P. Koumoutsakos, W. Petersen, P. Hadjidoukas
ETH Zentrum, CLT E 13
CH-8092 Zürich

Fall semester 2017

Set 2 - Roofline Model and Performance Measures

Issued: October 6, 2017
Hand in (optional): October 16, 2017 11:59pm

Understanding the characteristics of the platform on which performance tests are executed is of fundamental importance since it gives a context in which to read performance results. The primary objective of this exercise is to learn how to characterize computing hardware performance.

Peak Performance and System Memory Bandwidth

The number of executed floating point operations (FLOP) is a measure used to characterize the costs of scientific software, e.g. computational fluid dynamics codes, finite element analysis programs, computational chemistry and computational biology packages.

The peak floating point performance, hereafter simply called peak performance, is a measure of the quantity of FLOP that a machine can execute in a given amount of time. Typically, the peak floating point performance (PP) in FLOP/s can be computed as in the following:

$$PP\ [FLOP/s] = f\ [HZ = cycle/s] \times c\ [FLOP/cycle] \times v\ [-] \times n\ [-], \qquad (1)$$

where f is the core frequency in CPU cycles per second (given in Hz), c the number of FLOP executed in each cycle, v the SIMD width (in number of floats) and n the number of cores.

The exact values for the above features can be found at the hardware specifications or given by the system administrators. For example, for the Euler cluster we can find the specifications of the Intel Xeon E5-2680v3 here: `https://ark.intel.com/products/81908/`. Additional information is provided at the Euler wiki: `https://scicomp.ethz.ch/wiki/Euler`.

Typical floating point performance values are reported in GFLOP/s or TFLOP/s. Floating point performance is also used to assess the performance of a given algorithm implementation (`http://en.wikipedia.org/wiki/FLOPS`).

The bandwidth of the system memory is a measure of the speed of data movement from the system to caches and vice-versa. As the problems considered here in fit on a single node, we are mostly interested in quantifying the DRAM bandwidth.

Given the specifications of the DRAM memory, the theoretical memory bandwidth (PB) in B/s can be computed as follows:

$$PB\ [B/s] = f_{DDR}\ [Hz = cycle/s] \times c\ [channel] \times w\ [bit/channel/cycle] \times 0.125\ [B/bit], \quad (2)$$

where f_{DDR} is the DDR clock rate, c the number of memory channels and w the bits moved through a channel per cycle (typically 64 bits). Typical bandwidths are reported in MB/s, GB/s or TB/s (`http://en.wikipedia.org/wiki/Memory_bandwidth`).

Memories based on the DDR (Double Data Rate) technology, such as DDR-SDRAM, DDR2-SDRAM, and DDR3-SDRAM[1], transfer two data per clock cycle. As a result, they achieve double the transfer rate compared to traditional memory technologies (such as the original SDRAM) running at the same clock rate. Because of that, DDR-based memories are usually labeled with double their real clock rate. For example, DDR3-1866 memories actually work at 933 MHz transferring two data per clock cycle, and thus are labeled as being a "1,866 MHz" device, even though the clock signal does not really work at 1.866 GHz.

Question 1: Operational Intensity

The purpose of a compute architecture is to perform a certain operation on data. This requires a mechanism to send data along the memory hierarchy[2] to the execution units such that they can perform the operation(s). Technological advances in the past dictate that the rate π at which operations can be executed doubles roughly every 18 months (Moore's Law), while the rate β at which data can be transported doubles roughly ever 36 months. The operational intensity is a measure that relates the amount of work W (operations) to the number of bytes Q (data) that need to be transferred and is defined as

$$I = \frac{W}{Q} \quad [\text{ops/byte}]. \tag{3}$$

The operational intensity I is used to characterize a code and reveals information about the expected utilization of the architecture's execution units and memory bandwidth. In practice, operations are defined by FLOPs and memory transfer is defined by access from DRAM to the closest processor cache.

a) Determine the asymptotic bounds on the operational intensity $I(n)$ for the following matrix/vector operations, where n is the dimension of the vector. State your assumptions.
   1. DAXPY: $y = \alpha x + y \quad \alpha \in \mathbb{R}; x, y \in \mathbb{R}^n$
   2. SGEMV: $y = Ax + y \quad x, y \in \mathbb{R}^n; A \in \mathbb{R}^{n \times n}$
   3. DGEMM: $C = AB + C \quad A, B, C \in \mathbb{R}^{n \times n}$

b) Consider the 1D diffusion equation in a periodic domain with length L

$$u_t(x,t) = \alpha u_{xx}(x,t), \tag{4}$$

$$u_0(x) = u(x,0) = \sin\left(\frac{2\pi}{L}x\right), \tag{5}$$

where $0 \leq x < L$, $t > 0$ and $\alpha > 0$ is the diffusion coefficient. You are asked to solve this problem numerically by using a second order centered finite difference scheme and the explicit Euler method to advance in time. The domain is discretized with N uniformly spaced grid points. Discretization of Equation (4) leads to

$$u_i^{n+1} = u_i^n + \frac{\Delta t \alpha}{\Delta x^2}\left(u_{i-1}^n - 2u_i^n + u_{i+1}^n\right), \tag{6}$$

where $u_i^n \approx u(x_i, t^n)$ is the approximate solution with $t^{n+1} = t^n + \Delta t$ and $x_i = i\Delta x$. The constant time step size and uniform grid spacing are denoted by Δt and Δx, respectively. Determine the operational intensity $I(N)$ for the scheme given in Equation (6). Can you identify a possible hardware bottleneck for this scheme? State your assumptions.

[1] `http://en.wikipedia.org/wiki/DDR3_SDRAM`

[2] `https://en.wikipedia.org/wiki/Memory_hierarchy`

Question 2: Roofline Model

The roofline model[3] is a simple visual tool that can be used to understand performance on a hardware architecture. It relates the nominal peak performance π and the nominal peak bandwidth β of a given hardware and introduces the concept of operational intensity, studied in the previous question.

a) According to the wiki of the Euler cluster, each of the two sockets on a node hosts an Intel Xeon E5-2680v3 12-core Haswell CPU capable of delivering 480 GFLOP/s each. In addition, the theoretical memory bandwidth for each of the two sockets can reach 68.3 GB/s. Try to justify the above numbers.

b) What is the relation of the memory bandwidth β and operational intensity I to hardware performance? The hardware specification indicates a nominal peak performance π. Find a relation for the hardware peak performance P_{peak} which depends on β and π. For the operational intensity I_B the machine is in balance, which means that the memory bandwidth β and nominal floating point performance π are optimally utilized. The operational intensity I_B allows to group a hardware into a memory bound region ($I < I_B$) and compute bound region ($I > I_B$), which will eventually dictate the amount of optimizations that can be undertaken. Find a relation for the operational intensity I_B.

c) Draw the roofline for a full NUMA node of Euler.

d) Write a small C/C++ code that implements Equation (6). Choose a suitable number of grid points N, where $L = 1000$ and $\alpha = 10^{-4}$. For stability reasons, you must choose your time step such that

$$\Delta t < \frac{\Delta x^2}{2\alpha}. \tag{7}$$

Benchmark the loop execution on one Euler node by using a similar approach that has been used in Question 2 of Exercise 1 in order to determine an average time for one update in time. Use the operational intensity calculated in Question 1b and indicate the measured performance in the roofline plot. Is your implementation memory bound or compute bound? What is the maximum performance that your code can reach according to the model? Explain issues in case you are not able to fully utilize the hardware.

e) (Optional). The Swiss National Supercomputing Centre (CSCS) in Lugano is home to the Piz Daint supercomputer. The machine reaches a performance of 19.6 PFLOP/s based on the Linpack benchmark and is currently the third fastest computer in the world (`https://www.top500.org/lists/2017/06/`, as of June 2017). Piz Daint is composed of Cray XC50 compute nodes which are hybrid nodes with GPU accelerators. Go to `http://www.cscs.ch/computers/piz_daint/index.html` and find out the specifications of the GPUs. Based on the numbers you have researched, draw the roofline for the GPU. Compare this roofline with the one you have drawn in part 2c. Describe the differences you observe in terms of how it would impact you as application developer.

[3] Williams et al, 2009: the original paper can be found on the course website.

Question 3: Performance Measures

Sometimes the nominal peak of a platform is not sufficient as you might want a more realistic upper bound on what performance you can effectively reach with a program. The objective of this exercise is to get as high as possible with your current knowledge.

a) Write a small C/C++ benchmark code to measure the peak single precision performance of a given platform. Report the method you used to measure the peak performance. Try your code on a compute node of the Euler cluster. Can you reach the theoretical peak performance reported before? Please explain your observations.