Theoretical Peak Performance of a Computer

In order to assess how efficiently our program is running on a given computer, it is useful to know the theoretical peak performance of the computer (aka light speed of performance). By comparing the measured performance of the program, we will then know how many % of the theoretical peak we are achieving. For a numerical program in scientific computing, the peak performance is measured in terms of flop/s (floating-point operations per second), *i.e.*, number of floating-point (as opposed to integer) operations (like multiplication and addition) executed per second.

The peak flop/s of a computer is estimated as a product of (1) the number of multi-core processors in the computer, (2) number of cores (*i.e.*, light-weight processors that share resources like memory with the other cores within a processor) per processor, and (3) peak flop/s of each core. Each core's flop/s in turn is estimated as a product of its clock speed (*i.e.*, how many times the system clock ticks per second) and the number of operations executable per clock cycle. Most modern processors are equipped with a fused multiply-add (FMA) circuit, which can operate 1 multiplication and 1 addition operations per clock cycle. Furthermore, each multiply or add operation can be performed on vector registers, each holding multiple operands (**Fig. 1**). In scientific computing, we commonly consider operations on double-precision (or 64 bits) numbers. For example, a 512-bit vector register can hold 8 double-precision numbers.

Thus, the theoretical flop/s of a computer is computed as

$$Peak\_flop/s = (\#processors) \times (\#cores\_per\_processor) \times (clock\_speed\ [1/\mathrm{s}]) \times (2 \times \#FMA\_units) \times \frac{vector\_size\ [\mathrm{bits}]}{64},$$

where *#processors* is the number of processors that constitute a parallel computer, *#cores_per_processor* is the number of cores per multi-core processor, *clock_speed* is usually measured in GHz (or 10^9 cycles per second [1/s]), *#FMA_units* is the number of FMA units per core (factor 2 accounts for one multiply and one add operations), and the last term is the number of double-precision operands held in each vector register.

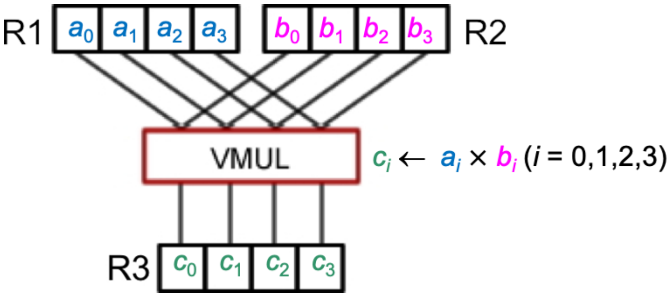

Fig. 1: An example of vector multiplier (VMUL), which (i) loads data from two vector registers, R1 and R2, each holding 4 double-precision numbers, (ii) concurrently performs 4 multiplications, and (iii) stores the results in vector register R3.

(Example) Fugaku (富岳) Computer

Fugaku is the world's fastest supercomputer as of June 2020 (**Fig. 2**).[*] Fugaku consists of 152,064 computing nodes, each with a 48-core ARM A64FX processor (**Fig. 3**).

[*] https://www.top500.org/system/179807/

Fig. 2: Fugaku computer.

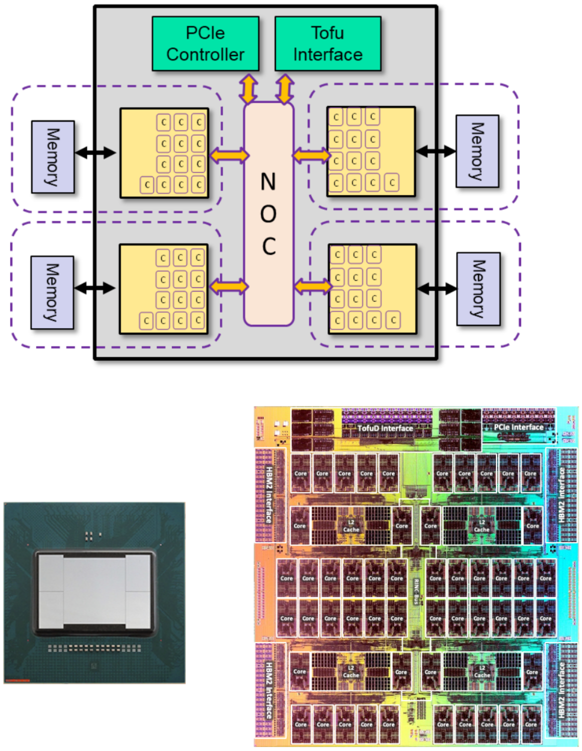

Fig. 3: Each ARM A64FX processor in Fugaku is composed of 48 cores.

Let us compute its theoretical peak flop/s using its basic data in the table below:

#processors	152,064
#cores_per_processor	48
clock_speed [GHz]	2.2
#FMA_units	2
vector_size [bits]	512

By substituting these numbers in the equation, we obtain

$$Peak\_flop/s = 152{,}064\ [\text{processors}] \times 48 \left[\frac{\text{cores}}{\text{processor}}\right] \times 2.2\ [\text{GHz}] \times 2 \times 2 \times \frac{512}{64} = 513{,}854{,}668.8\ [\text{Gflop/s}] = 514\ [\text{Pflop/s}].$$

Here, Gflop/s (gigaflop/s) is 10^9 floating-operations per second and Pflop/s (petaflop/s) is 10^{15} floating-point operations per second.

The measured flop/s performance for the basic linear algebra subprograms (BLAS) software is 415.53 Pflop/s, which is 80.9% of the theoretical peak. That is very high.

Reference

1. https://en.wikipedia.org/wiki/FLOPS.