Roofline model (Williams et al. 2008)

Resources in a processor that bound performance:
- Peak performance [flops/cycle]
- Memory bandwidth [bytes/cycle]
- Others

Platform model

<table>
<thead>
<tr>
<th>mem</th>
<th>Bandwidth $\beta$ [bytes/cycle]</th>
</tr>
</thead>
<tbody>
<tr>
<td>cache</td>
<td></td>
</tr>
</tbody>
</table>

Peak performance $\pi$ [ops/cycle]

Algorithm model ($n$ is the input size)

Operational intensity $I(n) = \frac{W(n)}{Q(n)} = \frac{\text{number of flops (cost)}}{\text{number of bytes transferred between memory and cache}}$ [ops/bytes]

$Q(n)$: assumes empty cache; best measured with performance counters
Runtime $T(n)$
Performance $P(n) = \frac{W(n)}{T(n)}$ [ops/cycle]

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

Roofline model

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

Bounds
- Based on $\pi$: $P \leq \pi$
- Based on $\beta$: $P \leq \beta I$
- Reason: $\beta \geq Q/T = (W/T)/(W/Q) = P/I$
- in log scale: $\log_2(P) \leq \log_2(\beta) + \log_2(I)$
- line with slope 1; $P = \beta$ 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)