Lattice QCD as a key benchmark for exascale systems

Antonin Portelli (The University of Edinburgh) / 15 November 2025 / R-CCS, Kobe, Japan

Roofline modelling of the Dirac-Wilson operator

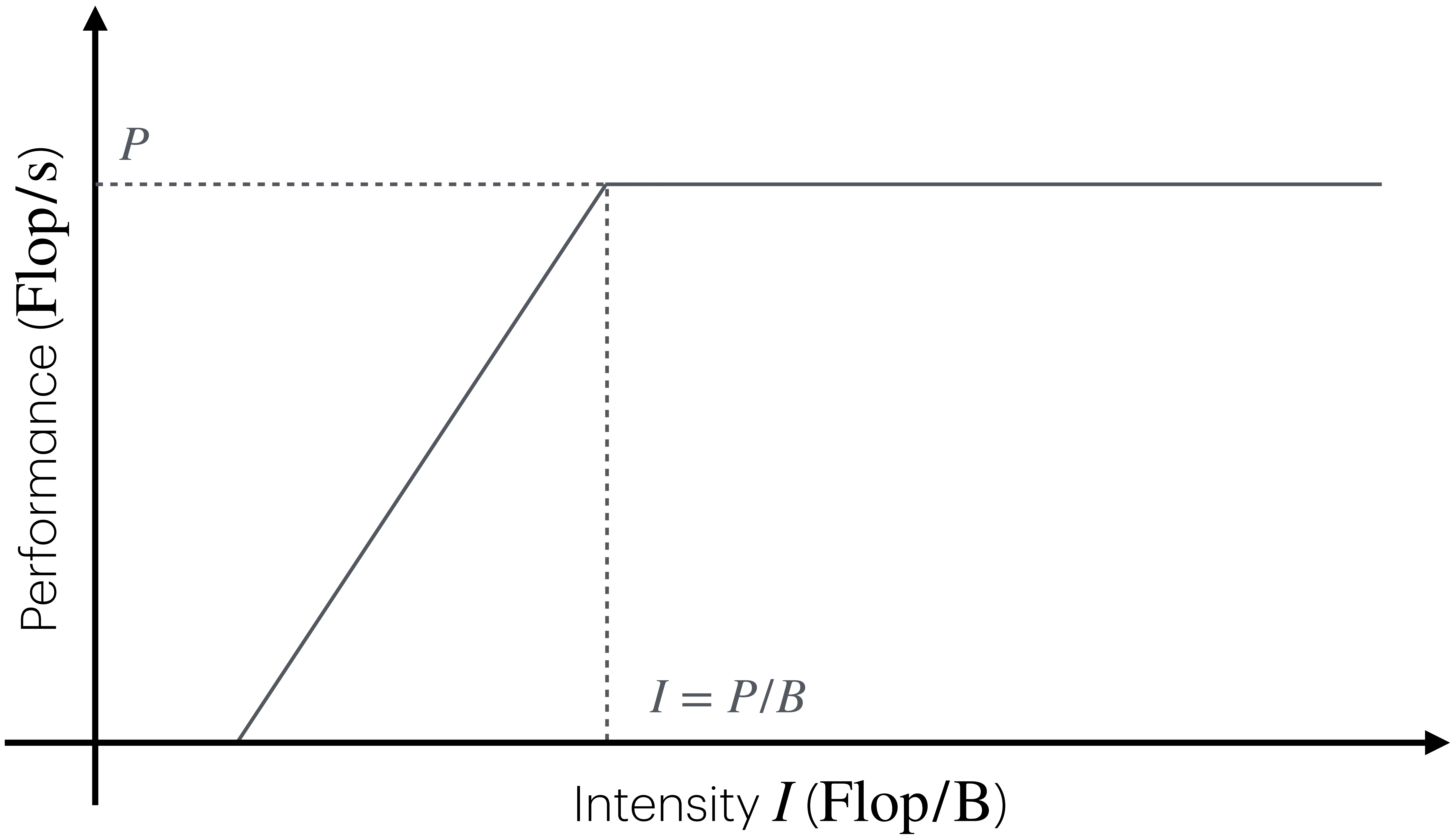
The roofline model

Basic principles

- **Core model assumption** — a computer does two things:
  1. It reads and writes numbers from and into memories
  2. It processes numbers into others using arithmetic operations
- Therefore the performance of a program is determined by
  1. How fast the computer can read/write numbers (in $\mathrm{B/s}$)
     and how fast it can process them (in $\mathrm{Flop/s}$)
  2. How much operations per byte the program needs to perform.
     This is the **arithmetic intensity** (in $\mathrm{Flop/B}$)

The roofline model

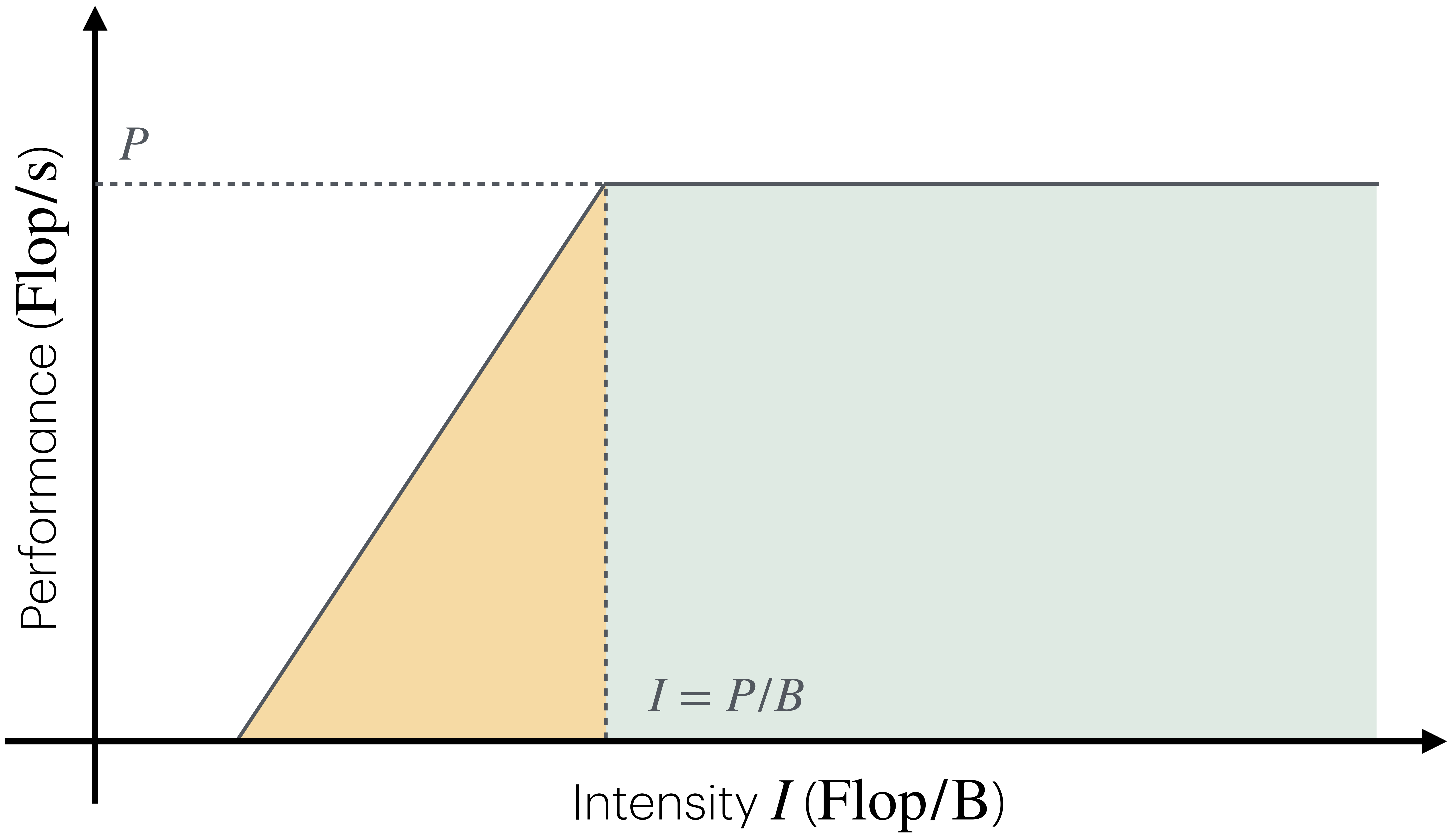
The “roofline”



- P: peak FP performance
- B: peak bandwidth

The roofline model

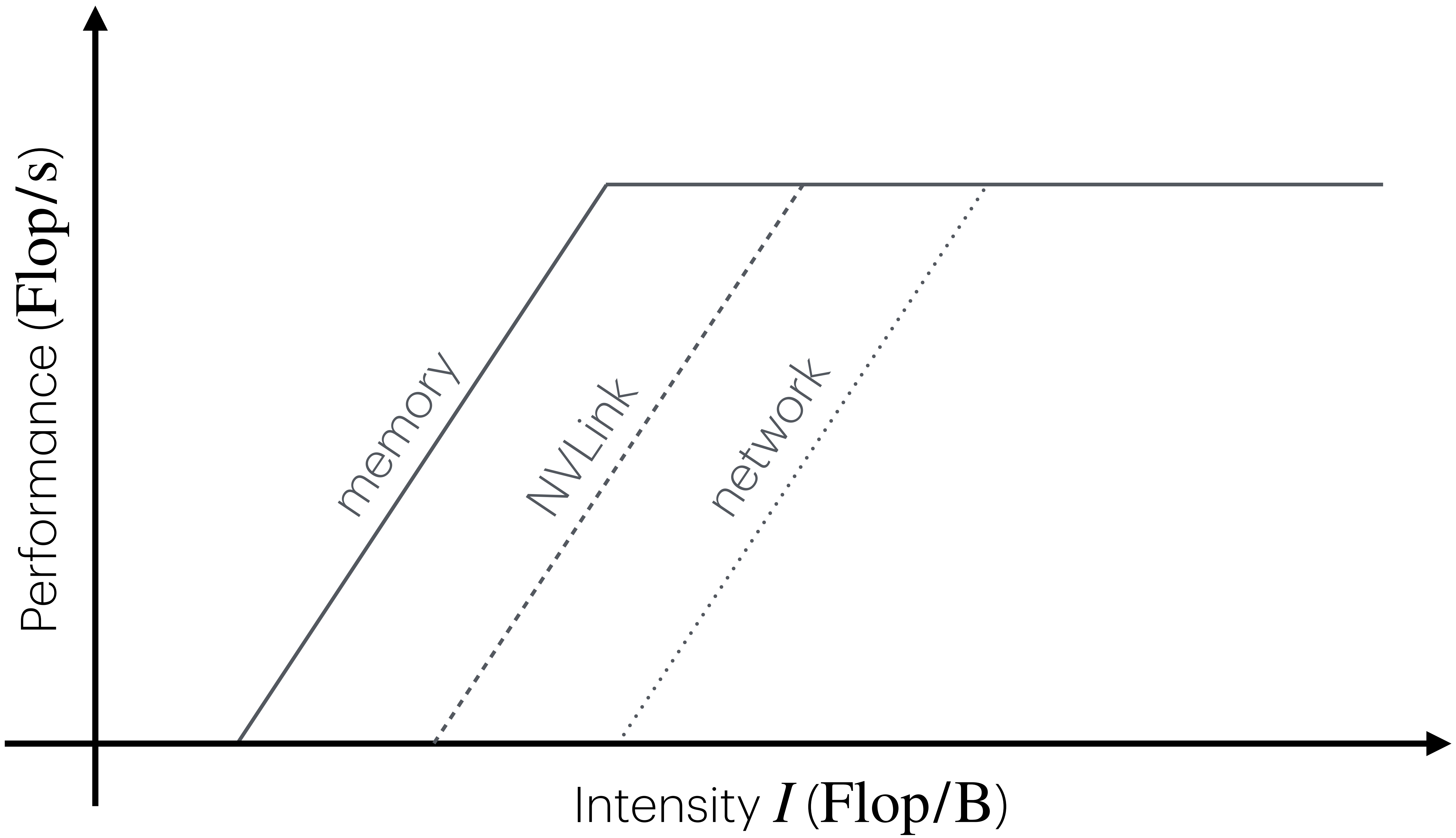
The "roofline"



- P: peak FP performance
- B: peak bandwidth
- : bandwidth-bound
- : compute-bound

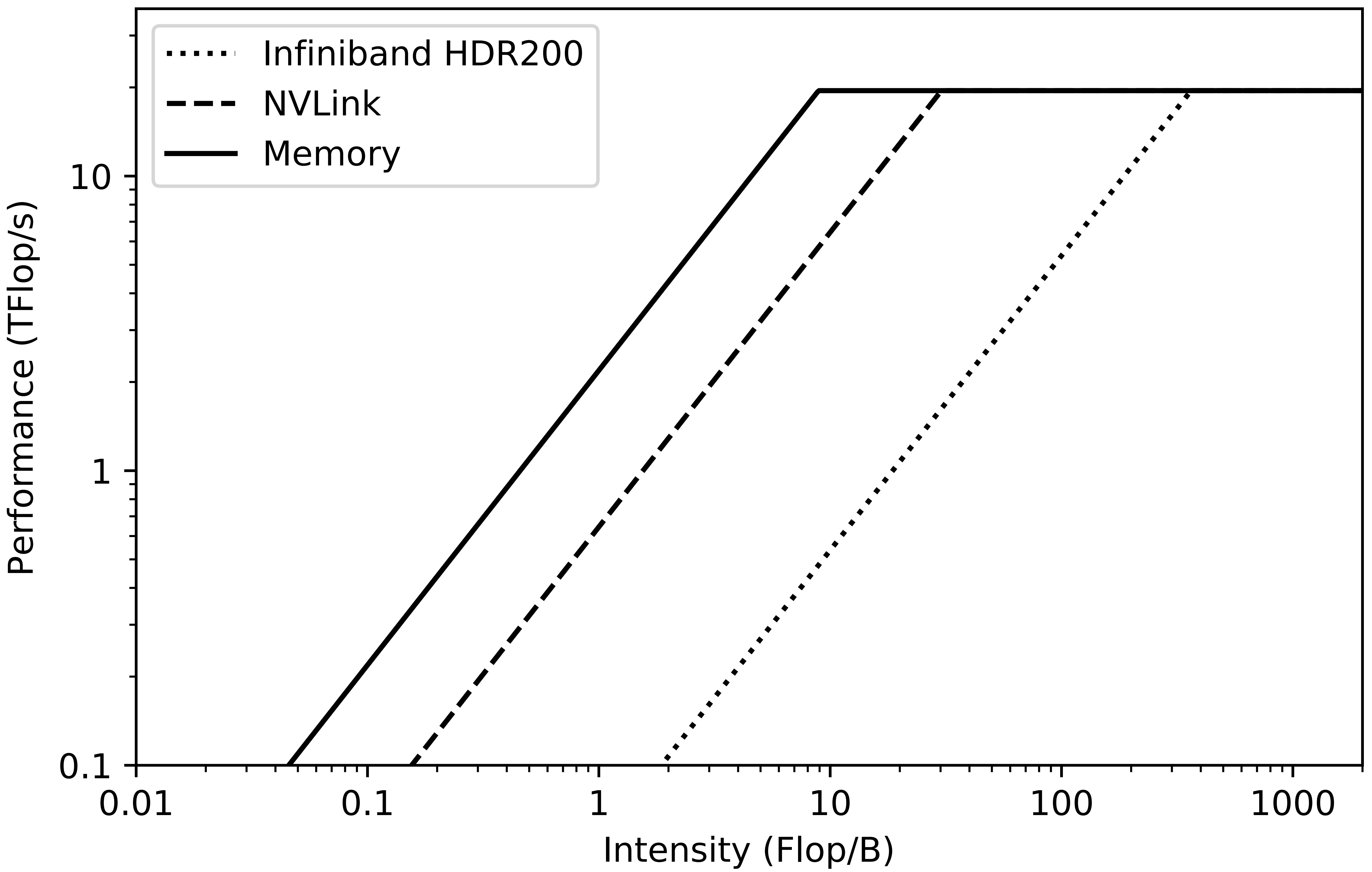
The roofline model

Multiple bandwidths



The roofline model

Example: NVIDIA A100-80 FP32



Top500 intensities

HPL & HPCG

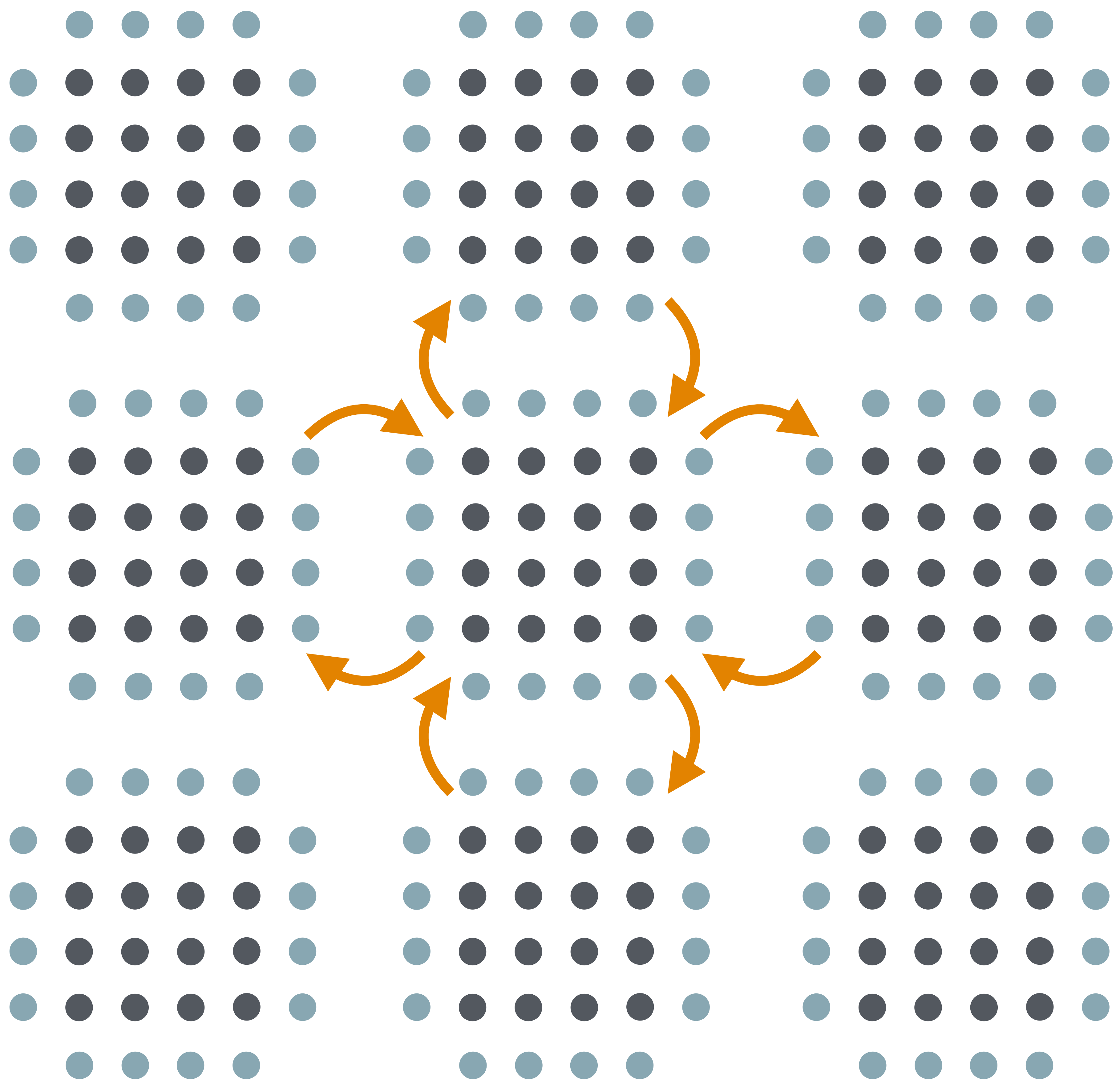
- HPL: variable intensity but $> 1000\ \mathrm{Flop/B}$ on contemporary systems
  Deep compute-bound regime, expect peak performance P

- HPCG: intensity $< 0.25\ \mathrm{Flop/B}$ typically estimated at $I_{\mathrm{HPCG}} = 0.1\ \mathrm{Flop/B}$
  Deep bandwidth-bound regime, expected performance $P_{\mathrm{HPCG}} = I_{\mathrm{HPCG}}\, B$

- What about medium-intensity benchmarks?

Intensities of the Dirac operator

Halo exchange



- Local d-dimensional lattice N^d
- Sparse matrix F $\mathrm{Flop/site}$
- Interior access: $\sim N^d$ read & write from local memory
- Exterior access: $\sim 2dN^{d-1}$ read & write between MPI processes
- HPCG: $d = 3$

  Lattice QCD: mainly $d = 4$

Intensities for the Dirac operator

Interior/exterior intensities

- Dirac-Wilson operator: $1344\ \text{Flop/site}$ (for $N_c = 3$)
- 12 complex numbers per site for spinors (interior)
  6 complex numbers per site for half-spinors in halos
- FP32 intensity for interior access $I_{\text{int}} = 7\ \text{Flop/B}$
- FP32 intensity for exterior access $I_{\text{ext}} = 1.75 \times N\ \text{Flop/B}$
- **More data to read in the interior, but much slower access for the surfaces**
- **For large jobs exterior dominates, so I_{ext} is the important number**

Dirac operator benchmark projections

- Benchmark projection $P_{\text{Wilson}} = 1.75 \times NB_{\text{net}}\ \text{Flop/s}$
- B_{net} is the peak bidirectional network bandwidth in B/s
- Assuming dominant exterior communication through network
- As long as bandwidth-bound regime is satisfied ($I_{\text{ext}} < P/B_{\text{net}}$) **performances approximately independent on GPU model**
- **Example:** NVIDIA A100-80, HDR200 network, $N = 32$: $P_{\text{Wilson}} = 3\ \text{TFlop/s}$

Benchmark results

In collaboration with Ryan Hill
Computing resources kindly provided by Maxwell T Hansen and Antonio Rago

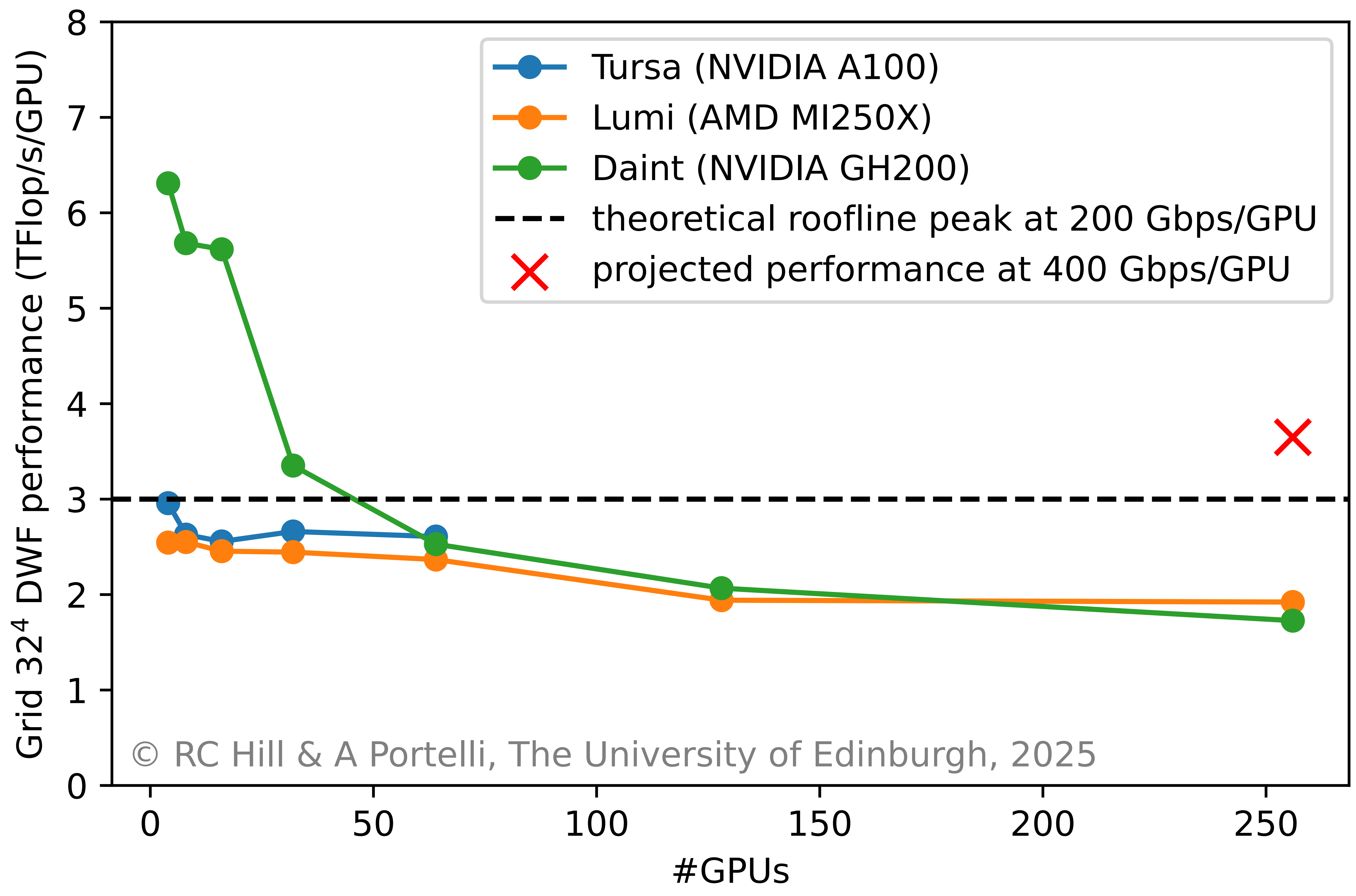
Context

UKRI Living Benchmark

- UKRI Living Benchmark (https://ukri-bench.github.io)
- Centralised **UK benchmark suite**
- Future **£750M system at University of Edinburgh**
- Grid benchmark (still in development)
  https://github.com/aportelli/grid-benchmark
- Forked from `Benchmark_ITT` in the Grid library

Benchmark results

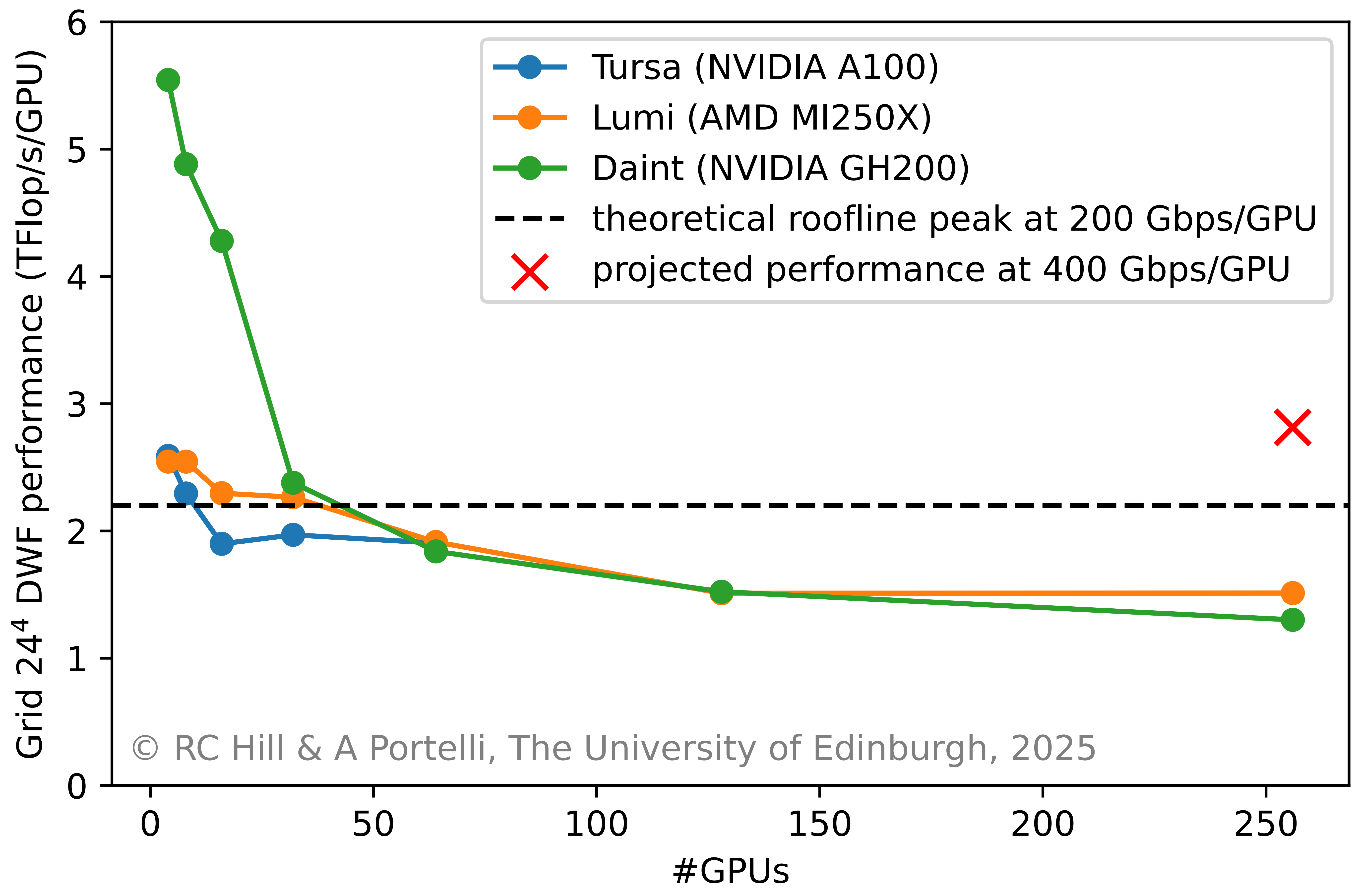
$N = 32$ scaling



- DWF operator better for bandwidth saturation
- Same intensity than Wilson with local 5th dimension
- Around 60% of roofline peak
- **Asymptotic independence from GPU model visible**

Benchmark results

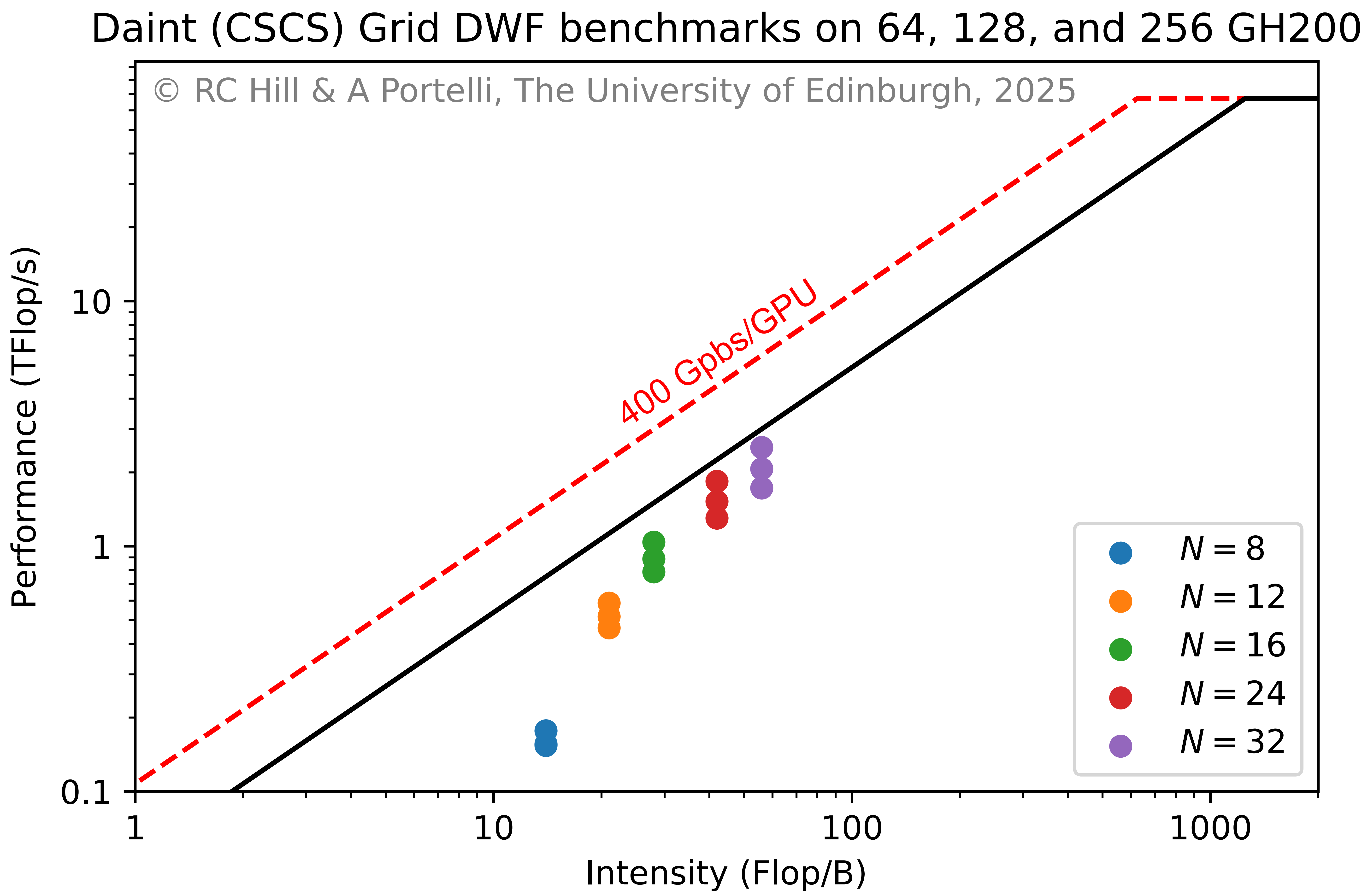
$N = 24$ scaling



- DWF operator better for bandwidth saturation
- Same intensity than Wilson with local 5th dimension
- Around 60% of roofline peak
- **Asymptotic independence from GPU model visible**

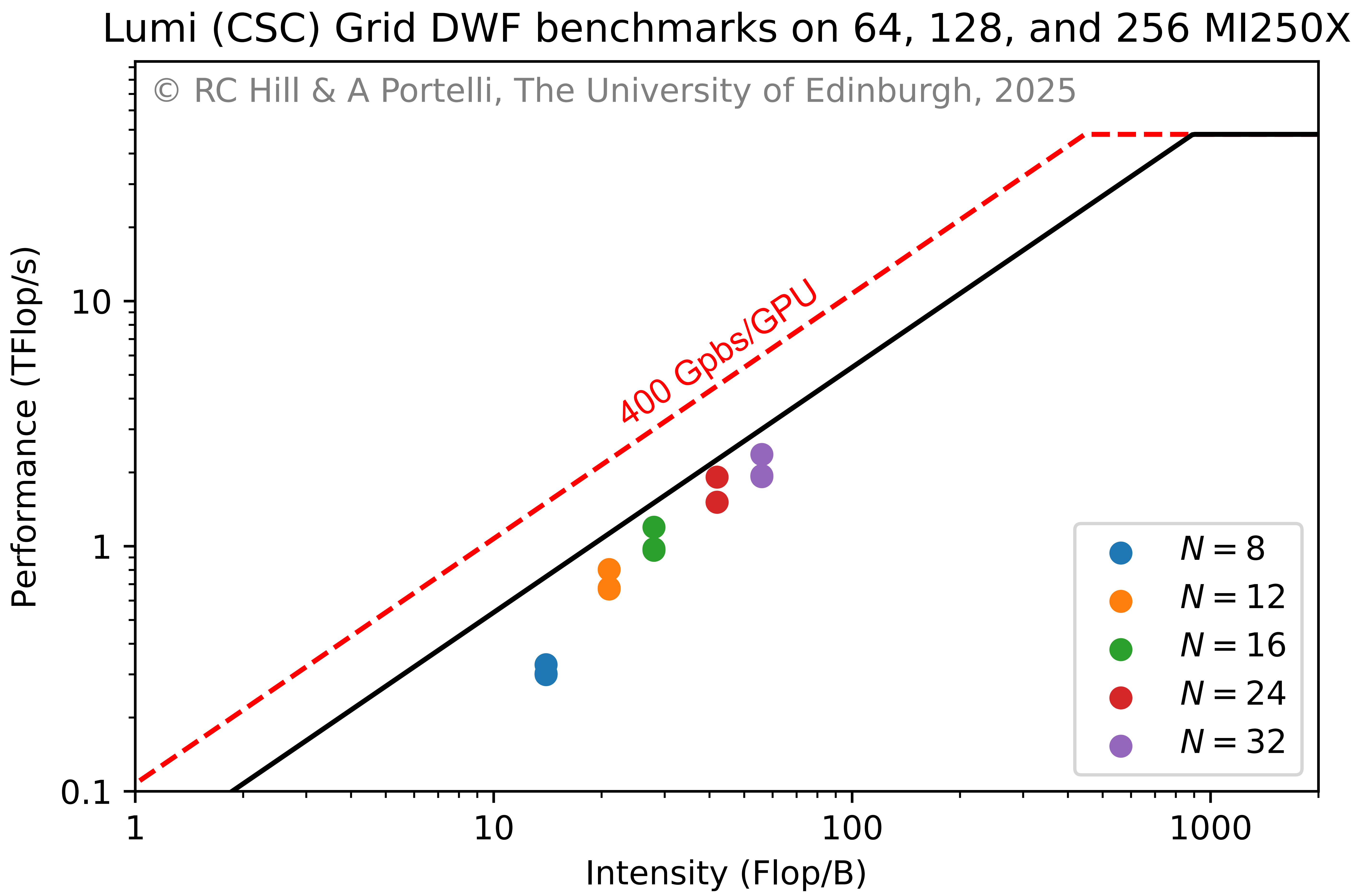
Benchmark results

Daint roofline view (NVIDIA GH200)



Benchmark results

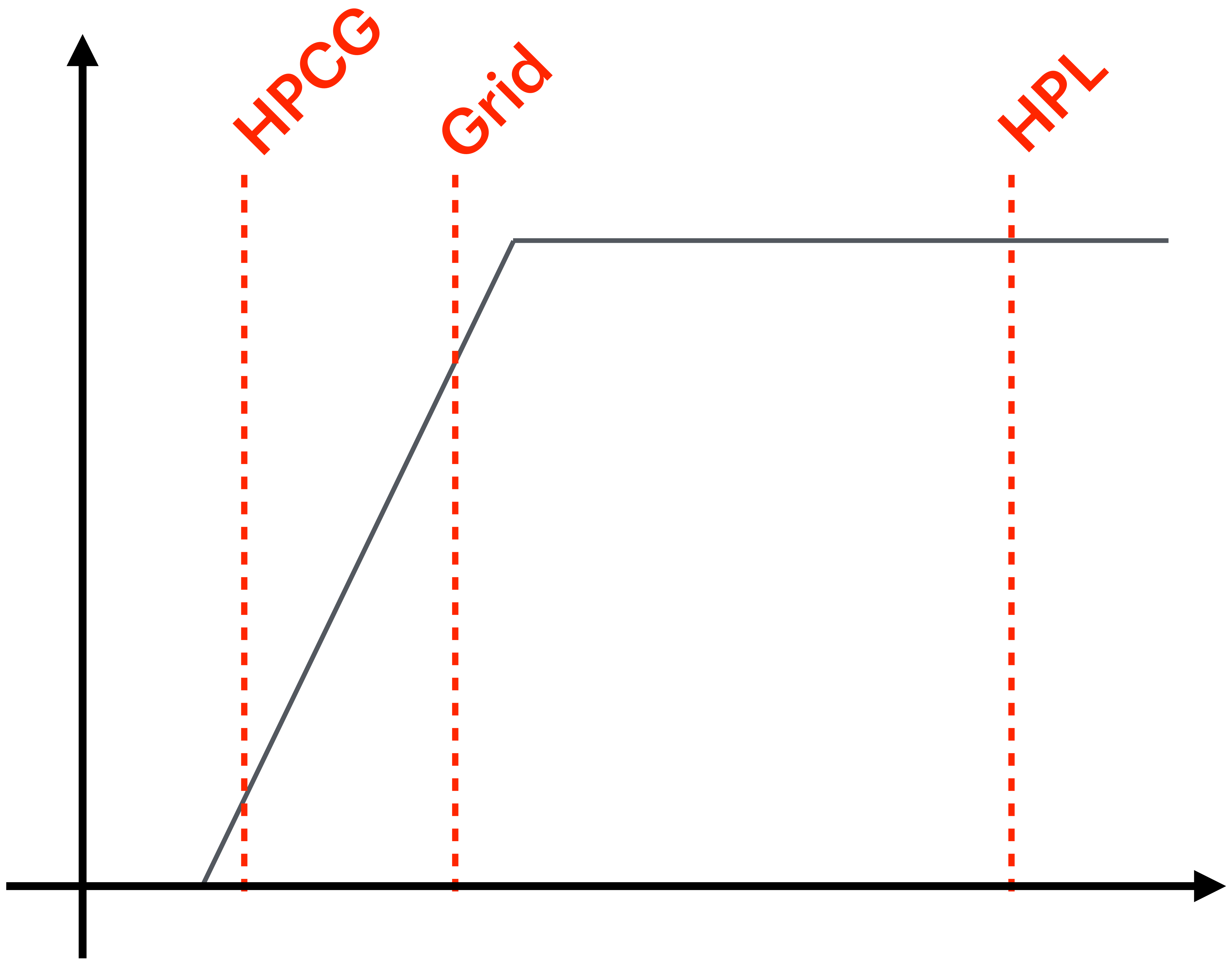
Lumi roofline view (AMD MI250X)



Performance model for exascale system design

Multi-intensity constraint

- Relevant dimensions of the roofline model can be constrained with **three benchmarks**:
  - **HPL** for high-intensity compute-bound
  - **Grid** for medium-intensity network-bound
  - **HPCG** for low-intensity memory-bound



HPL & HPCG projections

- The top 100 supercomputers achieve in average **70% of their peak for HPL** (source: Top500 June 2025 data https://top500.org/)
- Top500 systems reasonably well described by **memory-bound roofline peak for HPCG**
  - Fugaku: 158,976 nodes with 1 TB/s/node memory
    Roofline peak: 17 PFlop/s — Top 500: 16 PFlop/s (95% of peak)
  - El Capitan: 44,544 GPUs with 5.3 TB/s/GPU memory
    Roofline peak: 25 PFlop/s — Top 500: 17.4 PFlop/s (70% of peak)

Full projections

- Arbitrary system n computing units (GPU, CPU, etc...)
- Per unit: P compute peak, B_{mem} memory BW peak, B_{net} network BW peak
- **Performance projections at 70% of peak roofline performances:**
  - HPL: $P_{\mathrm{HPL}} = 0.7 \times nP$
  - Grid (FP32 DWF at $N = 32$): $P_{\mathrm{Grid}} = 0.7 \times 56nB_{\mathrm{net}}$
  - HPCG: $P_{\mathrm{HPCG}} = 0.7 \times 0.1 \times nB_{\mathrm{mem}}$

Conclusion & outlook

- Top500 benchmarks (HPL & HPCG) do not constrain the fabric on modern systems
- The Dirac-Wilson operator provide a strong medium-intensity network bound benchmark
- The roofline model describes reasonably well Grid performances for large local lattices
- The roofline provides simple projections for HPL, HPCG & Grid for system design
- **High network bandwidth (> 400Gbps per GPU) is absolutely critical for lattice QCD on contemporary GPU architecture**

ご清聴ありがとうございました！

宮島2025年11月15日