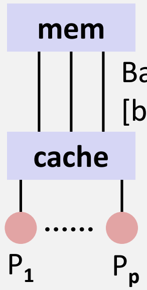
Roofline model (Williams et al. 2008)

Resources in a processor that bound performance:

- Peak performance [flops/cycle]
- Memory bandwidth [bytes/cycle]
- <others>

Platform model



Bandwidth β [bytes/cycle] ← carefully measured

- raw bandwidth from manual is unattainable (maybe 60% is)
- Stream benchmark may be conservative

Peak performance π [ops/cycle]

Algorithm model (n is the input size)

Operational intensity I(n) = W(n)/Q(n) =

$$\frac{\text{number of flops (cost)}}{\text{number of bytes transferred between memory and cache}} \quad \text{[ops/bytes]}$$

Q(n): assumes empty cache;
best measured with performance counters
Runtime T(n)
Performance P(n) = W(n)/T(n)

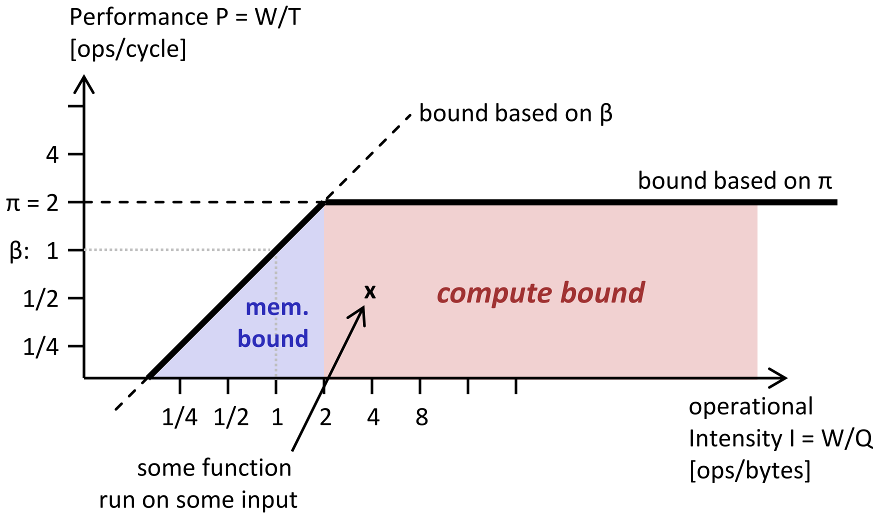
Notes

In general, Q and hence W/Q depend on the cache size m [bytes].
For some functions the optimal achievable I = W/Q is known:

FFT/sorting: Θ(log(m))
Matrix multiplication: Θ(sqrt(m))

Roofline model

Example: one core with π = 2 and β = 1 and no SSE
ops are double precision flops



Bounds

- Based on π: P ≤ π
- Based on β: P ≤ βI
- Reason: β ≥ Q/T = (W/T)/(W/Q) = P/I
- in log scale: $\log_2(P) \leq \log_2(\beta) + \log_2(I)$
- line with slope 1; P = β for I = 1

Variations

- vector instructions: peak bound goes up (e.g., 4 times for AVX)
- multiple cores: peak bound goes up (p times for p cores)
- program has uneven mix adds/mults: peak bound comes down (note: now this bound is program specific)
- accesses with little spatial locality: operational intensity decreases (because entire cache blocks are loaded)