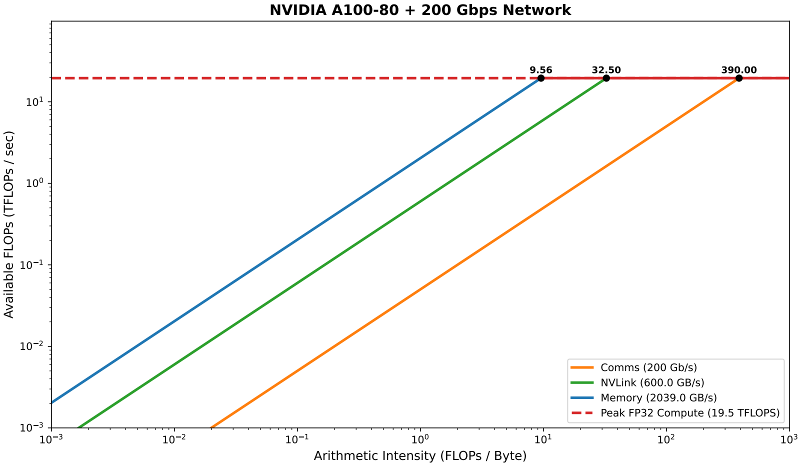


- Assumes processor performance depends on two metrics:
  - Speed at which arithmetic work can be performed (Flops/sec)
  - Volume of floating-point data can be read/written (B / sec)
- Natural input variable “Arithmetic Intensity”: FLOPs/Byte.
- Predicts the theoretical maximum FLOPs rate for a given AI.

The Roofline Model

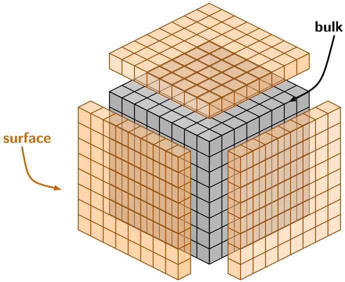


The Roofline Model



The Dirac Operator Roofline Model

- Dirac operator multiplies require **nearest-neighbour** elements.
- Calculation decomposed across **multiple MPI ranks**.
- Neighbours held on different MPI ranks transferred *via* the **Halo Exchange** algorithm:
  - 3D "surfaces" transferred between MPI ranks.
  - Calculation in 4D "bulk".



Flops	FP32 Memory Transfer	FP32 Comms Transfer
$L^4 L_s \times 1344$	$L^4 L_s \times 192$ Bytes	$L^3 L_s \times 768$ Bytes

- Memory Arithmetic Intensity: 7 Flops/Byte
- Comms Arithmetic Intensity: $1.75L$ Flops/Byte

The Dirac Operator Roofline Model

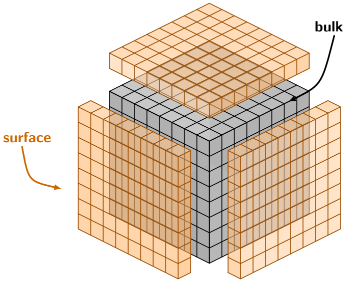
- Dirac operator multiplies require **nearest-neighbour** elements.
- Calculation decomposed across **multiple MPI ranks**.
- Neighbours held on different MPI ranks transferred *via* the **Halo Exchange** algorithm:
  - 3D "surfaces" transferred between MPI ranks.
  - Calculation in 4D "bulk".

Flops	FP32 Memory Transfer	FP32 Comms Transfer
$L^4 L_s \times 1344$	$L^4 L_s \times 192$ Bytes	$L^3 L_s \times 768$ Bytes

- Memory Arithmetic Intensity: 7 Flops/Byte
- Comms Arithmetic Intensity: $1.75L$ Flops/Byte

The Dirac Operator Roofline Model

- Dirac operator multiplies require **nearest-neighbour** elements.
- Calculation decomposed across **multiple MPI ranks**.
- Neighbours held on different MPI ranks transferred *via* the **Halo Exchange** algorithm:
  - 3D "surfaces" transferred between MPI ranks.
  - Calculation in 4D "bulk".



Flops	FP32 Memory Transfer	FP32 Comms Transfer
$L^4 L_s \times 1344$	$L^4 L_s \times 192$ Bytes	$L^3 L_s \times 768$ Bytes

- Memory Arithmetic Intensity: 7 Flops/Byte
- Comms Arithmetic Intensity: $1.75L$ Flops/Byte

The Dirac Operator Roofline Model

P_{mem}: $7B_{\text{mem}}$ Flops/s, P_{comms}: $1.75LB_{\text{comms}}$ Flops/s

- Roofline model predicts the *memory-bound* regime to depend only on the hardware memory bandwidth.
- In the *network-bound* regime, the performance scales with the network bandwidth **and** linearly with the local lattice dimension.

Benchmark

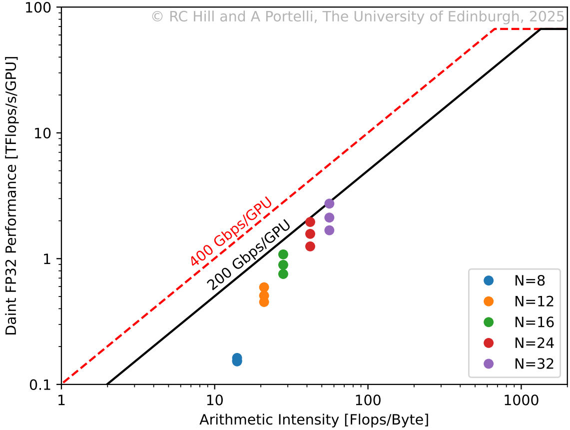
- Following results based on `grid-benchmark` https://github.com/aportelli/grid-benchmark
- All presented data publicly available in above repository alongside the benchmark itself.
- Forked from the `Benchmark_ITT` benchmark in the `Grid` repository.

GRID

Thanks to Antonio Rago and Maxwell T Hansen for contributing computing resources for this project.

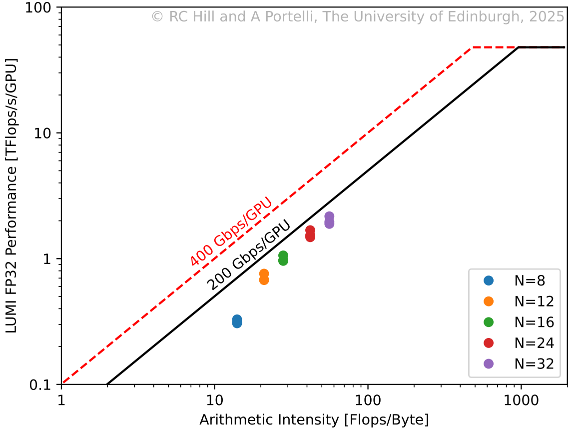
Benchmark Results

Daint (CSCS) Grid DWF benchmarks on 64, 128, and 256 GH200

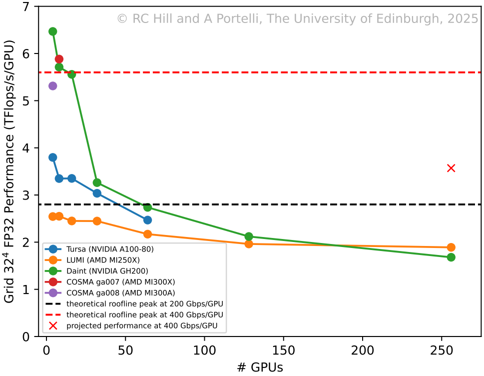


Benchmark Results

LUMI (CSC) Grid DWF benchmarks on 64, 128, and 256 MI250X

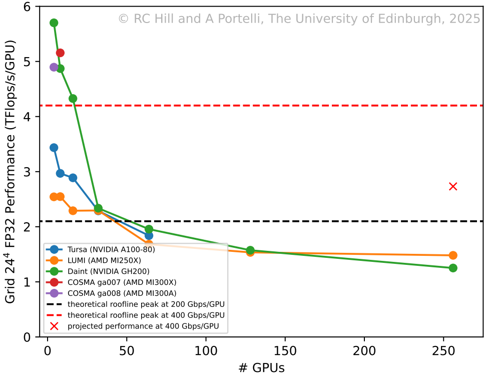


Benchmark Results



- Large-scale performance **independent of GPU model!**
- Contemporary hardware (MI300, GH200) is **held back** by 200 Gbps networks for lattice applications.
- **TOP500** rankings won't show this!

Benchmark Results

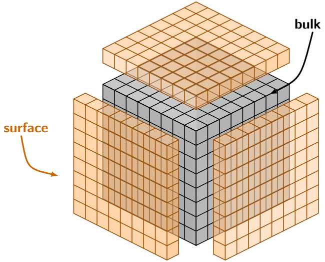


- Large-scale performance **independent of GPU model!**
- Contemporary hardware (MI300, GH200) is **held back** by 200 Gbps networks for lattice applications.
- **TOP500** rankings won't show this!

Lattice QCD as a System Benchmark

- TOP500 uses the HPL benchmark to rank supercomputers.
  - Designed to be highly **compute-bound**.
  - Decides if a system qualifies as *e.g.* "exascale".
- Alternatively ranked by the HPCG benchmark.
  - Low-intensity sparse matrix, tests bandwidth-bound regime.
  - However, HPCG solves a 3D problem, not 4D.
  - Typically **memory-bound** on modern systems, **not network-bound**.

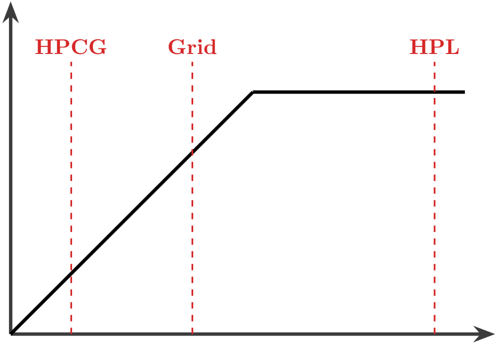
- **Neither** ranking effectively probes modern machine interconnects, the **main bottleneck** for lattice codes.



Lattice QCD as a System Benchmark

Taken together, these benchmarks explore the main regimes of the roofline model:

- **HPL**:
  → High intensity
  → Compute bound.
- **Grid**
  → Medium intensity
  → Network bound.
- **HPCG**
  → Low intensity
  → Memory bound.



UKNNSS-Benchmarks

This page provides links to documentation and resources associated with benchmarks for the UK NNSS procurement process.

The benchmarks will play a key role in the procurement process and span full application benchmarks to synthetic micro-benchmarks. Please ensure you read the benchmark descriptions carefully to understand the rules around source code modification, mandatory runtime configuration and options and the data that needs to be submitted at various stages in the procurement process.

This page and the benchmark descriptions will be maintained and updated throughout the procurement process. Updates will be recorded in the CHANGELOG.

Application Benchmarks

The application benchmarks have been chosen to exercise a range of system requirements and technical characteristics in areas of key importance to the UK research community.

- CP2K
- DOLFINx
- Grid
- LAMMPS

Latest Hardware

GPU Model	Memory (GB)	Memory BW (TB/s)	FP32 Compute (TFLOPs/s)
NVIDIA A100-80	80	2.0	19.5
AMD MI250X	128	3.2	47.9
NVIDIA GH200	96	4.0	67.0
NVIDIA H200	141	4.8	67.0
AMD MI300A	128	5.3	122.6
AMD MI300X	192	5.3	163.4
NVIDIA B200	192	8.0	75.0
AMD MI355X	288	8.0	157.3
(TBC) AMD MI430X	432	19.6	???
(TBC) NVIDIA Rubin GPU	288	22.0	130.0

- We need network speeds to keep pace with memory bandwidths to make full use of modern hardware.
- **Intraconnect** speeds also need to be carefully balanced against memory and interconnect.
- Huge (!) jump in memory bandwidth (claimed) in upcoming HBM4 memory chips.

Conclusions

- A relatively simple roofline model can give good network-bound performance estimates for Grid.
- Lattice QFT operates in a different regime to the TOP500 benchmarks.
- The UK's New National Supercomputing Service will include a Grid-based lattice QCD benchmark as one of the four official application benchmarks.
- **400Gbps+ networks are critical for modern lattice QFT.**