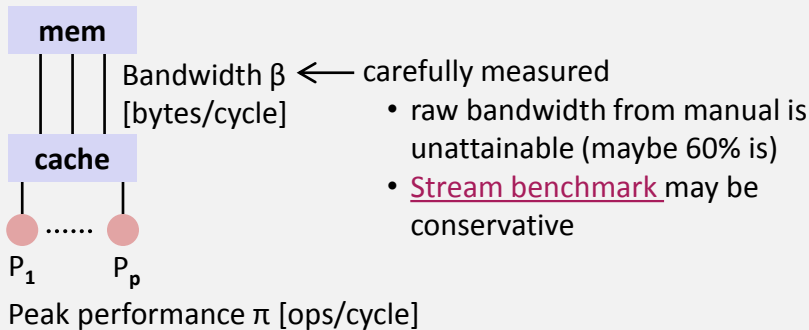


Roofline model ([Williams et al. 2008](#))

Resources in a processor that bound performance:

- peak performance [flops/cycle]
- memory bandwidth [bytes/cycle]
- <others>

Platform model



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

Notes

In general, Q and hence W/Q depend on the cache size m [bytes].

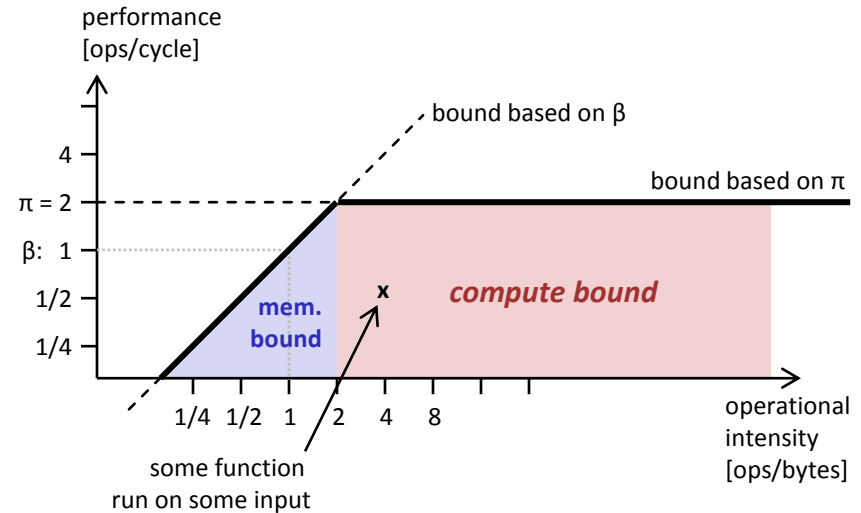
For some functions the optimal achievable W/Q is known:

FFT/sorting: $\Theta(\log(m))$

Matrix multiplication: $\Theta(\sqrt{m})$

Roofline model

Example: one core with $\pi = 2$ and $\beta = 1$ and no SSE
ops are double precision flops



Bound based on β?

- assume program as operational intensity of x ops/byte
- it can get only β bytes/cycle
- hence: performance = $y \leq \beta x$
- in log scale: $\log_2(y) \leq \log_2(\beta) + \log_2(x)$
- line with slope 1; $y = \beta$ for $x = 1$

More lines (bounds) can be added

- single core but with SSE (π goes up by $2x$)
- four cores (π goes up by $4x$)
- accesses with no spatial locality (β comes down by $8x$)