Design of Parallel and High-Performance Computing

Fall 2014

Lecture: Roofline

Instructor: Torsten Hoefler & Markus Püschel

TA: Timo Schneider & Arnamoy Bhattacharyya

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Operational Intensity

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

$$\textit{Operational intensity: } I(n) = \frac{W(n)}{Q(n)}$$

W(n): #flops (input size n)

Q(n): #bytes transferred cache ⟷ memory (for input size n)

- **Examples: Determine asymptotic bounds on I(n)**
  - Vector sum: y = x + y — **O(1)**
  - Matrix-vector product: y = Ax — **O(1)**
  - Fast Fourier transform — **O(log(n))**
  - Matrix-matrix product: C = AB + C — **O(n)**

Example MVM: $y = Ax + y$

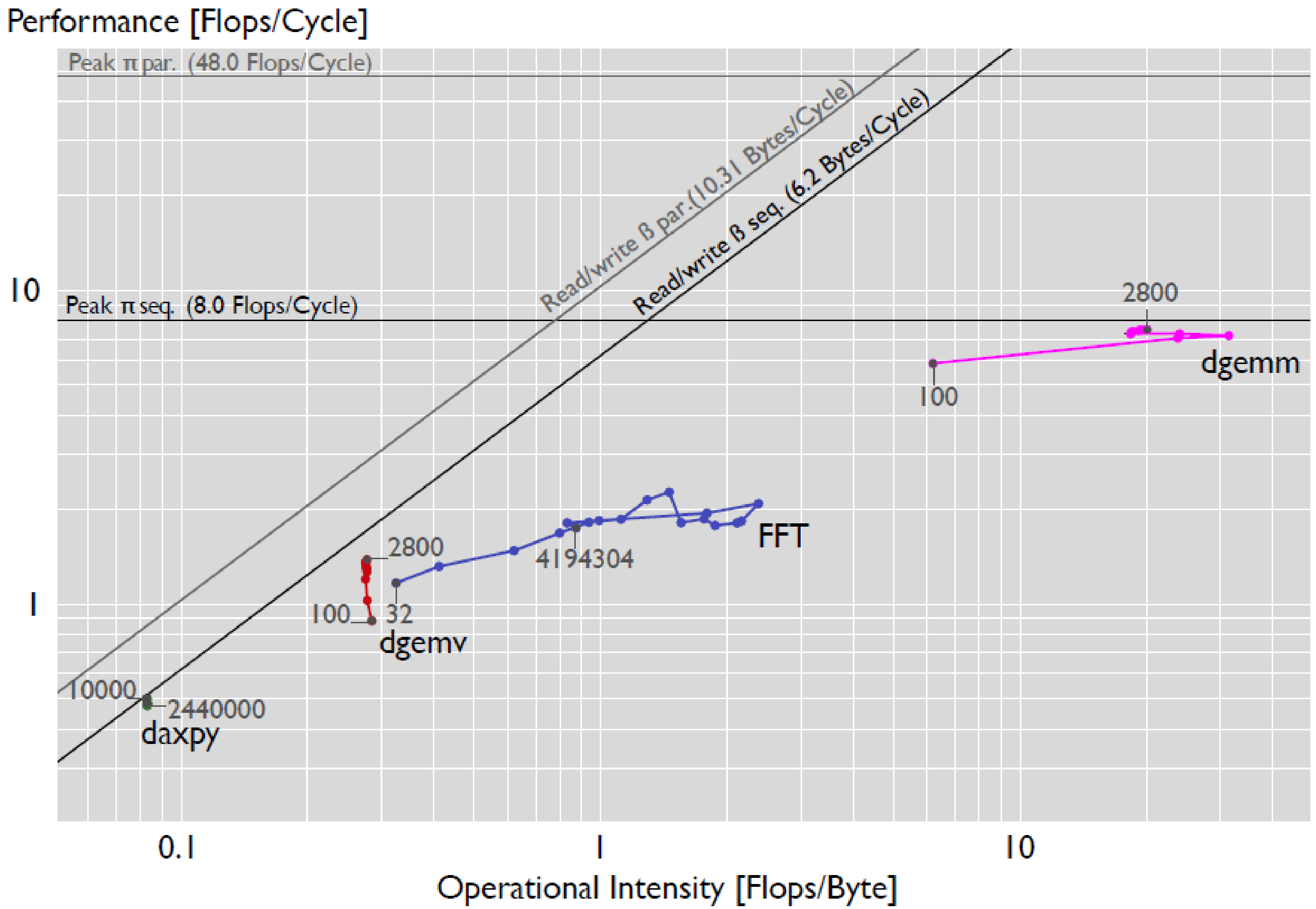
- **Number of flops?**
- **Number of compulsory misses (cold cache)?**
- **Upper bound on the operational intensity?**

Roofline Measurements

- **Tool developed in our group**
  (G. Ofenbeck, R. Steinmann, V. Caparros-Cabezas, D. Spampinato)
- **Example plots follow**
- **Get bounds on I:**
  - daxpy: $y = \alpha x+y$
  - dgemv: $y = Ax + y$
  - dgemm: $C = AB + C$
  - FFT

Roofline Measurements

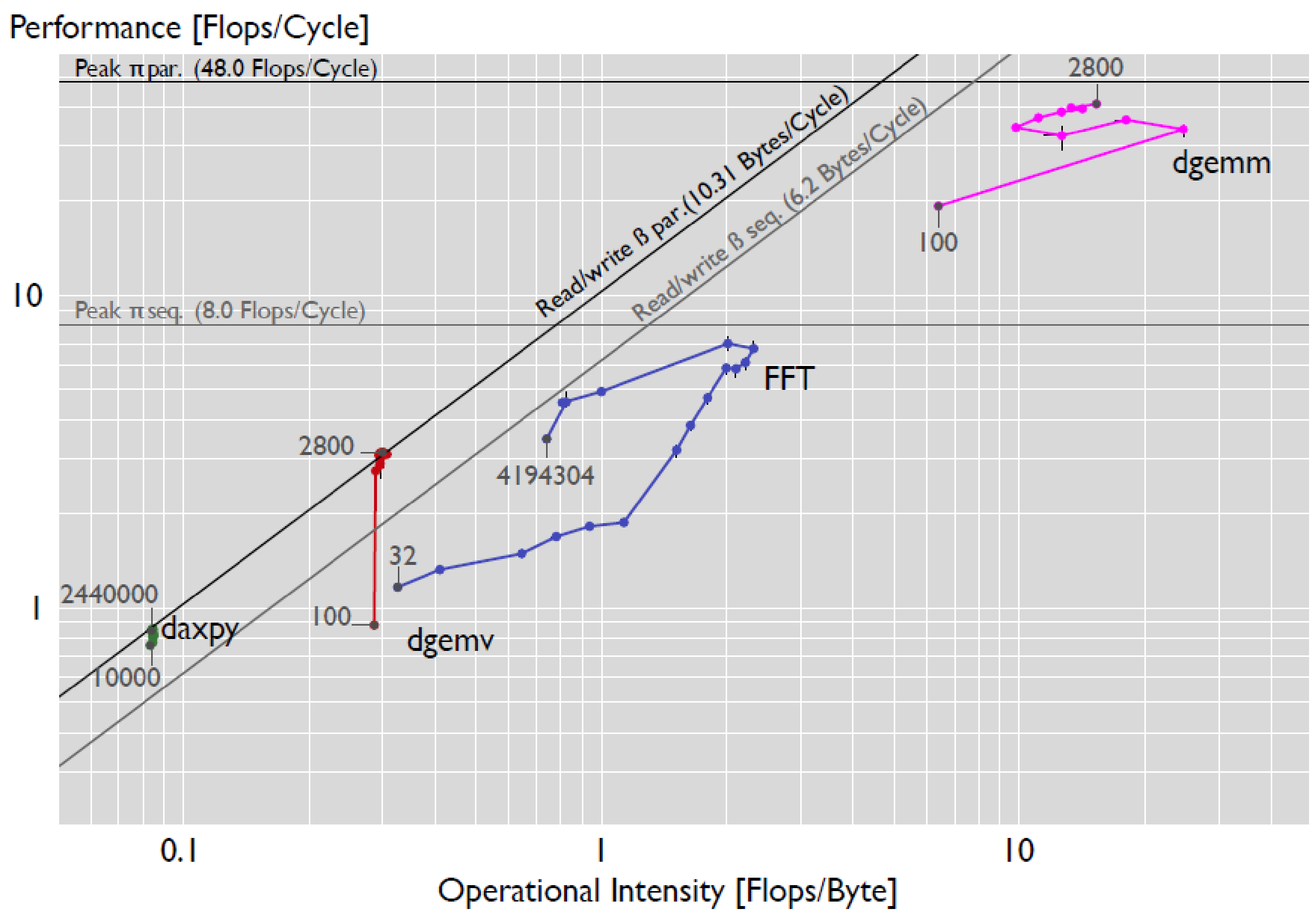
Core i7 Sandy Bridge, 6 cores
*Code: Intel MKL, **sequential***
Cold cache



What happens when we go to parallel code?

Roofline Measurements

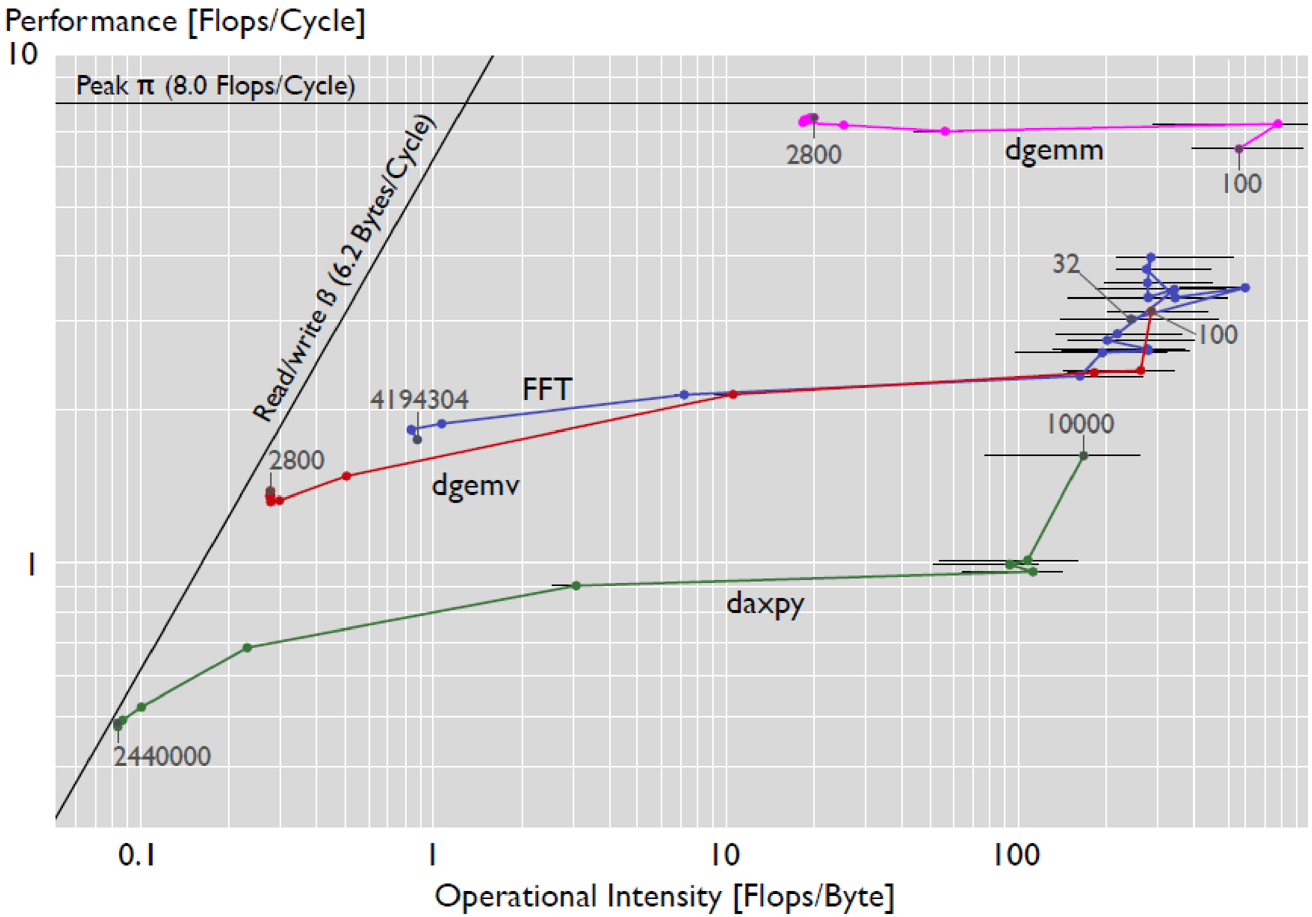
Core i7 Sandy Bridge, 6 cores
*Code: Intel MKL, **parallel***
Cold cache



What happens when we go to warm cache?

Roofline Measurements

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



Summary

- Roofline plots distinguish between memory and compute bound
- Can be used on paper
- Measurements difficult (performance counters) but doable