Roofline performance model for CPU, GPU and FPGA

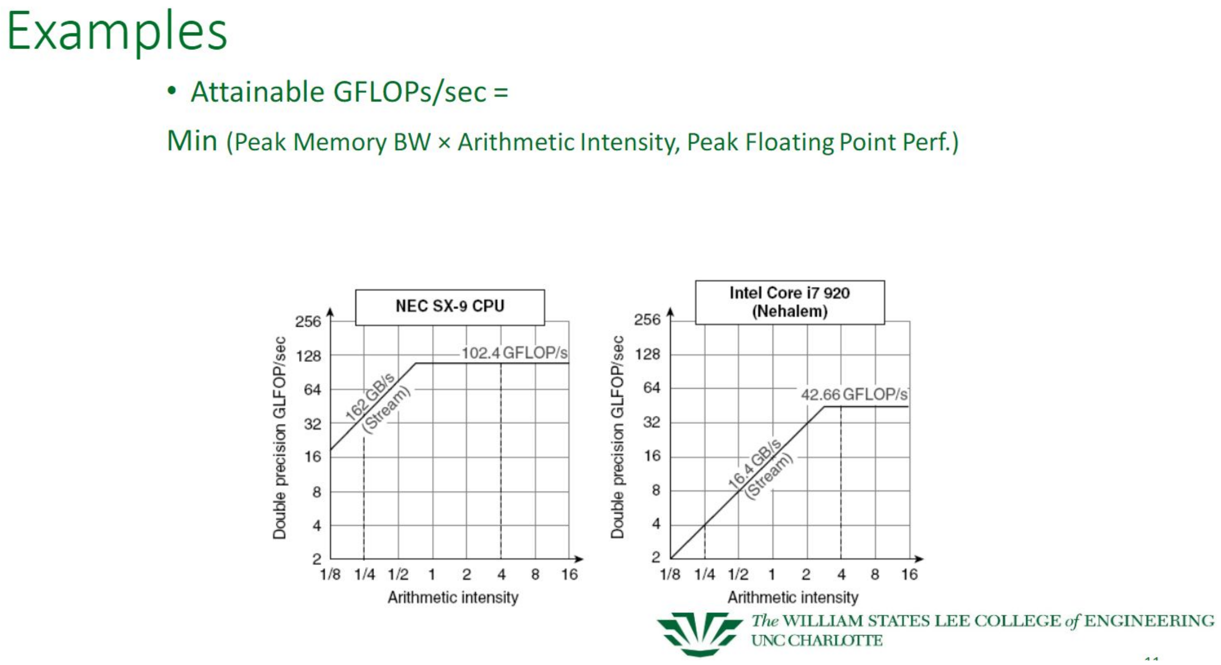
Roofline model - A performance model to assess the performance of numerical operations (integer or floating point) running on CPU or accelerators like GPU and FPGA.

Understanding: There are three key features to improve performance/

1. Computation - We want to increase the number of floating point operations per second (GFLOPs/sec)
2. Communication - we want the DDR bandwidth (GB/s) to be utilized efficiently.
3. Data locality - We want to minimize the cache hits and misses and effectively use the data from cache.

One model that helps to plot the performance - GFLOPs/sec in terms of arithmetic intensity - GFLOPs/byte is the roofline performance model.

Now what is arithmetic intensity - It is a measure of number of floating point operations performed relative to the amount of memory accesses in bytes.
Example: ¼ intensity means one operation requires 4 bytes of memory.



Examples

- Attainable GFLOPs/sec =

Min (Peak Memory BW × Arithmetic Intensity, Peak Floating Point Perf.)

The WILLIAM STATES LEE COLLEGE of ENGINEERING
UNC CHARLOTTE

For figure 1, we see that as the arithmetic intensity increases (means that we can do more floating point operations for given memory bytes) the performance increases. But as we keep on increasing the intensity from ½ to 1 and further, the performance does not increase anymore. Similarly, for figure 2, the performance saturates at intensity of 2.5 and does not increase thereafter.

This is because as the arithmetic intensity increases, there are more floating point operations one can do but there are no enough compute units (or ALUs) to compute these numericals. This is often referred to as compute bound.

Now, we want this compute bound to be as delayed as possible. This means that we want to delay the saturation and want to keep on increasing the arithmetic intensity and thus the performance.

So, when we compare figure 1 and 2, figure 2 has a better architecture as the arithmetic intensity saturates at 2.5 and thus has a compute bound at a delayed stage.

To further understand this, we look at 3 different computer architectures and assess the performance based on the Roofline model and see which one performs better and why.

Specs:
CPU: Intel Core i5-8250U processor

GPU: Nvidia K80

Xilinx FPGA: VU9P