TIMO SCHNEIDER <TIMOS@INF.ETHZ.CH>

DPHPC Recitation Session
Roofline Model

Last week:

- Amdahl's law
- Work/Depth model
- PRAM

- What is the work/depth of a matrix multiplication?
- Develop a PRAM algorithm for MM
- What kind of PRAM do we need for that?

Little's law

- In a queing system:
  - Latency (alpha)
  - Arrival rate (beta)
  - Waiting items (N)

  alpha * beta = N

  We can apply this model to memory!

  Latency * Throughput = Concurrency

Operational Intensity

- Operational Intensity I = #flops / #bytes
  - Can be meassured using perf. Counters
  - Can be modeled from the algorithm

  Example: Matrix multiplication

Roofline Model

- There are two fundamental limits:
  - Memory bandwidth (beta)
  - Computational bandwidth (pi)

Balance Principles

- An architecture is balanced
  - For a specific algorithm
  - On a specific input
- If pi/beta = W/Q
- If this holds, how would the roofline plot look like?

Balance Principles (Kung)

- An architecture is balanced
  - For a specific algorithm
  - On a specific input
- If pi/beta = W/Q
- If this holds, how would the roofline plot look like?
- What happens if pi/beta increases? Can we rebalence, i.e., matrix multiplication
- How realistic is an increase of pi, beta?