Roofline model \textit{(Williams et al. 2008)}

Resources in a processor that bound performance:
- peak performance [flops/cycle]
- memory bandwidth [bytes/cycle]
- <others>

Platform model

\begin{center}
\begin{tikzpicture}
\node[draw,rectangle] (mem) at (0,0) {mem};
\node[draw,rectangle] (cache) at (0,-1) {cache};
\draw[->] (mem) -- (cache);
\node[above] at (mem) {Bandwidth $\beta$ [bytes/cycle]};
\node[below] at (mem) {$P_1$ \text{ and } P_p$};
\node[below] at (cache) {Peak performance $\pi$ [ops/cycle]};
\end{tikzpicture}
\end{center}

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

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

Bound based on $\beta$?
- assume program as operational intensity of $x$ ops/byte
- it can get only $\beta$ 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 $x = 1$

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)