Michigan State University

CMSE/CSE 822 – Parallel Computing **Fall 2019**

Homework 2

Alexander Harnisch

---

1) Performance Modeling

a)

There is some missing information here and assumptions have to be made. However, it's not important since it only effects the numbers to plug in and get out, not the logic behind it.

The kernel performs 6 floating point operations (FLOP) in each iteration. Assuming no caching we have seven floating point number reading and one writing operation for each loop iterations. However, we can assume that the same variables are kept in the registers in each loop iterations, or at least in fast memory. Which means the entire y and z arrays both only have to be loaded once. Assuming single precision with 4 byte per float that translates to 12 byte of memory access and an arithmetic intensity I of

$$I = \frac{6\,\mathrm{FLOP}}{12\,\mathrm{byte}} = \frac{1}{2}\,\frac{\mathrm{FLOP}}{\mathrm{byte}}\,. \tag{1}$$

b)

In a simple roofline model for some I the critical peak performance π_{crit} is given by βI, where β is the peak memory bandwidth. So for $I = 0.5\,\frac{\mathrm{FLOP}}{\mathrm{byte}}$ from a) we get:

$$\pi_{\mathrm{crit}} = 30\,\frac{\mathrm{GB}}{\mathrm{s}} \cdot \frac{1}{2}\,\frac{\mathrm{FLOP}}{\mathrm{byte}} = 15\,\frac{\mathrm{GFLOP}}{\mathrm{s}}\,. \tag{2}$$

So in case the processor's peak performance is greater than 15 GFLOP/s the kernel is compute bound, otherwise memory bound.

c)

A simple roofline model plot is given by Figure 1. The performance for an arithmetic intensity of $I = 0.5\,\mathrm{FLOP/byte}$ is $15\,\mathrm{GFLOP/s}$.

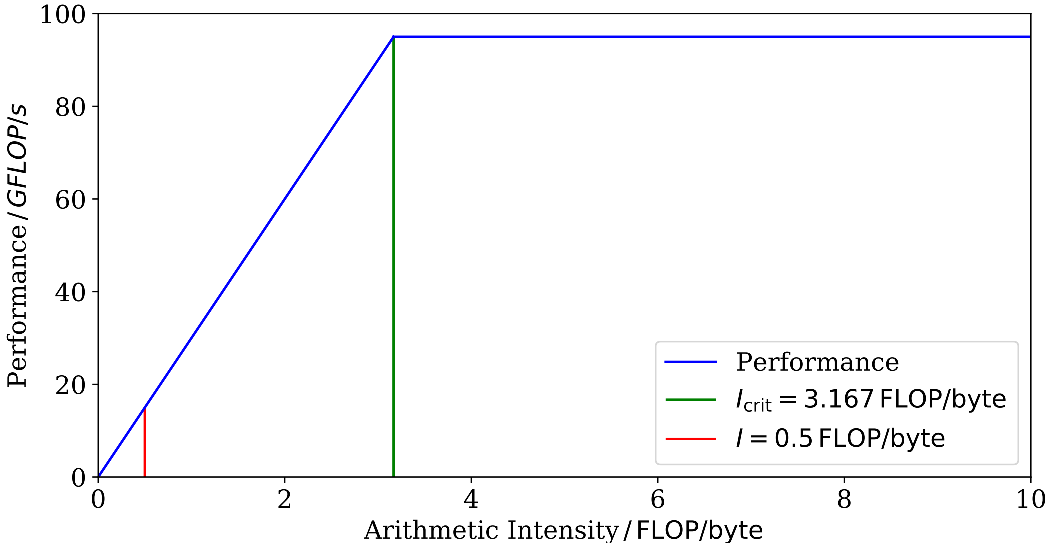

Figure 1: Simple roofline model.

2) Cache optimization: Matrix Vector Multiplication