**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 $\beta$ [bytes/cycle]
- care fully measured
  - raw bandwidth from manual is unattainable (maybe 60% is)
  - Stream benchmark may be conservative

**Algorithm model (n is the input size)**

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

$Q(n)$: 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: $O(\log(m))$
    - Matrix multiplication: $O(\sqrt{m})$

**Bound based on $\beta$?**
- assume program as operational intensity of $x$ ops/byte
  - it can get only $\beta$ bytes/cycle
  - hence: performance $\leq \gamma \leq \beta x$
  - in log scale: $\log_2(\gamma) \leq \log_2(\beta) + \log_2(x)$
  - line with slope $1$: $\gamma = \beta x$

**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)
Roofline Measurements

- Tool developed in our group
  (G. Ofenbeck, R. Steinmann, V. Caparros-Cabezas, D. Spampinato)
  [http://www.spiral.net/software/roofline.html](http://www.spiral.net/software/roofline.html)
- Example plots follow
- Get (non-asymptotic) bounds on I:
  - daxpy: \( y = \alpha x + y \)
  - dgemv: \( y = Ax + y \)
  - dgemm: \( C = AB + C \)
  - FFT

---

Roofline Measurements

Core i7 Sandy Bridge, 6 cores  
Code: Intel MKL, sequential  
Cold cache

---

What happens when we go to parallel code?
Roofline Measurements

Core i7 Sandy Bridge, 6 cores
Code: Intel MKL, parallel
Cold cache

What happens when we go to warm cache?

Roofline Measurements

Core i7 Sandy Bridge, 6 cores
Code: Intel MKL, sequential
Warm cache
**Roofline Measurements**

```
Core i7 Sandy Bridge, 6 cores
Code: Various MMM
Cold cache
```

**Summary**

- Roofline plots distinguish between memory and compute bound
- Can be used on paper
- Measurements difficult (performance counters) but doable
- Interesting insights: *use in your project!*

---

© Markus Püschel

Computer Science
References

- Samuel Williams, Andrew Waterman, David Patterson
  *Roofline: an insightful visual performance model for multicore architectures*
  Communications ACM 55(6): 121-130 (2012)

- Georg Ofenbeck, Ruedi Steinmann, Victoria Caparros, Daniele G. Spampinato and Markus Püschel
  *Applying the Roofline Model*

- Victoria Caparros and Markus Püschel
  *Extending the Roofline Model: Bottleneck Analysis with Microarchitectural Constraints*