Operational Intensity Again

- Definition: Given a program P, assume cold (empty) cache

\[ I(n) = \frac{W(n)}{Q(n)} \]

- Examples: Determine asymptotic bounds on I(n)
  - Vector sum: \( y = x + y \) \( \text{O}(1) \)
  - Matrix-vector product: \( y = Ax \) \( \text{O}(1) \)
  - Fast Fourier transform \( \text{O}(\log(n)) \)
  - Matrix-matrix product: \( C = AB + C \) \( \text{O}(n) \)
Compute/Memory Bound

- A function/piece of code is:
  - **Compute bound** if it has high operational intensity
  - **Memory bound** if it has low operational intensity

- The roofline model makes this more precise
- Blackboard

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)

- You can use it in your project
- 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

MMM: Try to guess the basic shapes
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!