HPC-LEAP School Jülich: Exercises Performance Modelling

Dirk Pleiter

24.01.2016

1 Information Exchange Functions

1.1 Background

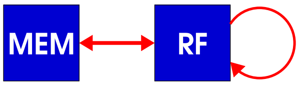
Consider the multiplication of dense matrices:

$$C_{i,j} \leftarrow \sum_{k=0}^{N-1} A_{i,k} \cdot B_{k,j} \quad (i,j = 0, \ldots, N-1) \tag{1}$$

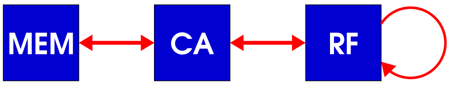
where the elements of A, B and C are double-precision floating-point numbers.

We consider the following abstract machine models:

Machine A Simple architecture comprising external memory (MEM) and a register file (RF), to which an arithmetic pipeline is attached:



Machine B Similar architecture but now including a data cache (CA):



You may assume that the RF is large enough to hold intermediate results.

1.2 Tasks

Determine for the given computational task and abstract machine model the following quantities:

1. Information exchange function describing the exchange of data between MEM and RF (machine A) or CA (machine B): $I_{\text{mem}}(N)$ (in units of Byte).

2. Information exchange function describing the amount of information moved through the arithmetic pipeline connecting the register file with itself: $I_{\text{fp}}(N)$ (in units of Flop).

3. Determine the arithmetic intensity of this computational task for both machines.

1.3 Solution

1. For each (i, j) we have to load N elements of A and B, each, and store one element of C:

   $$I_{\mathrm{mem}}(N) = N^2(2N + 1)8\,\mathrm{Byte} = (16N^3 + 8N^2)\,\mathrm{Byte}.$$

   If we could keep all elements of A and B in the processor we have to load these only once and thus:

   $$I_{\mathrm{mem}}(N) = (2N^2 + N^2)8\,\mathrm{Byte} = 24N^2\,\mathrm{Byte}.$$

2. For each (i, j) we have to perform N multiplications and $N - 1$ additions:

   $$I_{\mathrm{fp}}(N) = N^2(2N - 1)\,\mathrm{Flop} = (2N^3 - N^2)\,\mathrm{Flop}.$$

3. The Arithmetic Intensity is given by

   $$\mathrm{AI} = \frac{I_{\mathrm{fp}}(N)}{I_{\mathrm{mem}}(N)} = \frac{2N^3 - N^2}{16N^3 + 8N^2}\frac{\mathrm{Flop}}{\mathrm{Byte}} \simeq \frac{1}{8}\frac{\mathrm{Flop}}{\mathrm{Byte}}.$$

   If we again assume that all elements of A and B can be hold in the processor then

   $$\mathrm{AI} = \frac{2N^3 - N^2}{24N^2}\frac{\mathrm{Flop}}{\mathrm{Byte}} = \frac{2N - 1}{24}\frac{\mathrm{Flop}}{\mathrm{Byte}}.$$

2 Semi-Empirical Performance Modelling

2.1 Background

We consider the same computational task as in the previous exercise (dense matrix-matrix multiplication), but focus only on Machine B, which includes a cache.

2.2 Tasks

1. Formulate 2 semi-empirical performance models using (a) I_{fp} and (b) I_{mem} as determined in the previous task.

2. Collect performance numbers using the benchmark program that can be found here: `http://bit.ly/1SFeYrW`. The program measures time in units of base core clock cycles.

3. Compare the results with the performance models and argue about the observations.

4. Determine the model parameters and compare these with the capabilities of the hardware that was used.

It is recommended to keep N small, e.g. $8 \leq N \leq 120$.

2.3 Solution

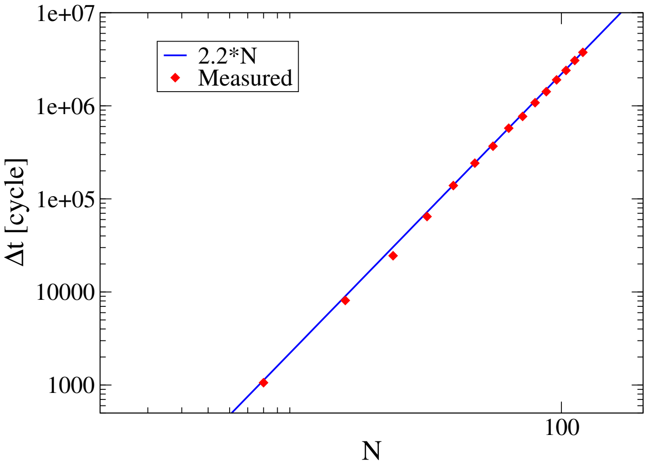
1. Performance models are

   $$\Delta t_a(N) = a_0 + a_1 I_{\mathrm{fp}}(N) = a_0 + a_1(2N^3 - N^2)$$

   and

   $$\Delta t_b(N) = b_0 + b_1 I_{\mathrm{mem}}(N) = b_0 + b_1(24N^2).$$

2. Runs executed on zam344 with an Intel Core i5-5200U processor:

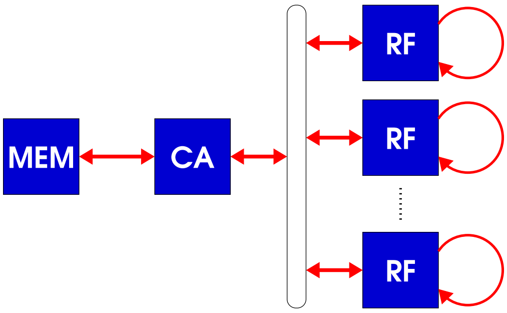


3. Comparison of data and model confirms that Δt_a is the better model. Which was to expected to be expected because
   - I_{fp} scales with N^3 and thus will dominate for large N.
   - The arithmetic intensity is quickly very large, i.e. time for memory accesses should become negligible.
4. The observed performance data can be parametrized by $\Delta t = 2.2N^3$cycles. Therefore $a_1^{-1} = 0.9091\,\text{Flop/cycle} = 2\,\text{GFlop/s}$. The used processor has a nominal clock speed of 2.20 GHz and can peform $2 \cdot 4$ double-precision FMA per cycle, i.e. has a peak performance of 35 GFlop/s per core.

3 BSP Model Application

3.1 Background

We consider the same computational task as in the previous exercises (dense matrix-matrix multiplication), but now consider parallelization on P processors cores using the Cannon algorithm. We assume a machine model with a sufficiently large cache, i.e. transfers from/to external memory MEM can be ignored:



We assume that the architecture can be described using BSP models with the following (application independent) parameters:

g	2 cycle
L	10,000 cycle

3.2 Tasks

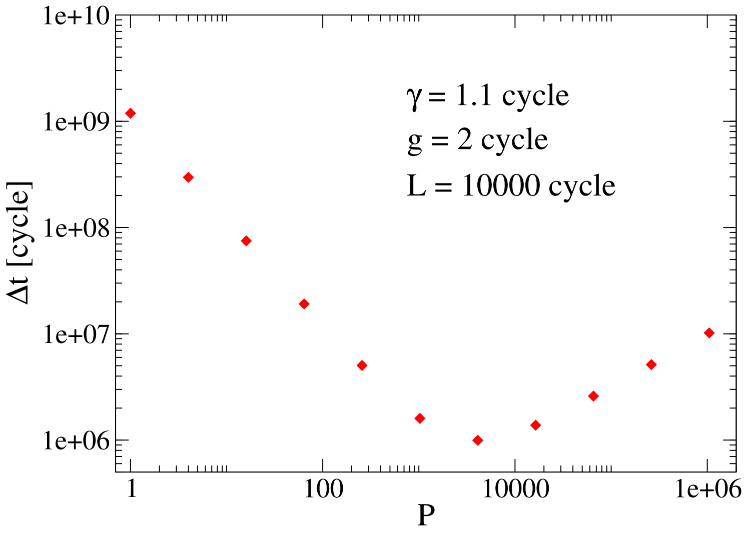
1. Use results from the previous exercise to construct a BSP model to estimate $\Delta t(P)$.
2. Determine the number of cores P for which $\Delta t(P)$ starts to increase when increasing P for $N = 1024$ (a graphical solution suffices).

3.3 Solution

In the performance modelling lecture the following formula was provided:

$$\Delta t = \gamma \frac{N^3}{P} + g\, 2\, \frac{N^2}{\sqrt{P}} + L\,\sqrt{P}$$

Graphical solution using results obtained on zam344 with an Intel Core i5-5200U processor:


γ = 1.1 cycle
g = 2 cycle
L = 10000 cycle
Δt [cycle]
1e+10
1e+09
1e+08
1e+07
1e+06
1e+06
P