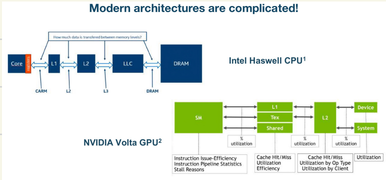
• 用作比較 SIMD architecture 下 FD 效能的 performance model

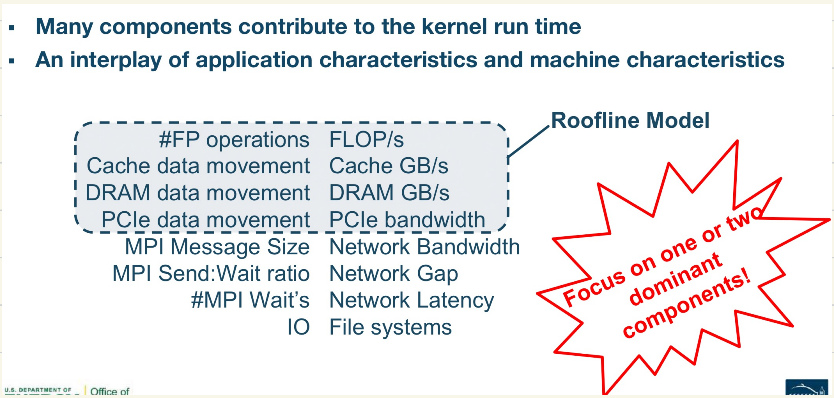
∵ 要考量 CPU core 算力外, 還要考慮 mem bandwidth




• 用 2D graph 呈現：

1. FP performance
2. MEM performance
3. arithmetic intensity : $\frac{\text{FP operations}}{\text{byte of mem accessed}} \Rightarrow \frac{\text{FP op for program}}{\text{Data Bytes transfer to mem during program execution}}$

用 Roofline model 可知 upper bound of FP operations / sec



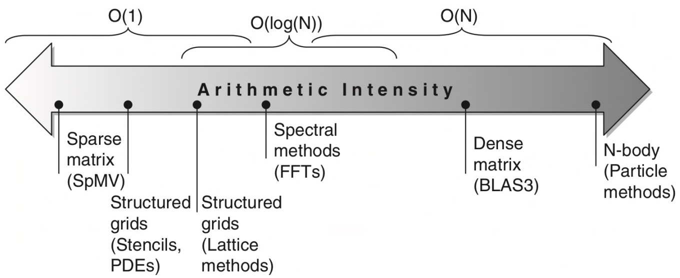



**Figure 4.10** Arithmetic intensity, specified as the number of floating-point operations to run the program divided by the number of bytes accessed in main memory [Williams et al. 2009]. Some kernels have an arithmetic intensity that scales with problem size, such as dense matrix, but there are many kernels with arithmetic intensities independent of problem size.

Roofline model 如下: 會對 computer system 分析建構 Roofline model

依照 program 之 arithmetic intensity 得到 FP operation / sec limit

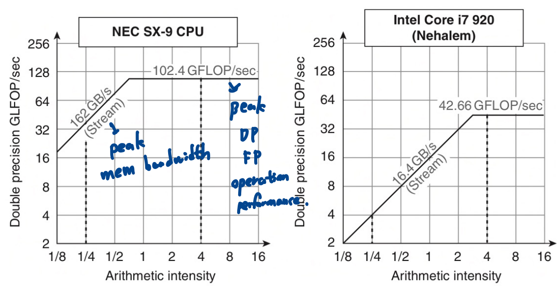



**Figure 4.11** Roofline model for one NEC SX-9 vector processor on the left and the Intel Core i7 920 multicore computer with SIMD Extensions on the right [Williams et al. 2009]. This Roofline is for unit-stride memory accesses and double-precision floating-point performance. NEC SX-9 is a vector supercomputer announced in 2008 that costs millions of dollars. It has a peak DP FP performance of 102.4 GFLOP/sec and a peak memory bandwidth of 162 GBytes/sec from the Stream benchmark. The Core i7 920 has a peak DP FP performance of 42.66 GFLOP/sec and a peak memory bandwidth of 16.4 GBytes/sec. The dashed vertical lines at an arithmetic intensity of 4 FLOP/byte show that both processors operate at peak performance. In this case, the SX-9 at 102.4 FLOP/sec is 2.4× faster than the Core i7 at 42.66 GFLOP/sec. At an arithmetic intensity of 0.25 FLOP/byte, the SX-9 is 10× faster at 40.5 GFLOP/sec versus 4.1 GFLOP/sec for the Core i7.

∵ arithmetic intensity $= \dfrac{\text{FP op for program}}{\text{Data Bytes transfer to mem during program execution}}$

當 a.i. 大時, 表示 FP 運算量 >> mem 搬移量 ⇒ depends on peak FP performance

a.i. 小時, << ⇒ depends on peak mem bandwidth
資料要搬入才能運算

∵ X 軸為 FPOP / byte　Y 軸為 FPOP / sec

∴ slope $= \dfrac{\text{FPOP / sec}}{\text{FPOP / byte}}$ = byte / sec ⇒ peak mem performance

故 FPOP/sec 的計算為: min{ peak mem bandwidth × arithmetic intensity, peak FP performance }

∴ 只有在 arithmetic intensity 大時, 才能達到 max performance

且可看出 SIMD machine 的 mem bandwidth 高於 MIMD core

## Roofline Performance Model

- Sustainable performance is bound by

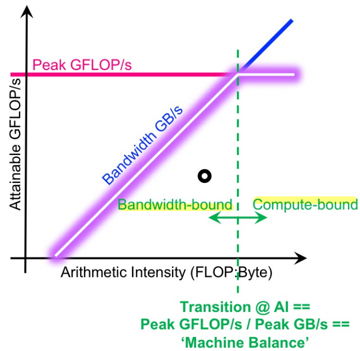
$$\text{GFLOP/s} = \min\begin{cases}\text{Peak GFLOP/s}\\ \text{AI} * \text{Peak GB/s}\end{cases}$$

- Arithmetic Intensity (AI) =

  FLOPs / Bytes

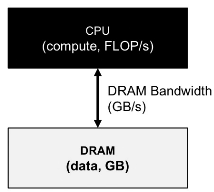
- How did this come about?

  → A CPU DRAM example




## (CPU DRAM) Roofline

- One could hope to always attain peak performance (FLOP/s)
- However, finite locality (reuse) and bandwidth limit performance.
- Assume:
  - Idealized processor/caches
  - Cold start (data in DRAM)




$$\text{Time} = \max\begin{cases}\text{\#FP ops / Peak GFLOP/s}\\ \text{\#Bytes / Peak GB/s}\end{cases}$$

$$\frac{\text{Time}}{\text{\#FP ops}} = \max\begin{cases}\text{1 / Peak GFLOP/s}\\ \text{\#Bytes / \#FP ops / Peak GB/s}\end{cases}$$

$$\frac{\text{\#FP ops}}{\text{Time}} = \min\begin{cases}\text{Peak GFLOP/s}\\ \text{(\#FP ops / \#Bytes) * Peak GB/s}\end{cases}$$

$$\text{GFLOP/s} = \min\begin{cases}\text{Peak GFLOP/s}\\ \text{AI * Peak GB/s}\end{cases}$$

Arithmetic Intensity (AI) = FLOPs / Bytes (as presented to DRAM )

## Roofline Performance Model

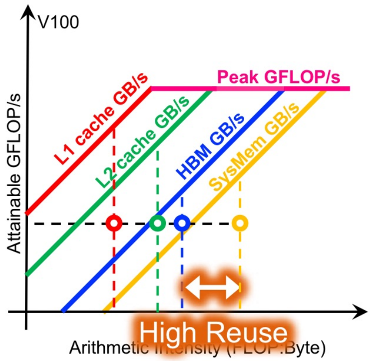
- A throughput-oriented model
  - tracks rates not times, i.e. GFLOP/s, GB/s, not seconds
- An abstraction over
  - architectures, ISA (CPU, GPU, Haswell, KNL, Pascal, Volta)
  - programming models, programming languages
  - numerical algorithms, problem sizes
- In log-log scale to easily extrapolate performance along Moore's Law

## Hierarchical Roofline

- Superposition of multiple Rooflines
  - Incorporate full memory hierarchy
  - Arithmetic Intensity =

    FLOPs / Bytes$_{L1/L2/HBM/SysMem}$

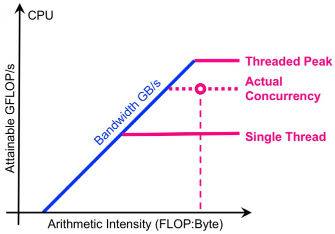
- Each kernel will have multiple AI's but one observed GFLOP/s performance




- Hierarchical Roofline tells you about **cache locality**

## Multiple Compute Ceilings

- Impact of **execution configuration**
- Concurrency affects your peak
  - OpenMP thread concurrency
  - SM occupancy
  - load balance
  - threadblock/thread configuration




- Performance is bound by the **actual concurrency** ceiling

# Multiple Compute Ceilings

- Impact of **instruction mix**
- Applications are usually a mix of FMA.f64, ADD.f64, MUL.f64…
- Performance is a **weighted** average … bound by a **partial FMA** ceiling