Information Roofline

A Model for the Era of Datatype Optimization and Emulation

Max Hawkins and Rich Vuduc
Georgia Institute of Technology

Measure Information. Not flop/s. Not bit/s.

1 Gallon Jug
Water content: ???

1 Gallon Jug
1 Gallon of Water

1 Gallon Jug
1/2 Gallon of Water

64 Physical Bits
Info content: ???

64 Physical Bits
64 Bits of Info

64 Physical Bits
55 Bits of Info

A bit is a container of information, but it's not necessarily full.

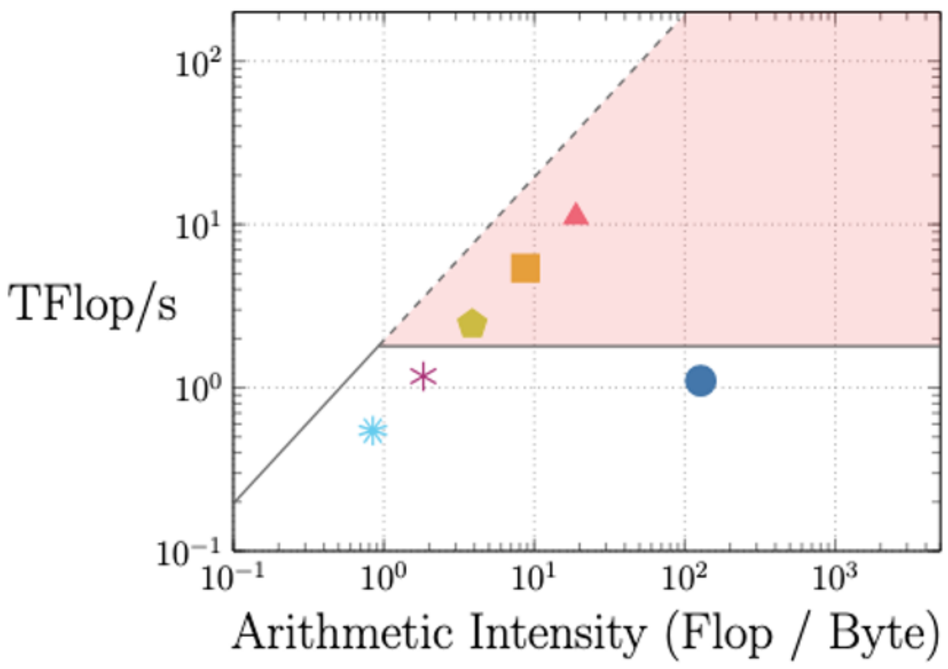
- A PHYSICAL bit != an INFORMATIONAL bit
- Examples:
  - Sign bit on a datatype storing unsigned data
  - Exponent bits of FP64
- We care about these inefficiencies now!
  - Datatype optimization, mixed precision, and emulation

Current Performance Models are Strained

How do you place/compare these on a traditional roofline model?

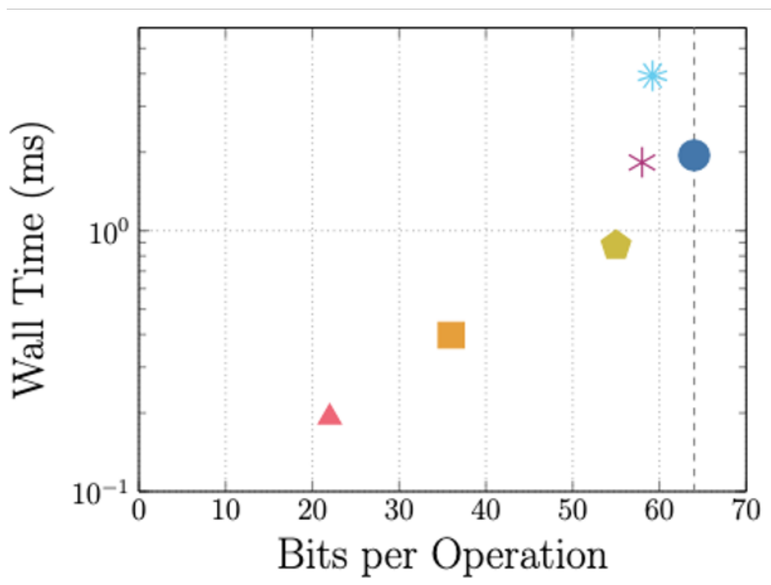
- Mixed-precision - FP8 vs FP64 vs Int1
  - Ginkgo: Remove bit-width from the x-axis: Values/s
- Emulation - Ozaki scheme 1 with 3 slices vs 18
- Noise - Lossless vs 1% error vs 90% error
- Ternary hardware - trits vs bits

The homogeneous 'op' treats everything the same. The physical bit assumes every bit transferred or computed on is needed.



Not all emulated operations are equal.

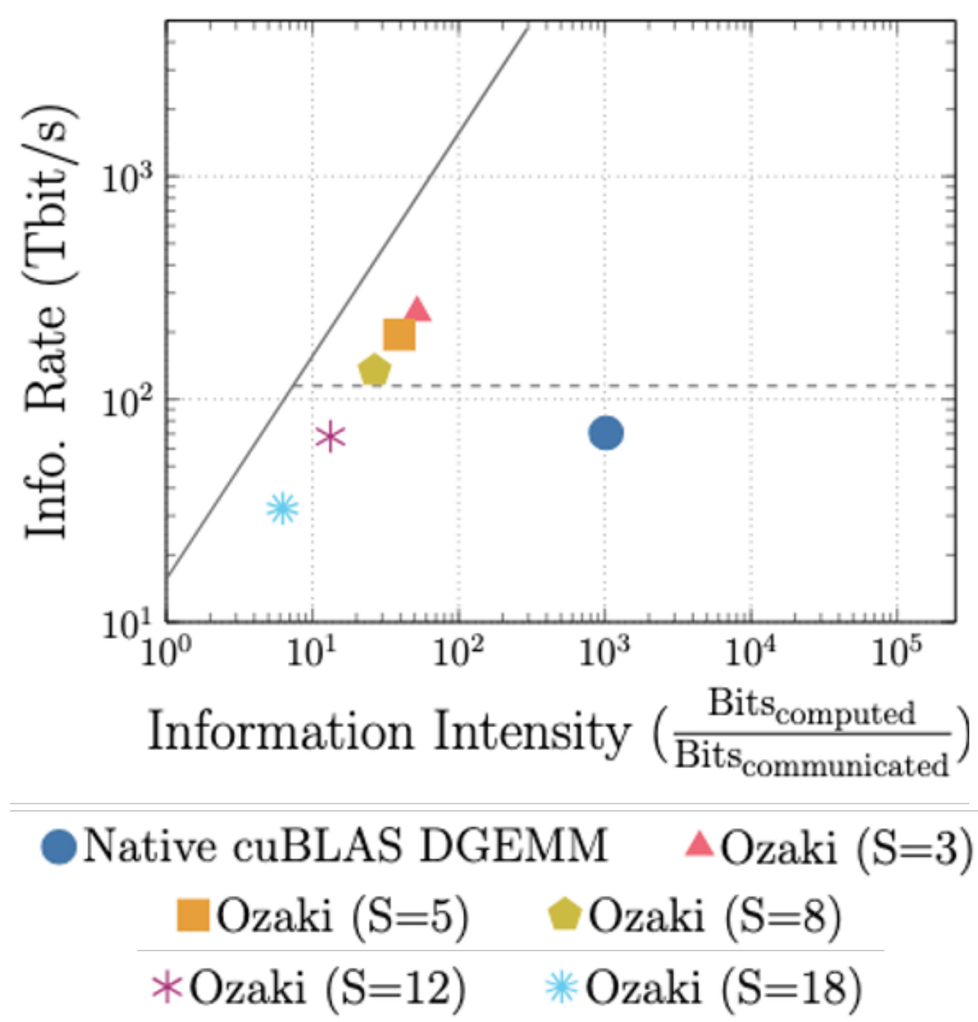
- Ozaki emulation builds a large integer (with varying slices, S) to map onto a float with piecewise cost.
  - Mantissa: linear bit cost. Exponent: exponential.



Information Roofline

- A von Neumann machine is two info channels
- X-axis: Info. rate of communication channel
- Y-axis: Info. Rate of computation channel
- Info Intensity: Unitless ratio of these info rates
- Roofline inequality simply states one of these channels limits the overall info processing rate: like water in pipes.

All plots show GEMM performance on an RTX 6000 Pro Blackwell for N=M=K=1024. ozIMMU emulation was used [1].



Measuring Information: Shannons vs Bits

- Quantifying entropy and information is weird.
- Bits: Historical precedence and common usage
  - Overloaded with physical bits, precision, range, etc
- Shannons: The technical quantity
  - Spec: IEC 80000-13:2025

Is this a hill worth dying on?

1. Ootomo et al. "DGEMM on integer matrix multiplication unit", 2024