High-**P**erformance **C**omputing 20**24**

Basics of Modern CPUs, SIMD, Memory Hierarchy, and Performance Considerations

Aryan Eftekhari, Olaf Schenk | **U**niversità della **S**vizzera **i**taliana (USI) | **I**nstitute of **C**omputing (CI)

Overview

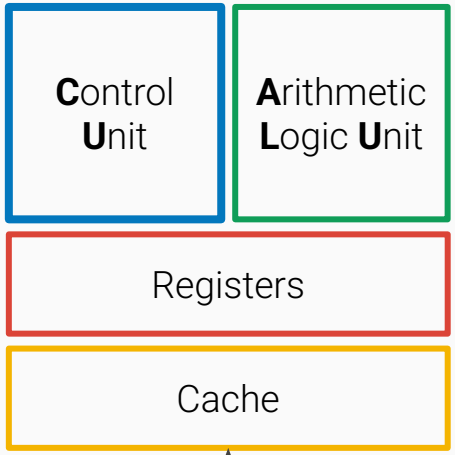
The key component of a modern CPU and performance considerations.

Next Lecture, Prof. Schenk will discuss the Roofline Model (Next Tuesday).

Modern CPU

Von Neumann Architecture–CPU + Main Memory

Central Processing Unit



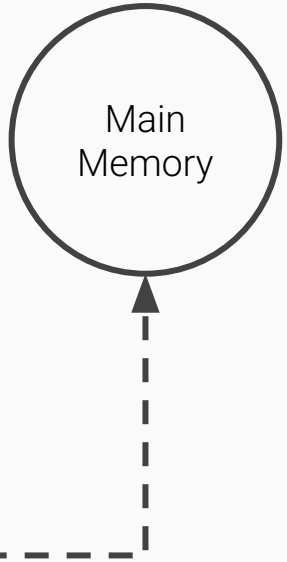
CU: Interprets instructions for execution and manages the ALU, memory transfers, and other components.

ALU: Performs arithmetic and logical operations like addition, subtraction, AND, OR, NOT, and more.

Registers: storage where data must be stored before processing.

Cache: high-speed memory for storing copies of frequently accessed data to improve performance. Serves as a buffer between the registers and main memory.

Main Memory: Also known as primary or Random Access Memory (RAM). RAM is very slow compared to the CPU cache.



… we are ignoring the *input/output* devices and *disk*.

Modern CPU
SIMD–Single Instruction Multiple Data

An **instruction** is carried out on (multiple) **inputs** resulting in (multiple) **outputs**.

E.g., *Advanced Vector Extension* (AVX512[1]) (*NEON* for ARM[2]).

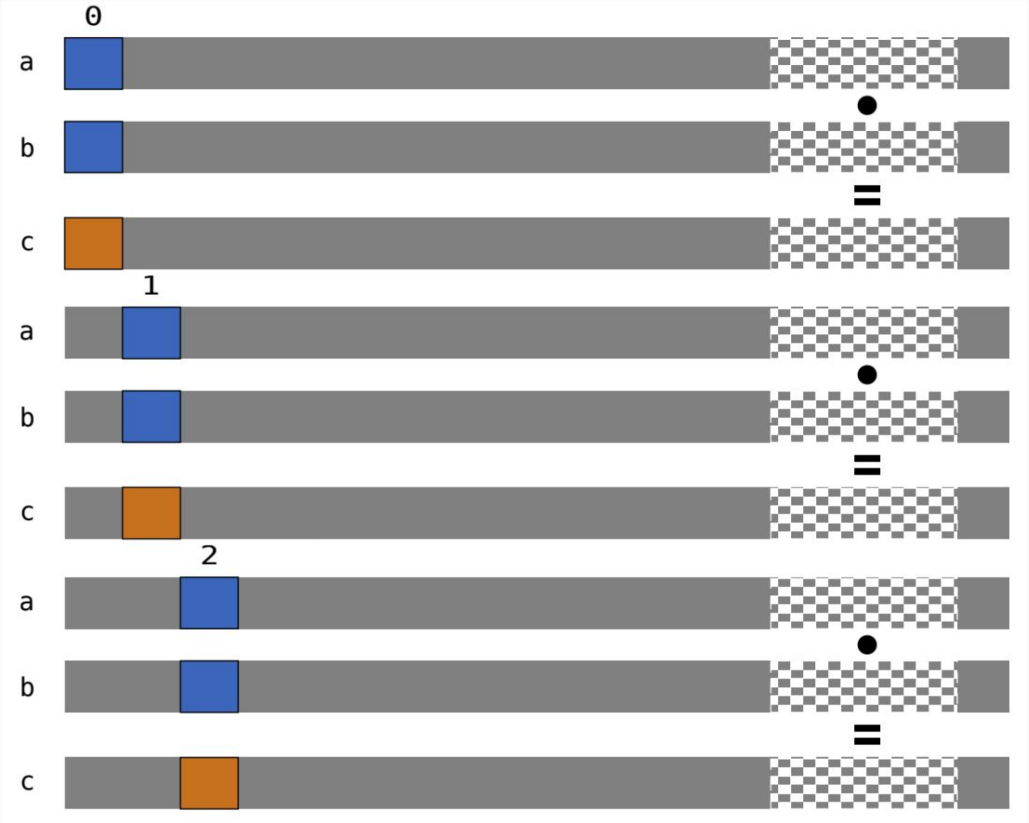
There are other levels of parallelism, such as instruction-level, thread-level, process-level, etc.

[1] https://www.intel.com/content/dam/develop/external/us/en/documents/31848-compilerautovectorizationguide.pdf
[2] https://developer.arm.com/Architectures/Neon

Modern CPU
SIMD–Single Instruction Multiple Data

Scalar Operation

```
for (size_t i = 0; i < n; i++){
  c[i]=a[i]*b[i]
}
```



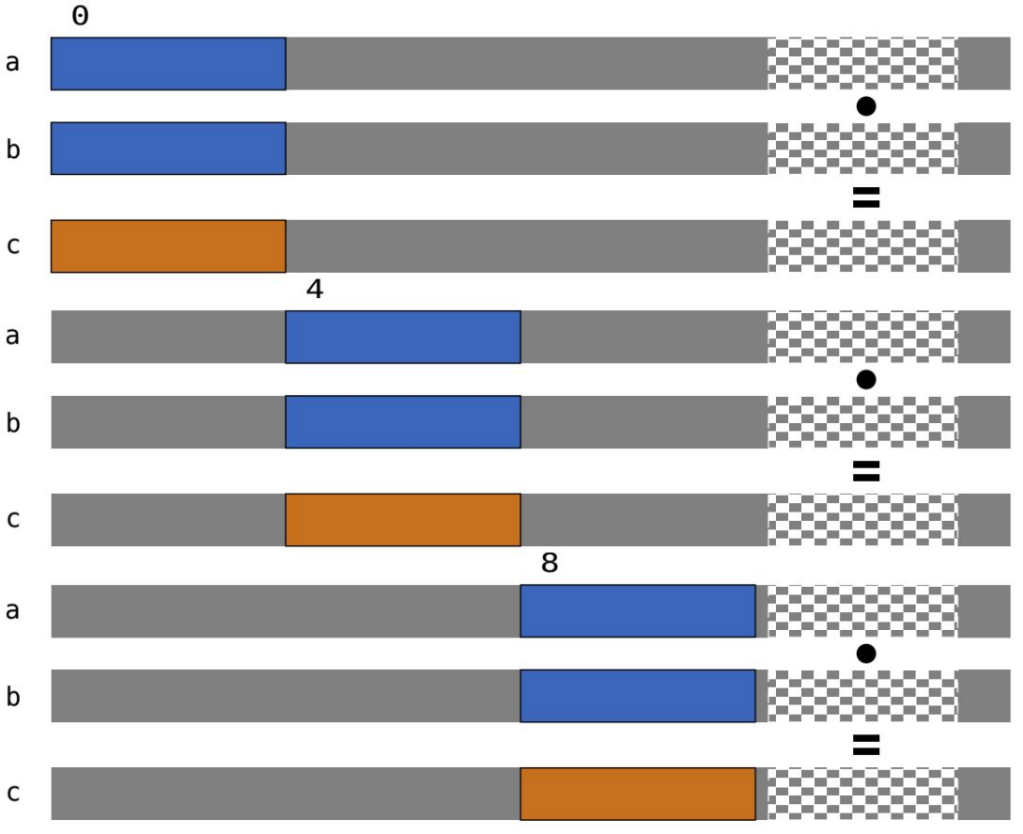
Modern CPU
SIMD–Single Instruction Multiple Data

SIMD Operation

```
for (size_t i = 0; i < n; i++){
  c[i]=a[i]*b[i];
}
```

… it executes “like this” … but each iteration happens all at once.

```
for (size_t i = 0; i < n/3; i++){
  // This happened all at once!
  c[i*3+0]=a[i*3+0]*b[i*3+0];
  c[i*3+1]=a[i*3+1]*b[i*3+1];
  c[i*3+2]=a[i*3+2]*b[i*3+2];
}
```



Modern CPU
SIMD-AVX Intrinsics

```
int main() {

double a[] = {1.0, 2.0, 3.0, 4.0};
double b[] = {5.0, 6.0, 7.0, 8.0};
double c[4];

// Load arrays into 256-bit registers
__m256d vec_a = _mm256_loadu_pd(a);
__m256d vec_b = _mm256_loadu_pd(b);

// Perform addition
__m256d vec_c = _mm256_add_pd(vec_a, vec_b);

// Store the result back into array c
_mm256_storeu_pd(c, vec_c);

// Print the result
for (int i = 0; i < 4; ++i) {
  std::cout << c[i] << ' ';
}
std::cout << std::endl;

return 0;
}
```

Some Data Types:

- `__m256`: Represents a 256-bit data type holding 8 *single-precision* floating-point numbers.
- `__m256d`: Represents a 256-bit data type holding 4 *double-precision* floating-point numbers.

See intel.com/content/www/us/en/docs/intrinsics-guide/index.html for details.

Compilers can optimize most general-purpose applications using SIMD (even advanced ones) … *but it's difficult to predict exactly how they will behave.*

Compiler: Unroll loop + packed SIMD instructions.

Flag	Optimization level	Execution Time	Compile Time
-O0	Optimization: compilation time (default)	+	-
-O1 or -O	Optimization: code size (small executable) and execution time	-	+
-O2	Optimization (more): code size and execution time (vectorization)	--	++
-O3	**Optimization (even more): code size and execution time (vectorization)**	**---**	**+++**
-Os	Optimization: code size (small executable)		++

We hope that the compiler will automatically optimize loops, although this is not guaranteed.

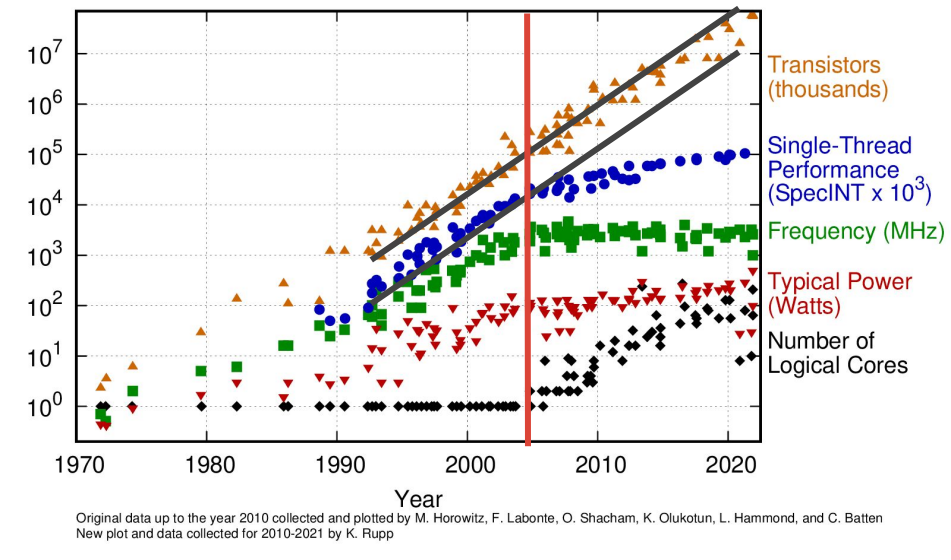
Modern CPU
Multi or Manycore Parallelism

Computers have stopped getting “faster” since the mid 2000’s !

Provocative statement to get attention.

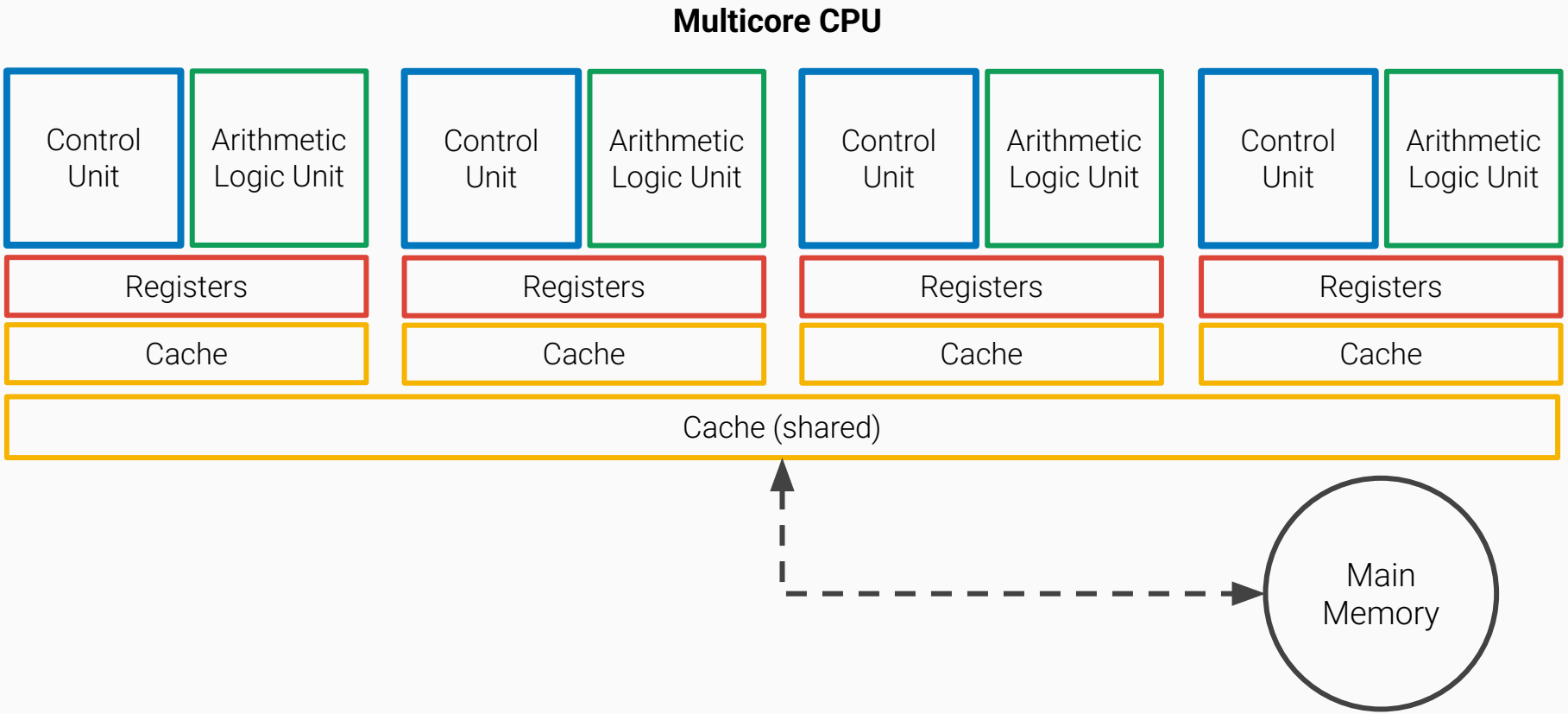
1970 to Mind 2000’s - Increasing Clock Speed
No longer an option due to heat and power.

Mind 2000’s to Current - Increasing Cores
More than one core, i.e., multicore processors.



Modern CPU

Multicore or Manycore CPU



Considerations for runtime …

A program requires:
(i) "*a*" **arithmetic operations**, each taking time "t_a", and
(ii) "*m*" **memory operations** each taking time "t_m".

Overall Runtime = { **Arithmetic Operations** $a \cdot t_a$ and ***Memory Operations*** $m \cdot t_m$ }

We have only been talking about **Arithmetic** component of runtime.

What about **Memory** ? Note $t_m \neq 0$!

Definitions:

- FLOPs = Number of Floating Point Operations
- FLOP**S** = FLOPs per second
- Peak FLOPS = Number of Cores × Clock Speed (cycle/sec) × Operations Per Cycle (FLOPs/cycle)

For example, a SIMD-enabled CPU with a 256-bit register width can multiply 8 floats (each 32 bits) in one cycle, making the operations per cycle count 8. What if we had doubles instead of floats?

- **Arithmetic Intensity** = FLOPs / Number of bytes of data transfer
- **Bandwidth** = Bytes of data transferred / second

Performance costs for doing the **FLOPs** and the **data transfer** …
we *will* have bottlenecks.

Performance & Memory
Attainable Performance

Consider a machine with a peak performance of 170 GFLOPS (FLOPs per second) and a memory bandwidth of 50.0 GB/s.

Arithmetic Intensity: Characteristic of a program [FLOPs/sec · $(B/sec)^{-1}$ = FLOPs/B]:
- At best, the machine can execute approximately 3.4 FLOPs per byte loaded from memory.

Compute Bound: Program A does 5 FLOPs per byte. For every gigaflop:
- Time spent on memory operations: ~1/(5 x 50), [$(FLOPs/B \cdot B/sec)^{-1}$ = sec/FLOPs]
- Time spent on arithmetic operations: ~1/170 (Bottleneck). [$(FLOPs/sec.)^{-1}$ = sec/FLOPs]

Memory Bound: Program B does 1 FLOP per byte. For every gigaflop:
- Time spent on memory operations: ~1/(1 x 50) (Bottleneck),
- Time spent on arithmetic operations: ~1/170.

This is the crux of the “roofline model.”

Performance & Memory

Bandwidth–A sense of scale

The rate of read/write/transfer.

Cache

L1/L2/L3 cache has a bandwidth in the ~100s of GB/s.

Main Memory

The bandwidth of ram is on the order of ~10 GB/s.

Local Network

High-performance networks in data centers have a bandwidth of ~1GB/s.

Disk

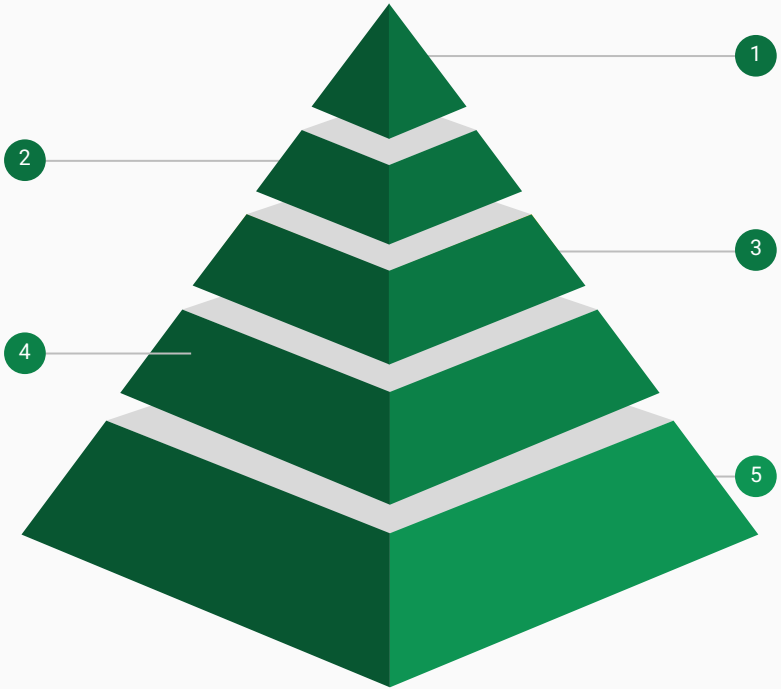
The bandwidth of disks can vary depending on type of SSD ~0.1GB/s.

Internet

4G is about ~0.01 GB/s.

Think about bottlenecks …

Performance & Memory
Memory Hierarchy



Registers ~ 64–512 bit
Integral to the CPU, lowest latency. Size varies: x86-64 general-purpose registers are 64-bit, Apple's M1 SIMD registers are 128-bit, and some Intel CPUs have 512-bit SIMD registers.

L1 Cache ~16–128 KB
Extremely fast with only few cycles latency.

L2 Cache ~256KB–1MB
Very fast, slower than L1 but much faster than main memory.

L3 Cache ~1–8 MB
Fast, slower than L2 but much faster than main memory. *Shared with all cores.*

Main Memory ~32–128 GB
Compared to the L3 this is slow.

No free lunch…fast memory is expensive and limited in capacity.

The values above are for reference and can vary significantly between products and over time.

Motivation for Assignment 1 …

“How good is your code?”

Understanding the maximum performance capabilities of your machine will help answer this question.

MSc Course - High Performance Computing

Single Processor Machines: **Memory Hierarchies, Memory Bandwidth, Processor Features and Locality**

Olaf Schenk
Institute of Computing, INF, USI
October 01, 2024

Outline

- Introduction memory hierarchy
- Memory access
- Case study: Matrix-matrix multiplication
- Berkeley Roofline Model of Performance

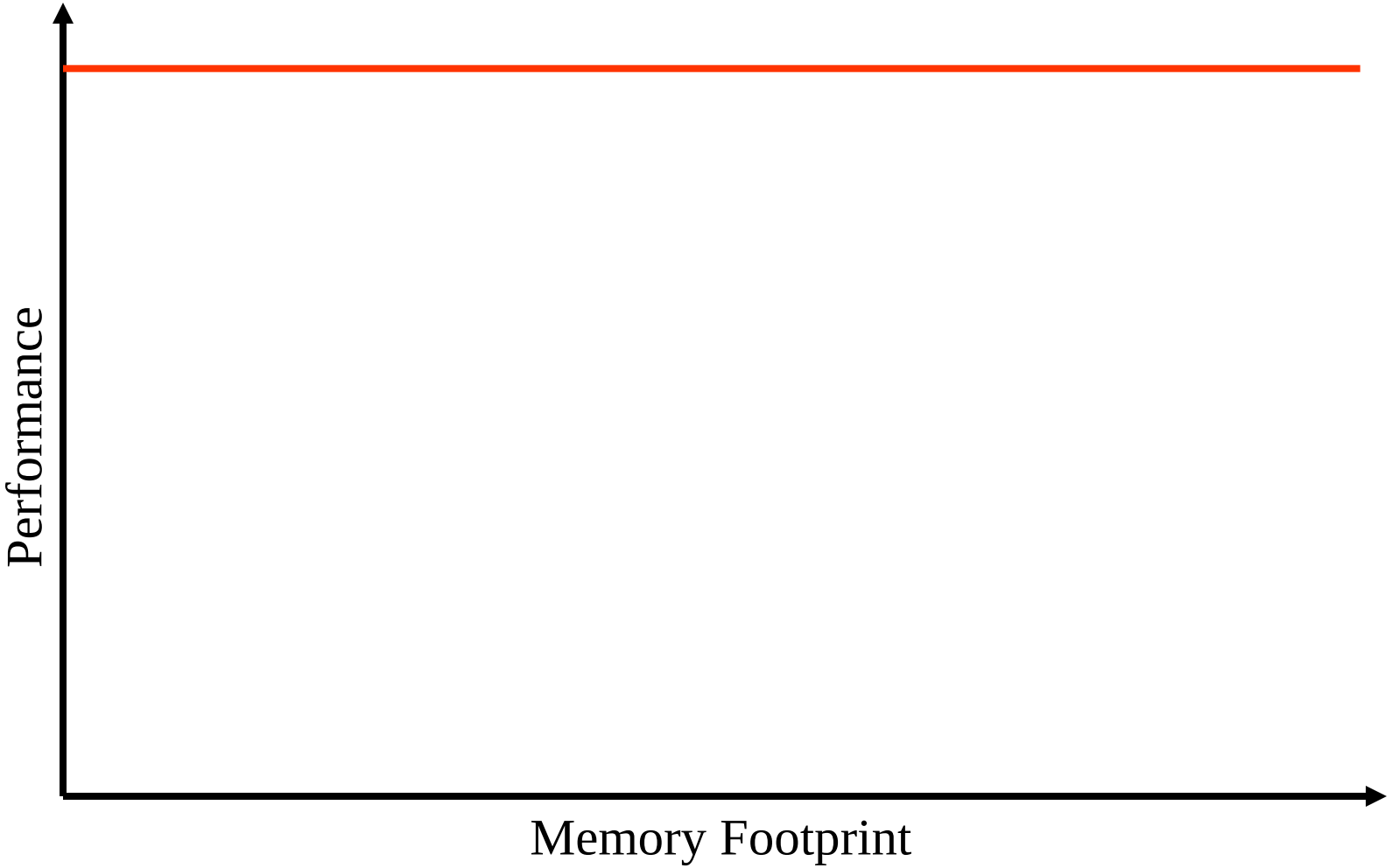
Motivation

- Most applications run at < 5% of the “peak” performance of a system
  - **Peak is the maximum the hardware can physically execute**

- Much of this performance is lost on a single core, i.e., the code running on one core often runs at only 5% of the processor peak

- Most of the single processor performance loss is in the memory system
  - **Moving data takes much longer than arithmetic and logic**

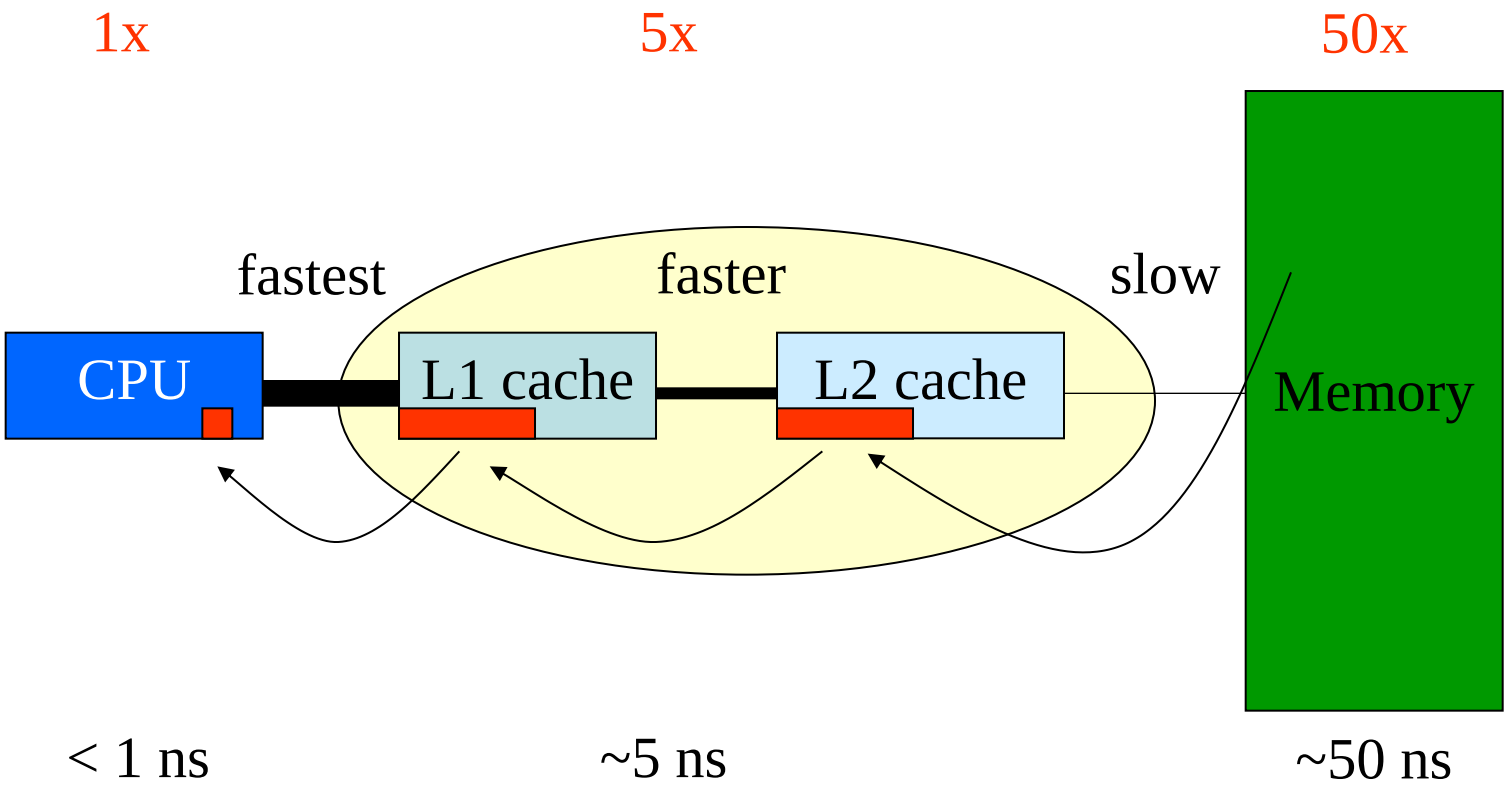
Intuitive Performance Graph



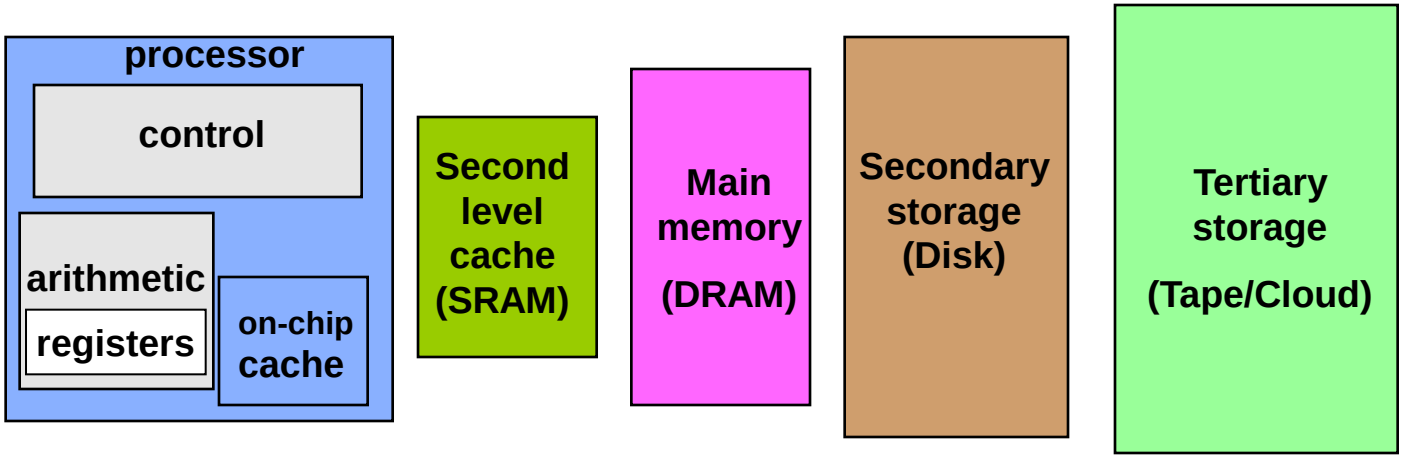
About Memory

- Memory plays a crucial role in performance
- Not accessing memory in the right way will degrade performance on all computer systems
- The extent of the degradation depends on the system
- Knowing more about some of the relevant memory characteristics will help you to write codes such that the problem will be non-existent, or at least minimal

Typical Cache Based System

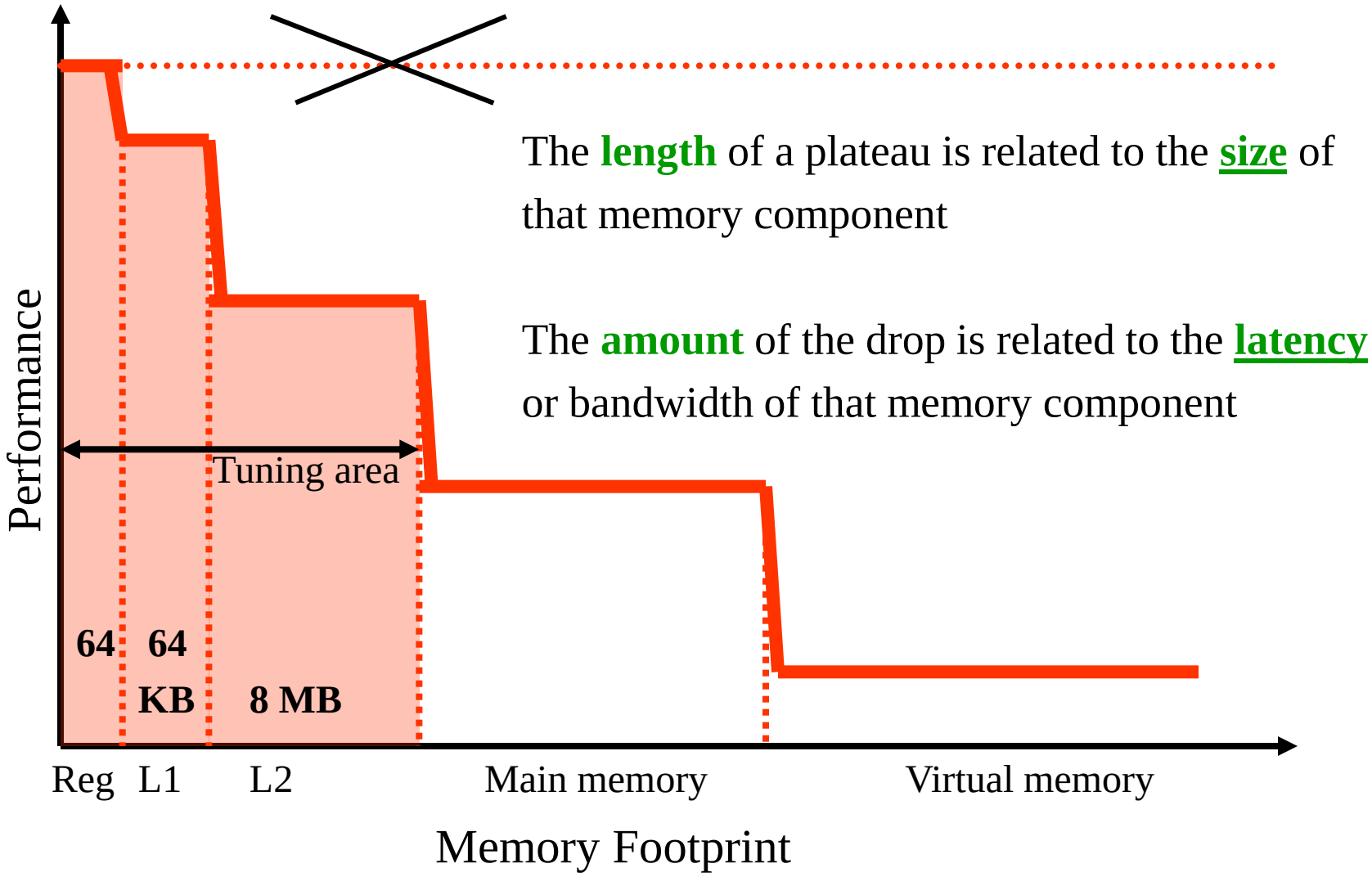


The Memory Hierarchy



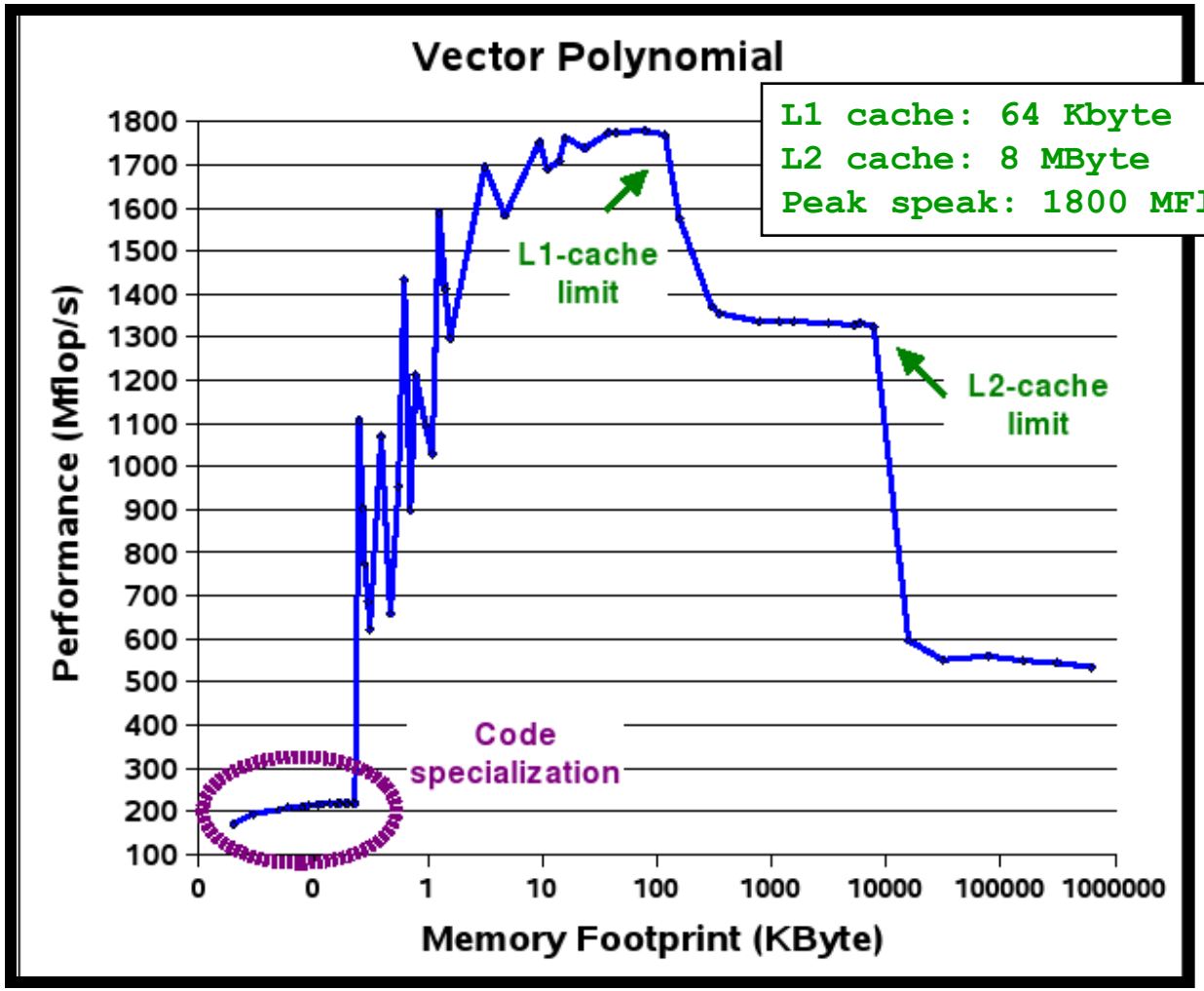
Speed	1ns	10ns	100ns	10ms	10sec
Size	KB	MB	GB	TB	PB

Performance is not Uniform



Example – 13th degree polynomial

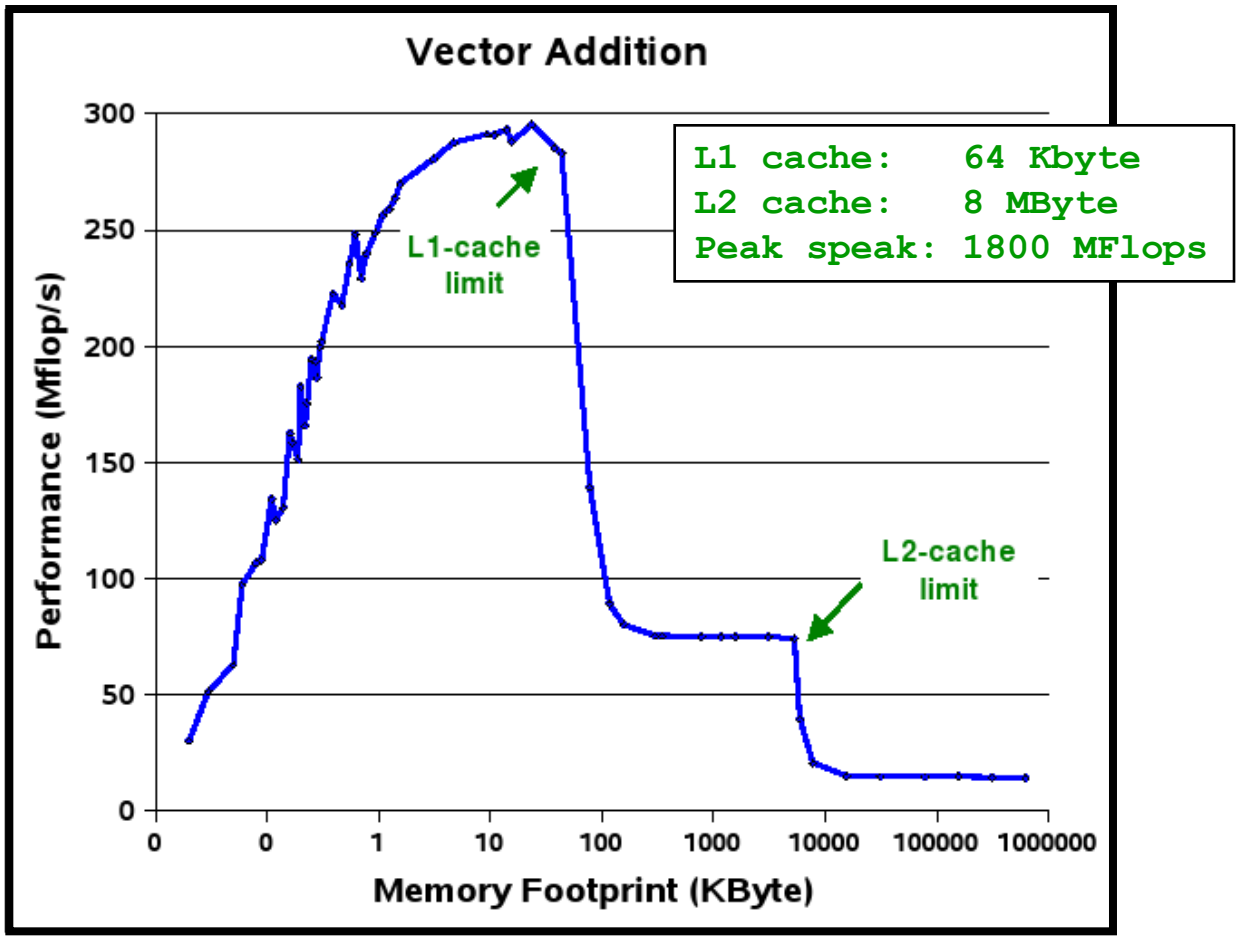
```
for  ( i=0 ; i<vlen; i ++)
   p[i]=c[0]+q[i]*(c[1]+q[i]*(c[2]+q[i]*(c[3]+ q[i] …
```



- This operation is **Compute bound** i.e. there are much more floating point operations than memory references
- The system realizes over 98% of the absolute peak performance
- Note that's start-up effect and the performance drop for larger problems

Example – Vector Addition

```
for  ( i=0 ; i<vlen; i ++)
   p[i]= q[i] + r[i];
```



- This operation is **Memory bound** i.e. there are much more memory references than floating point operations
- Note the start-up effect and the performance drop for larger problems

Outline

- Intro memory hierarchy
- Cache mappings
- Memory access
- Case study: Matrix-matrix multiplication

Cache Choices

- Modern systems use a wide variety of caches
- Typically there are at least 3 types of caches
  - Instruction cache
  - Data cache(s)
  - Page Address Cache (TLB)
- Some key design criteria
  - Size
  - Architecture/Mapping
  - Usage and cost are amongst the key decision factors

Cache Line Replacement

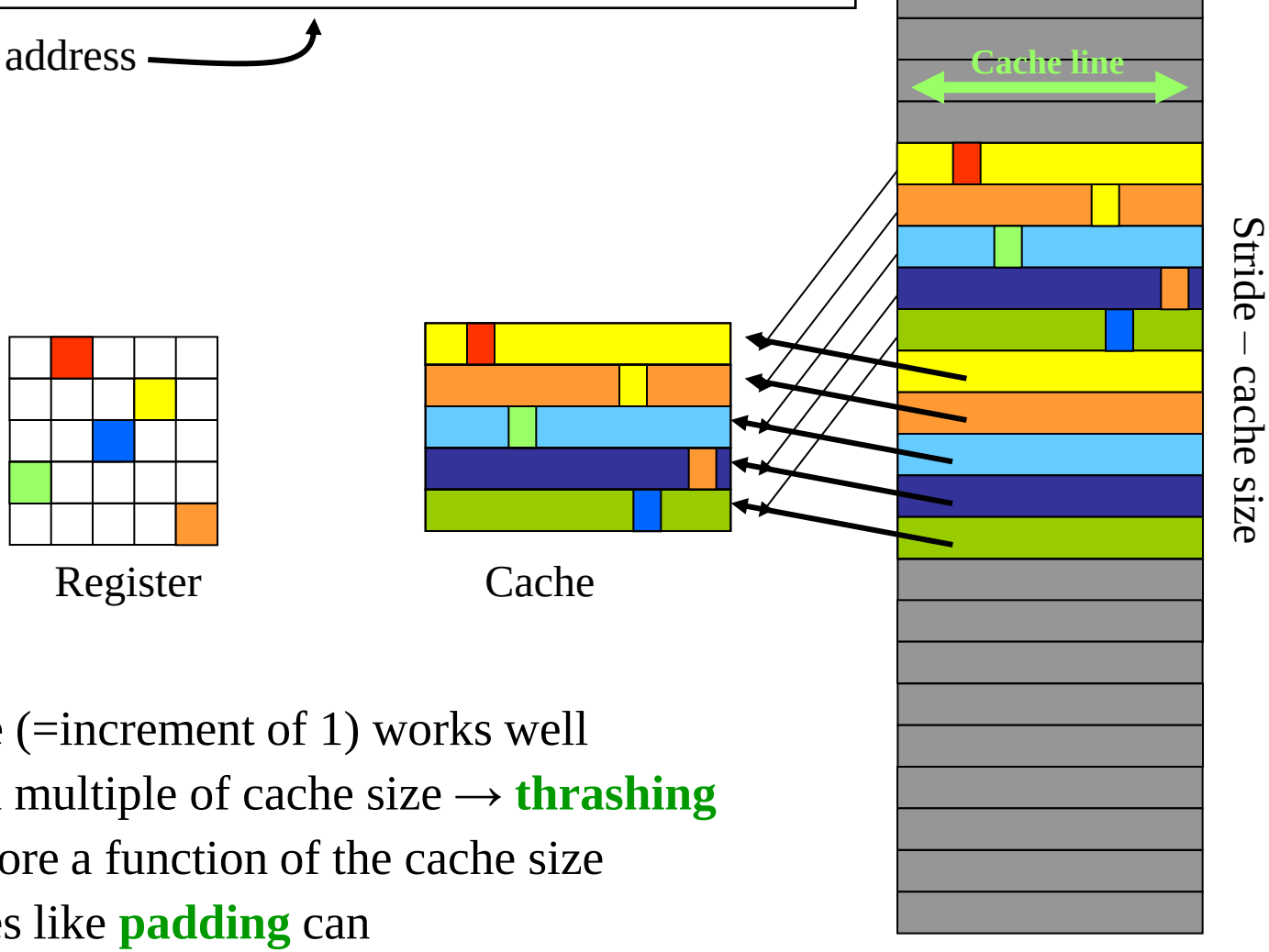
- Caches have a finite size
- A natural problem is how to replace data
- Three typical methods of doing this:
  - **Direct Mapped**
    Location: Memory location modulo cache size
  - **Fully Associative**
    Location: Least Recently Used (LRU, oldest cache line)
  - **X-way-Set Associative***
    Location: Choose a set first; direct mapped in set

*Note: X is typically 2, 4 or 8

A Direct Mapped Cache

```
Cache Location = Memory Address % Cache Size
```

Virtual or physical address



- **Comments**
  - Unit stride (=increment of 1) works well
  - **Stride** is a multiple of cache size → **thrashing**
  - It is therefore a function of the cache size
  - Techniques like **padding** can reduce or eliminate trashing

Example Direct Cache Mapping

```
float a[4096], b[4096];
for  (i=0 ; i< 4096; i ++)
   sum += a[i]*b[i];
```

Assumption:
- A cache line 8*4 Byte = 32 Byte

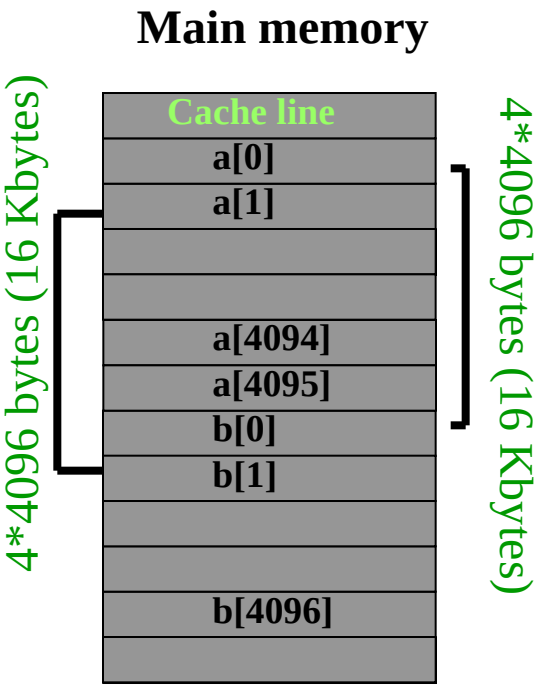
```
load a[0] into a reg.      miss: brings a[0] … a[7] into cache
load b[0] into a reg.      miss: brings b[0] … b[7] into cache
sum +=a[0]*b[0]
load a[1] into a reg.      hit: brings a[1] still in cache
load b[1] into a reg.      hit: brings b[1] still in cache
sum +=a[1]*b[1]
                    . . . etc . . .
load a[7] into a reg.      hit: brings a[7] still in cache
load b[7] into a reg.      hit: brings b[7] still in cache
sum +=a[7]*b[7]
load a[8] into a reg.      miss: brings a[8] … a[15] into cache
load b[8] into a reg.      miss: brings b[8] … b[15] into cache
sum +=a[8]*b[8]
load a[9] into a reg.      hit: brings a[9] still in cache
load b[9] into a reg.      hit: brings b[9] still in cache
sum +=a[9]*b[9]
```

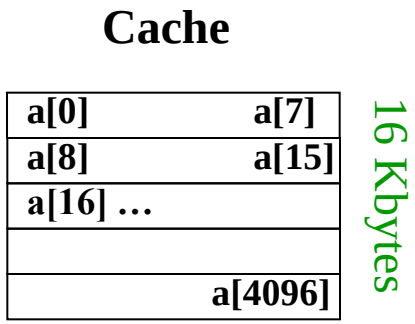
Assume Different Memory Lay-out

```
float a[4096], b[4096];
for  (i=0 ; i< 4096; i ++)
   sum += a[i]*b[i];
```

Assumption:

- A cache line 8*4 Byte = 32 Byte
- The cache size is 16 Kbyte





The cache is completely filled in a[4096]

Every Reference is a Cache Miss!

```
float a[4096], b[4096];
for (i=0 ; i< 4096; i ++)
   sum += a[i]*b[i];
```

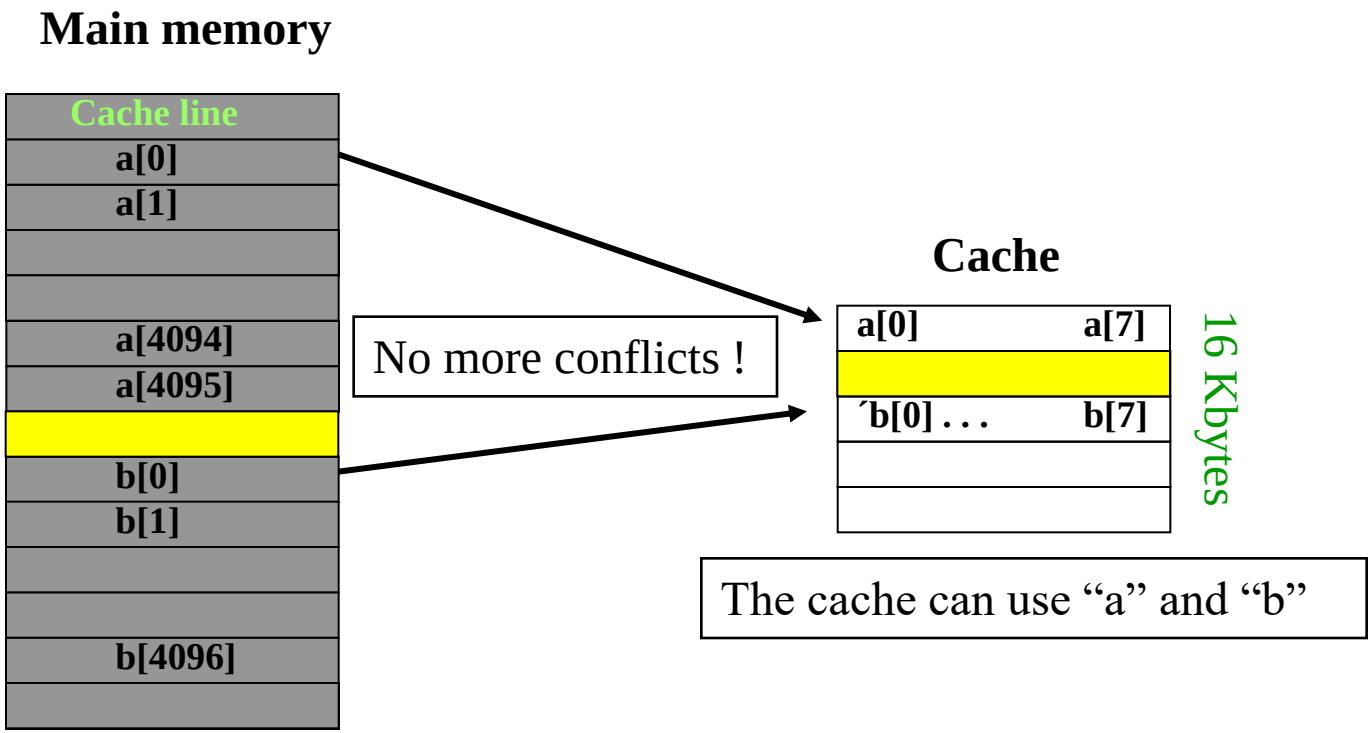
Assumption:

- A cache line 8*4 Byte = 32 Byte
- The cache size is 16 Kbyte

```
load a[0] into a reg.      miss: brings a[0] … a[7] into cache
load b[0] into a reg.      miss: brings b[0] … b[7] into cache
                           load of b[0] removes a[0]..a[7]
sum +=a[0]*b[0]
load a[1] into a reg.      miss: a[1] has just been removed
                           load of a[1] removes b[0]..b[7]
load b[1] into a reg.      miss: b[1] has just been removed
                           load of b[1] removes a[0]..a[7]
sum +=a[1]*b[1]
load a[2] into a reg.      miss: a[2] has just been removed
                           load of a[2] removes b[0]..b[7]
load b[2] into a reg.      miss: b[2] has just been removed
                           load of b[2] removes a[0]..a[7]
sum +=a[2]*b[2]
```

Padding – Change Address

```
float a[4096], padd[], b[4096];
for (i=0 ; i< 4096; i ++)
   sum += a[i]*b[i];
```

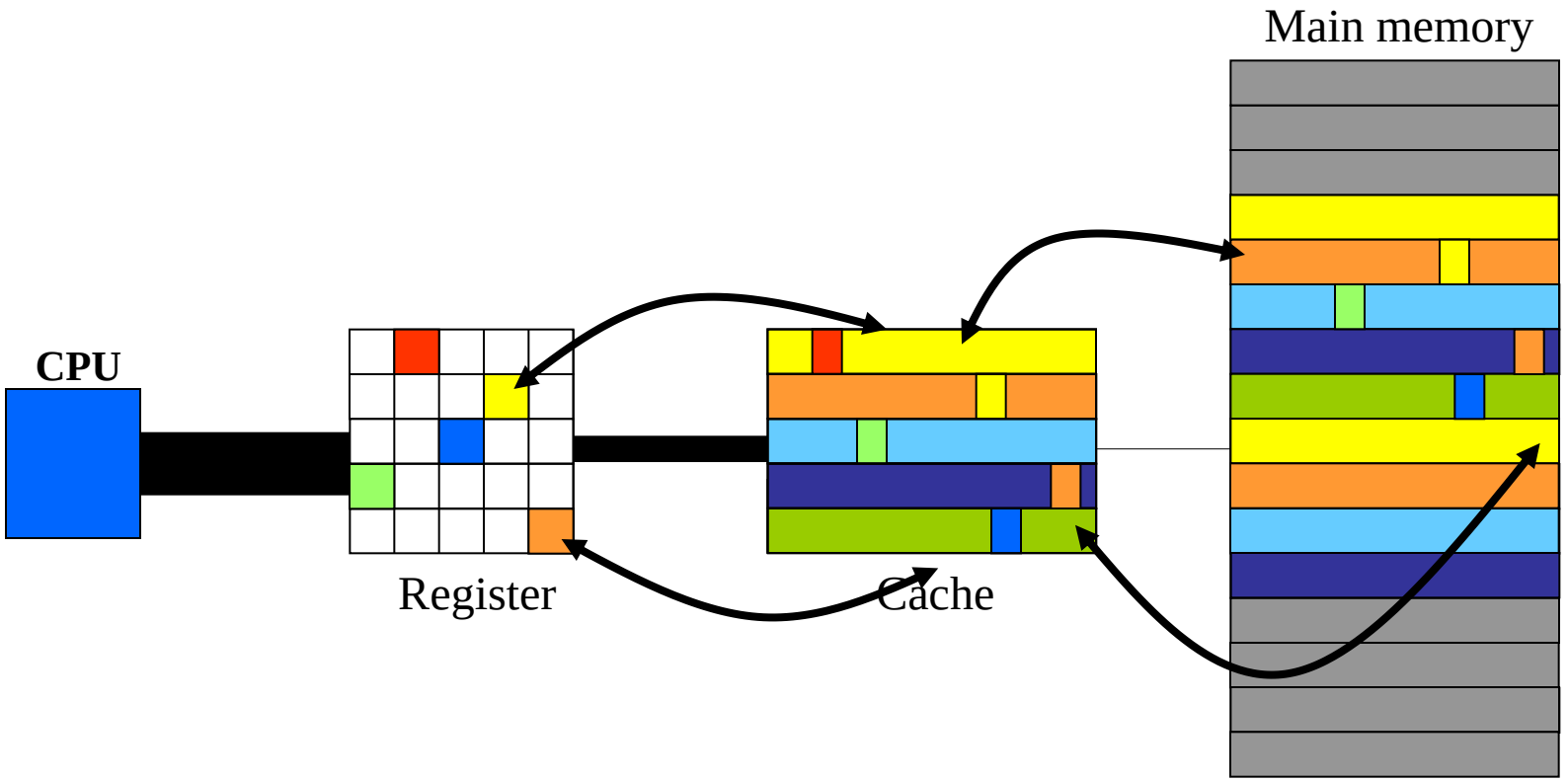


Outline

- Intro memory hierarchy
- Cache mappings
- Memory access
- Case study: Matrix-matrix multiplication

Cache lines

- For **good performance**, it is crucial to use the cache(s) in the intended (=optimal) way
- Recall that the unit of transfer is a cache line
- A cache line is a linear structure i.e. it has a fixed length (in bytes) and a starting address in memory



Cache Line Utilization

- Two Key **Rules for Performance** – **Maximize**

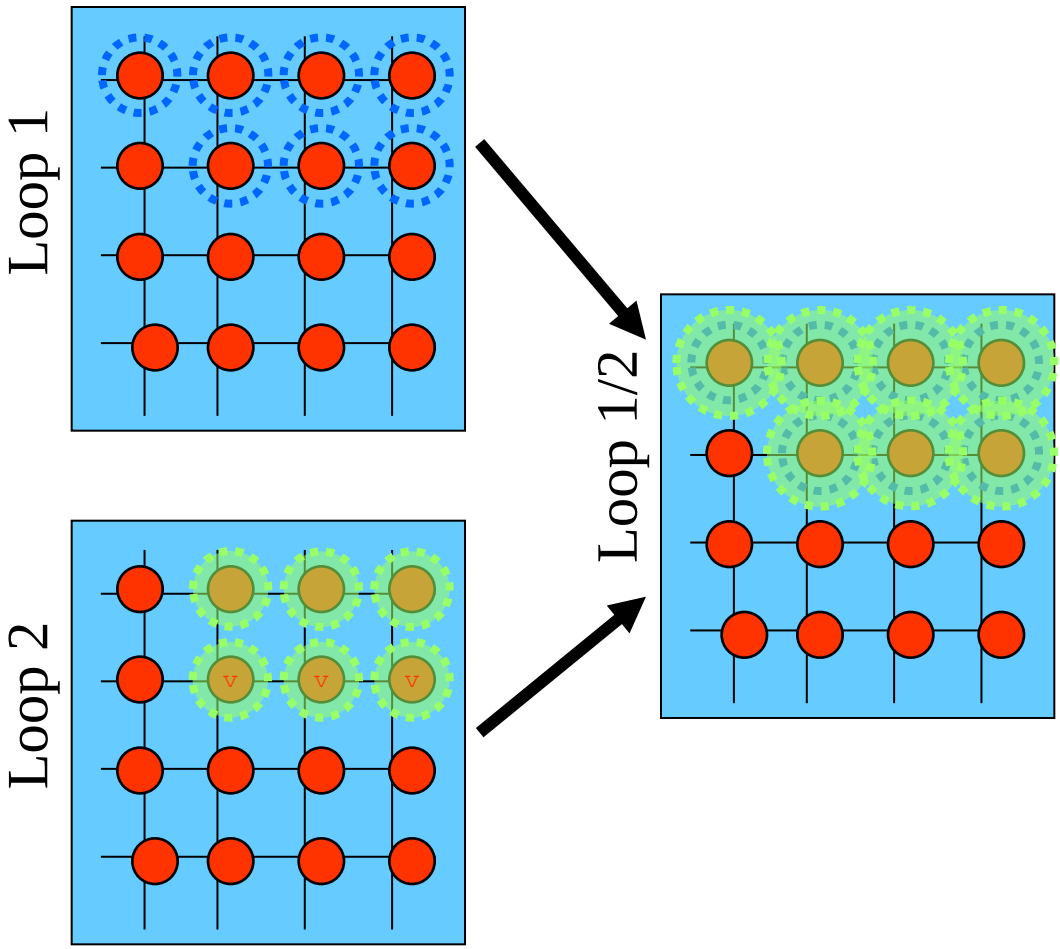
- **Spatial Locality** − Use all data in one cache line

  This strongly depends on the storage of your data and the access patterns

- **Temporal Locality** − Re-use data in a cache line

  This mainly depends on the algorithm used

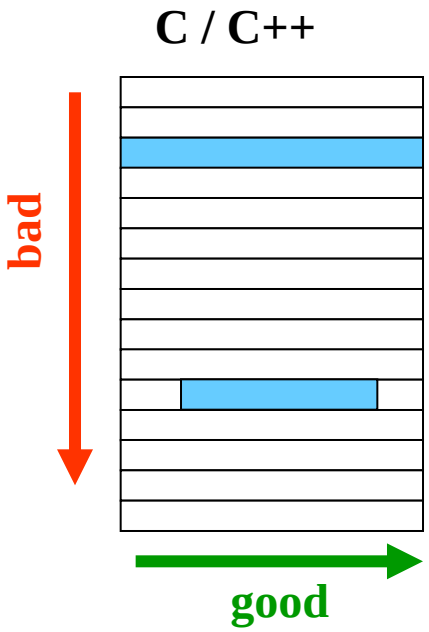
Cache Line Re-Use



- On the left we show a typical vector style of coding
- It is not a good approach for cache based systems: all grid elements have to be reloaded for each loop
- It is more beneficial to pre-calculate expression on the already loaded grid points

Memory Access

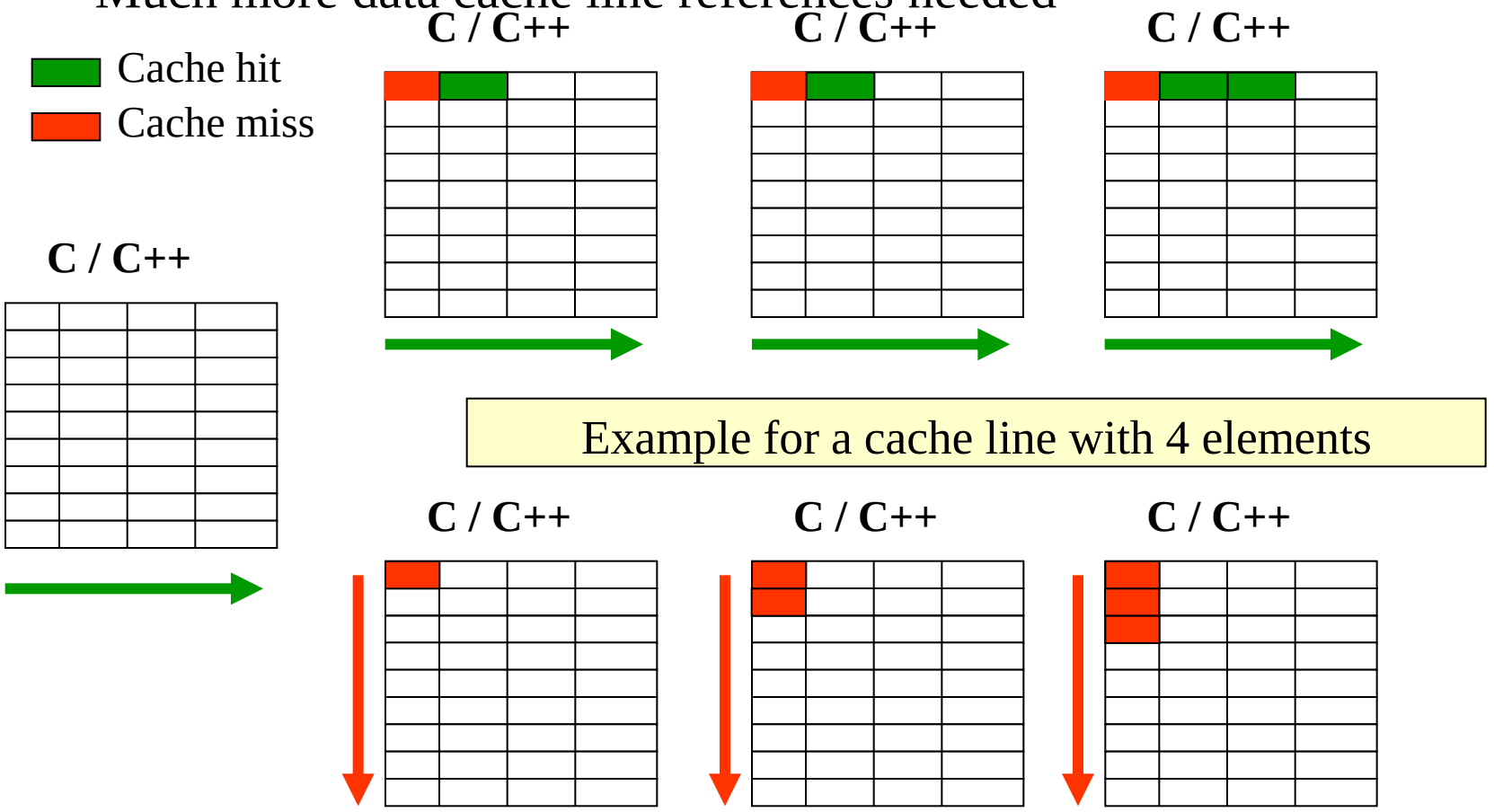
- Memory has always a 1D linear structure
- Access to multi-dimensional arrays depends on the way data is stored
- This is language dependent:



Bad Memory Access has a Huge Impact on Performance

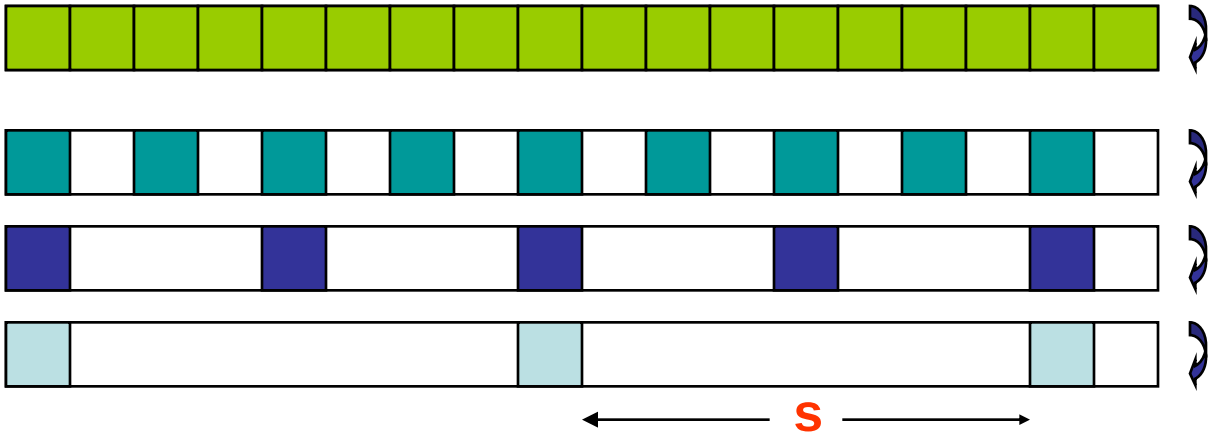
Non-unit Stride is always Bad

- If one follows the storage order in memory, all is well; This is called unit stride or (stride=1)
- If data is accessed with **non-unit stride**:
  - Much more data cache line references needed



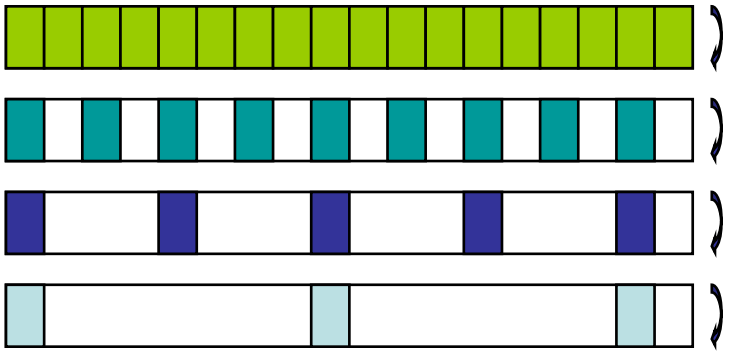
Experimental Study of Memory

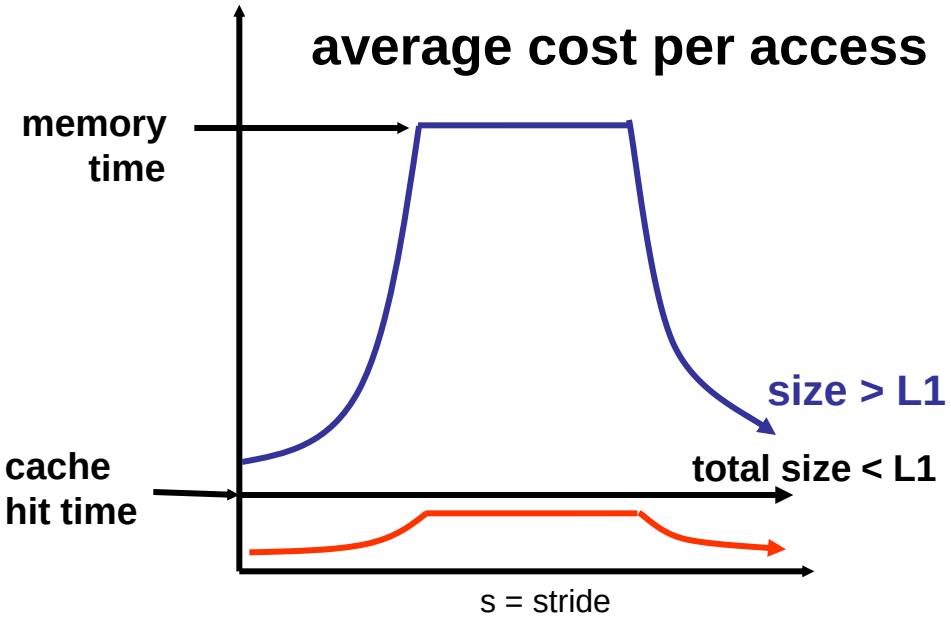
- Micro benchmark for memory system performance




```
for array A of length L from 4KB to 8MB by 2x
   for stride s from 4 Bytes (1 word) to L/2 by 2x
          // time the following loop
          // repeat many times and average
          for i from 0 to L by s
              load and store A[i] from memory (4 Bytes)
```

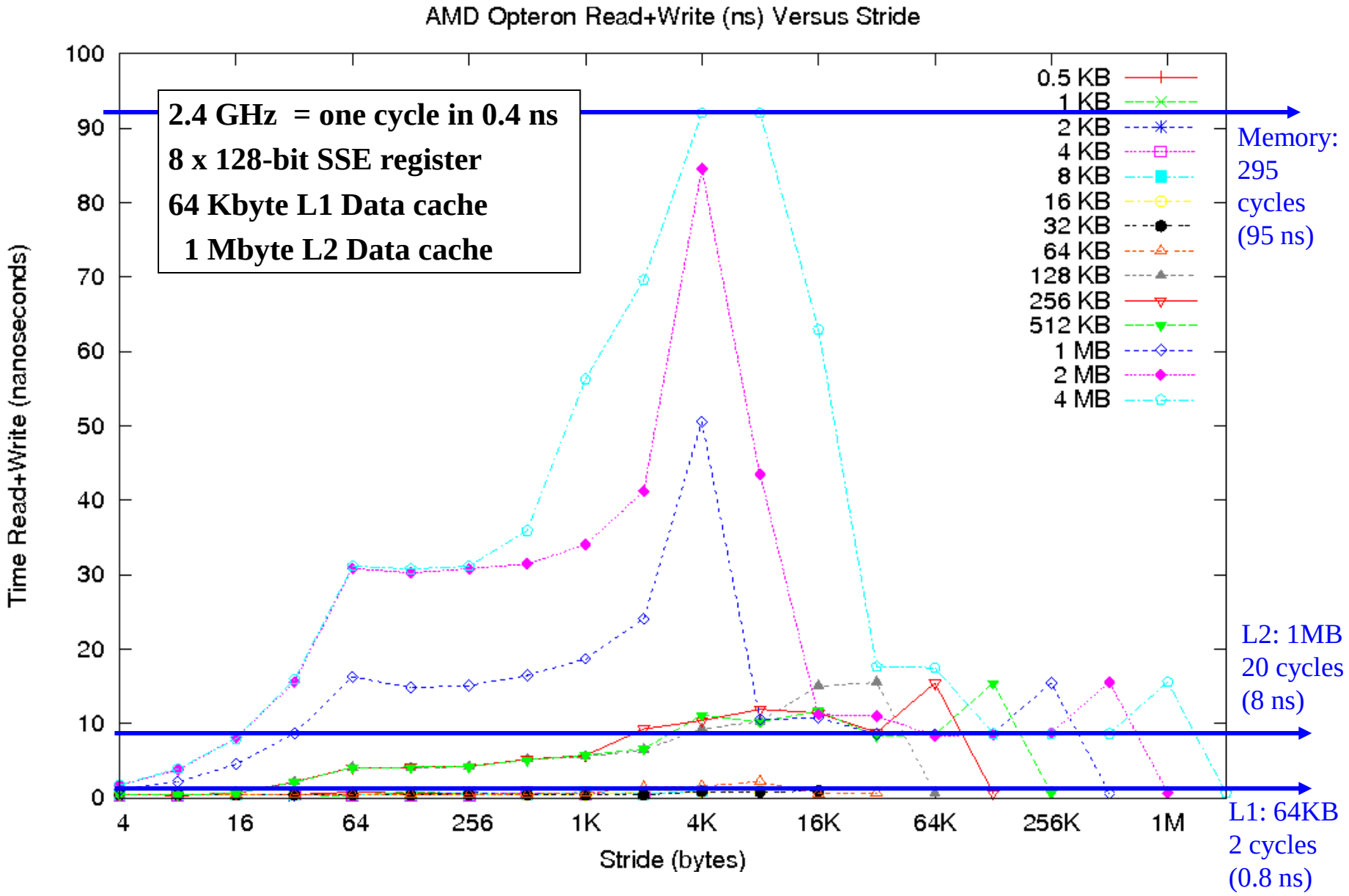
Experimental Study of Memory: What to expect





- Consider the average cost per load
  - **Plot one line** for each array length, **time vs. stride**
  - **Small stride is best**: if cache line holds 4 words, at most ¼ miss
  - If array is smaller than a given cache, all those accesses will hit (after the first run, which is negligible for large enough runs)
  - Picture assumes only one level of cache
  - Values have gotten more difficult to measure on modern procs

Experimental Study of Memory: AMD Opteron Processor



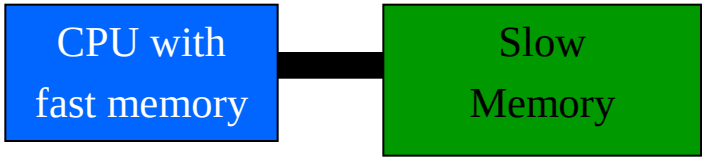
Outline

- Intro memory hierarchy
- Cache mappings
- Memory access
- Case study: Matrix-matrix multiplication

Case study: Matrix-Matrix Multiplication

- Case study: **Matrix-Matrix Multiplication**
  - Use of performance models to understand performance
  - Simple cache model
- Why Matrix Multiplication?
  - Appears in many linear algebra algorithms
- Optimization ideas can be used in other problems
- The best case for optimization payoffs
- The most-studied algorithm in high performance computing

Case study: Matrix Multiplication



- Assume just **2 levels** in the **memory** hierarchy, **fast** and **slow**
- All data initially in slow memory
  - m = number of memory elements (words) moved between fast and slow memory
  - t_m = time per **slow memory** operation
  - f = number of arithmetic operations
  - t_f = time per **arithmetic operation** in **fast memory** << t_m
  - q = f / m average number of flops per slow memory access
- Minimum possible time = f* t_f when all data in fast memory
- Actual time
  - f * t_f + m * t_m =f * t_f * (1 + t_m/t_f * m/f) = f * t_f * (1 + t_m/t_f * 1/q)
- **Larger q** means time closer to minimum f * t_f
  - **q ≥ t_m/t_f** needed to get at least half of peak speed

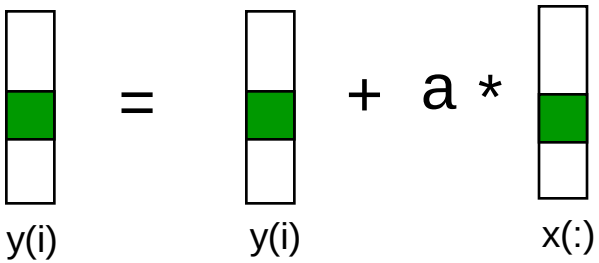
Computational Intensity: Key to algorithm efficiency

Machine Balance:Key to machine efficiency

Warm up: Vector-vector operation

- **Assumption:** Fast memory large enough to store 2 vectors, “a” is a scalar, x, y are vectors

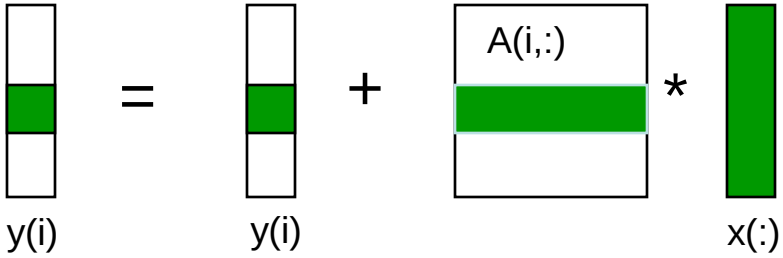
```
read x(1:n) into fast memory
read y(1:n) into fast memory
for j = 1:n
   y(i) = y(j) + a*x(i)
write y(1:n) back into slow memory
```



- m = number of slow memory accesses → m = 3n
- f = number of floating point operations → 2n
- q = average number of flops per slow memory access = f / m = 2/3
- Vector-vector addition is dominated by slow memory access

Warm up: Matrix-vector multiplication

```
// implements y = y + A*x
for i = 1:n
        for j = 1:n
                y(i) = y(i) + A(i,j)*x(j)
```



Warm up: Matrix-vector multiplication

Assumption: Fast memory large enough to store 2 vectors y and x and one row of matrix A

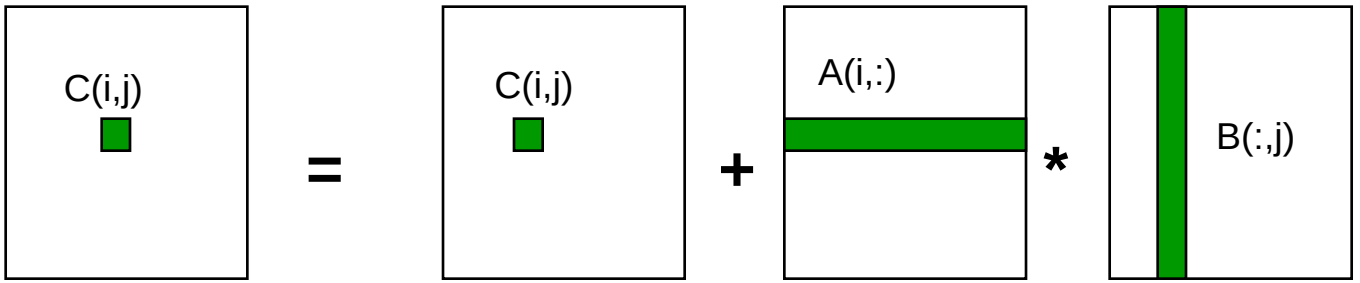
```
read x(1:n) into fast memory
read y(1:n) into fast memory
for i = 1:n
       read row i of A into fast memory
       for j = 1:n
          y(i) = y(i) + A(i,j)*x(j)
write y(1:n) back to slow memory
```

- m = number of slow memory refs → $m = 3n + n^2$
- f = number of arithmetic operations → $f = 2n^2$
- q = f / m ~= 2
- Matrix-vector multiplication is limited by slow memory speed

Naive Matrix-Matrix Multiplication

```
// computes C = C + A*B
for i = 1 to n
   for j = 1 to n
       for k = 1 to n
              c(i,j) = c(i,j) + a(i,k) * b(k,j)
```

- Algorithm has $2*n^3 = O(n^3)$ Flops and operates on $4*n^2$ words of memory → q potentially as **large** as $2*n^3 / 4*n^2 =$ **O(n)**



Naive Matrix-Matrix Multiplication

Assumption: Fast memory large enough to store one row of the matrices A, B and C.

```
// implements C = C + A*B
for i = 1 to n
read row i of A into fast memory
    for j = 1 to n
            read scalar C(i,j) into fast memory
            read column j of B into fast memory
            for k = 1 to n
                    C(i,j) = C(i,j) + A(i,k) * B(k,j)
            write scalar C(i,j) back to slow memory
```

- Number of **slow memory** references on unblocked matrix multiply

  $\mathbf{m = n^3}$ to read each column of B n^2 times

  $\mathbf{+ n^2}$ to read each row of A once

  $\mathbf{+ 2n^2}$ to read and write each element of C once

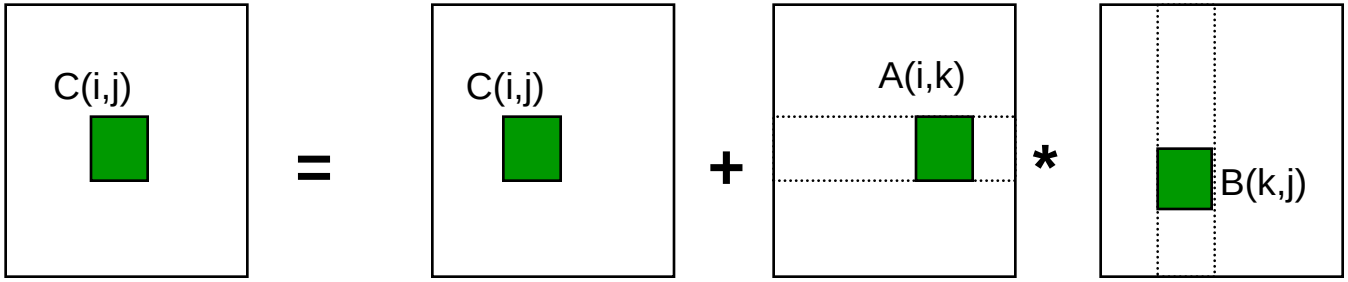
  $\mathbf{= n^3 + 3n^2}$

  So $q = f / m = 2n^3 / (n^3 + 3n^2) \sim= 2$ for large n, **no improvement** over matrix-vector multiply, but on previous slide it was O(n)!

Blocked Matrix Multiply

- **Assumption:** Fast memory large enough to store small subblocks row of the matrices A, B and C.

```
// Consider A,B,C to be n-by-n matrix viewed as
// N-by-N matrices of b-by-b subblocks where
// b=n/N is called the block size
for I = 1 to N
    for J = 1 to N
        read block C(I,J) into fast memory
        for K = 1 to N
        read block A(I,K) into fast memory
        read block B(K,J) into fast memory
        // do a matrix multiply on blocks
        C(I,J) = C(I,J) + A(I,K) * B(K,J)
        write block C(I,J) back to slow memory
```



Blocked Matrix Multiply

- Recall:
  - m is amount **memory traffic** between slow and fast memory
  - matrix has nxn elements, and NxN blocks each of size bxb
  - f is **number of floating point** operations, **$2n^3$** for this problem
  - q = f / m is the measure of algorithm efficiency in the memory model
- So:

  m = $N*n^2$ read each block of B N^3 times ($N^3*b^2 = N^3*(n/N)^2 = N*n^2$)

  + $N*n^2$ read each block of A N^3 times

  + $2n^2$ read and write each block of C once

  = $(2N + 2) * n^2$

- **The computational intensity is**

  $q = f / m = 2n^3 / ((2N + 2) * n^2) \sim= n / N = b$ for large n

  So we can improve performance by **increasing** the blocksize **(cachesize)** b and can be much faster than matrix-vector multiply (q=2)

Using Analysis to Understand Machines

The blocked algorithm has **computational intensity q ~= b**

- The larger the block size, the more efficient our algorithm will be
- Limit: All three blocks from A,B,C must fit in fast memory (cache), so we cannot make these blocks arbitrarily large
- Assume your fast memory has size M_{fast}

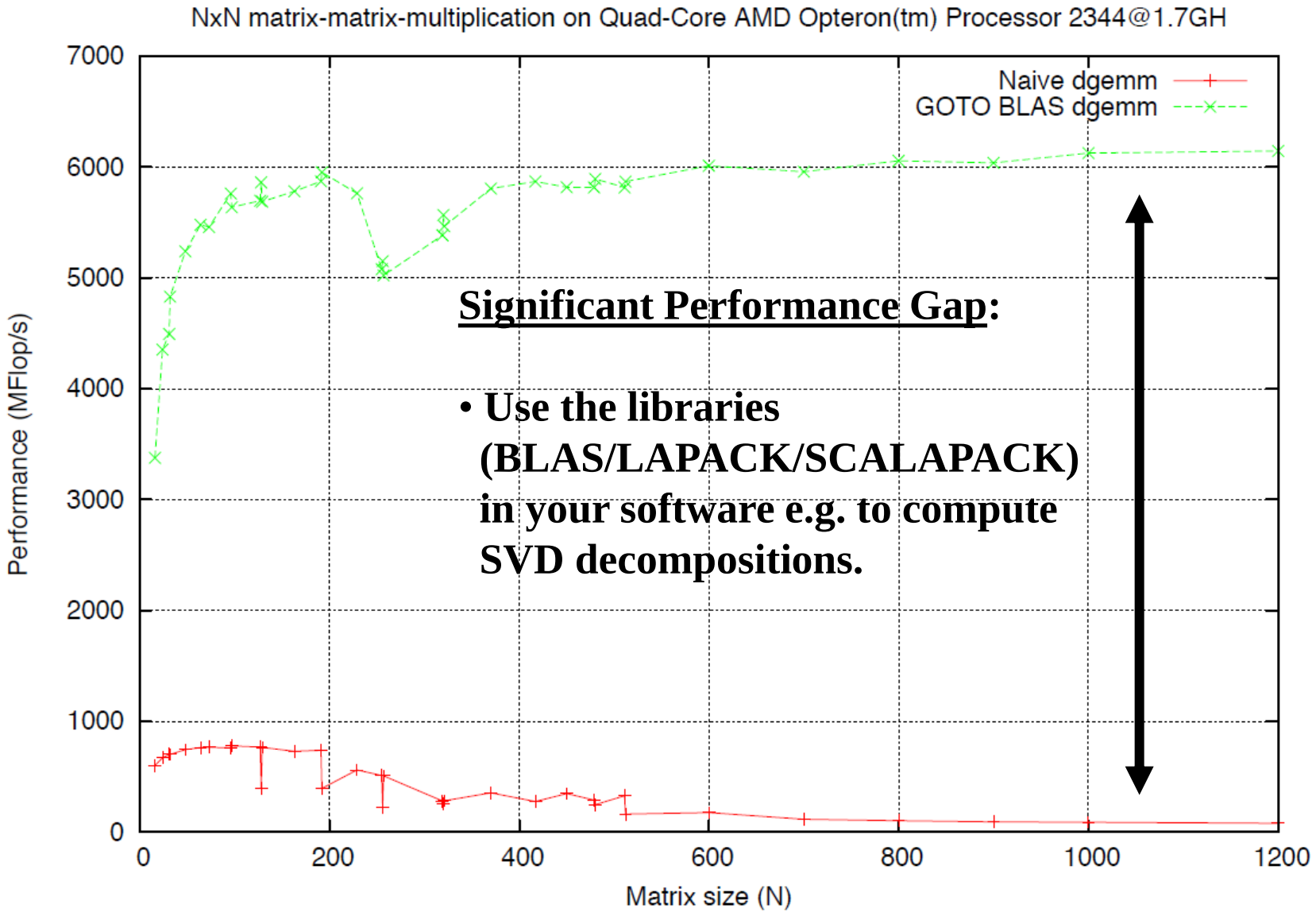
  $3b^2 <= M_{fast}$, so $q \sim= b <= sqrt(M_{fast}/3)$

- What if more levels of memory hierarchy?
  - Apply blocking recursively, once per level

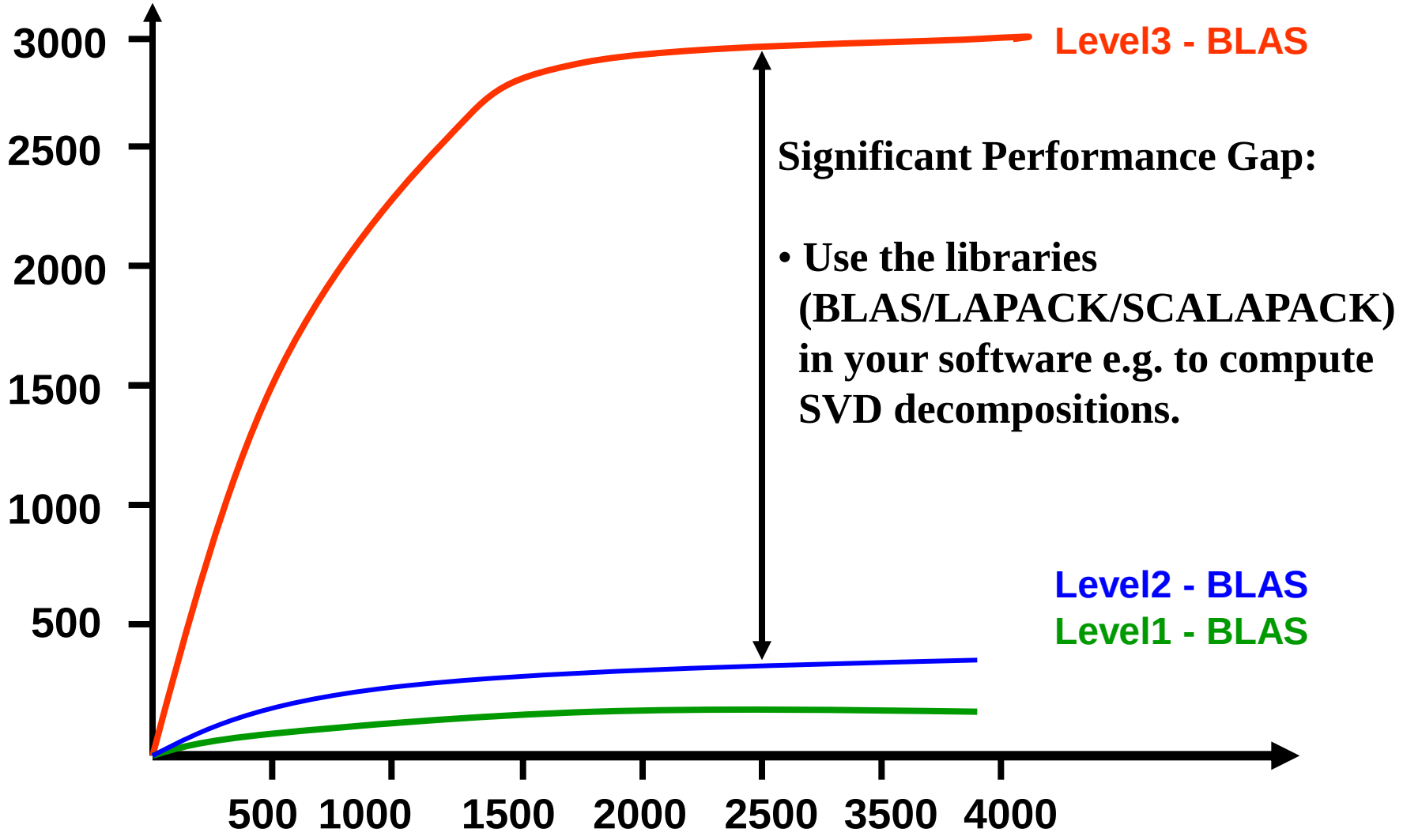
Use optimized libraries for matrix-/vector operations

- Industry library standard interface: **www.netlib.org/blas**
- Vendors, others supply highly optimized implementations
  - **BLAS – Basic Linear Algebra Subroutines**
    - **BLAS1 (1970s):**
      - vector operations: dot product, saxpy (y=a*x+y), etc
      - m=2*n, f=2*n, q ~1 or less
    - **BLAS2 (mid 1980s)**
      - matrix-vector operations: matrix vector multiply, etc
      - $m=n^2$, $f=2*n^2$, q~2, less overhead
      - somewhat faster than BLAS1
    - **BLAS3 (late 1980s/earlier 90)**
      - matrix-matrix operations: matrix-matrix multiply, etc
      - $m <= 3n^2$, $f=O(n^3)$, so q=f/m can possibly be as large as n, so BLAS3 is potentially much faster than BLAS2
- Good algorithms used BLAS3 when possible (LAPACK & ScaLAPACK)
- **LAPACK: Linear Algebra Package**
  ScaLAPACK: Scalable Linear Algebra Packages (parallel)

Real Matrix-Matrix Performance on ROSA Cluster



BLAS Performance on Xeon SSE2 with 1.5 GHz



A last note on the complexity of matrix multiply

- The traditional algorithm (with or without blocking) has $O(n^3)$ flops
- n x n Matrix = 2 x 2 Matrix with n/2 x n/2 block matrices

```
Let M =  | m11 m12  | =  | a11 a12 |   *  | b11 b12 |
         | m21 m22  |=   | a21 a22 |      | b21 b22 |
         m11 = a11*b11 + a12*b21,  m12 = a11*b12 + a12*b22
         m21 = a21*b11 + a22*b21,  m22 = a21*b12 + a22*b22
```

- Reduction to four sub problems with
  - 2 multiplications von n/2 x n/2 matrices
  - 1 addition von n/2 x n/2 matrices
  - Total: 8 Mult, 4 Adds
- **Complexity:** $T(n) = 8 * T(n/2) + \theta(n^2)$

  #subproblems / size of subproblems / complexity for addition

  $= 8 (n/2)^3 = n^{\log_2 8} = n^3$

  $= \theta(n^3)$ **No reduction!**

A last note on the complexity of matrix multiply

- Upper complexity bound for matrix-matrix multiply?
- Lower complexity bound for matrix-matrix multiply?

Complexity of Strassen's Matrix Multiply

- Strassen discovered an algorithm with asymptotically lower flops
  - $O(n^{2.81})$
- Consider a 2x2 matrix multiply, normally takes 8 multiplies, 4 adds
  - Strassen does it with **7 multiplies and 18 adds**

```
Let M =  | m11 m12  | =  | a11 a12 |   *  | b11 b12 |
         | m21 m22  |=   | a21 a22 |      | b21 b22 |
Let p1 = (a12 - a22) * (b21 + b22)     p5 = a11 * (b12 - b22)
    p2 = (a11 + a22) * (b11 + b22)     p6 = a22 * (b21 - b11)
    p3 = (a11 - a21) * (b11 + b12)     p7 = (a21 + a22) * b11
    p4 = (a11 + a12) * b22
Then  m11 = p1 + p2 - p4 + p6
      m12 = p4 + p5
      m21 = p6 + p7
      m22 = p2 - p3 + p5 - p7
```

Complexity:

$T(n) = 7 * T(n/2) + \theta(n^2)$

#subproblems / size of subproblems / complexity for addition

$= 7 (n/2)^3 = n^{\log_2 7} + \theta(n^2)$

$= \theta(n^{2.81})$

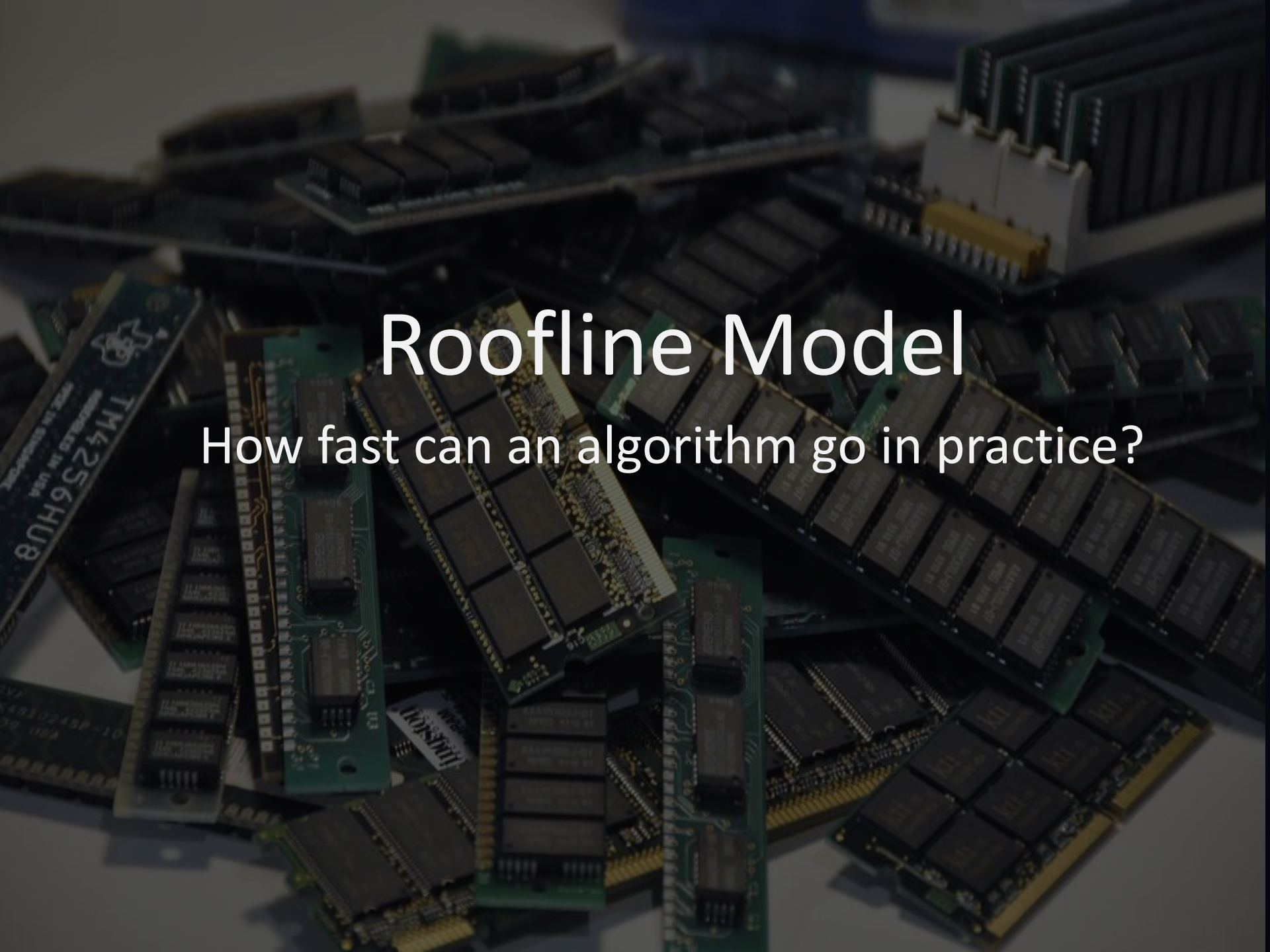
A last note on the complexity of matrix multiply

- World's record was O (n 2.375477...)
  Coppersmith & Winograd, 1987
- New Record! Exponent reduced to 2.3729269
  Virginia Vassilevska Williams, UC Berkeley & Stanford, 2011
- Newer Record! Exponent reduced to 2.3728639
  Francois Le Gall, 2014
- Latest Record! Exponent reduced to 2.371552

  Virginia Vassilevska Williams, Yinzhan Xu, Zixuan Xu, Renfei Zhou, 2023

Outline

- Introduction memory hierarchy
- Memory access
- Case study: Matrix-matrix multiplication
- Berkeley Roofline Model of Performance



Roofline Model

How fast can an algorithm go in practice?

What’s in the Roofline Model?

Three pieces: 2 for machine and 1 for application

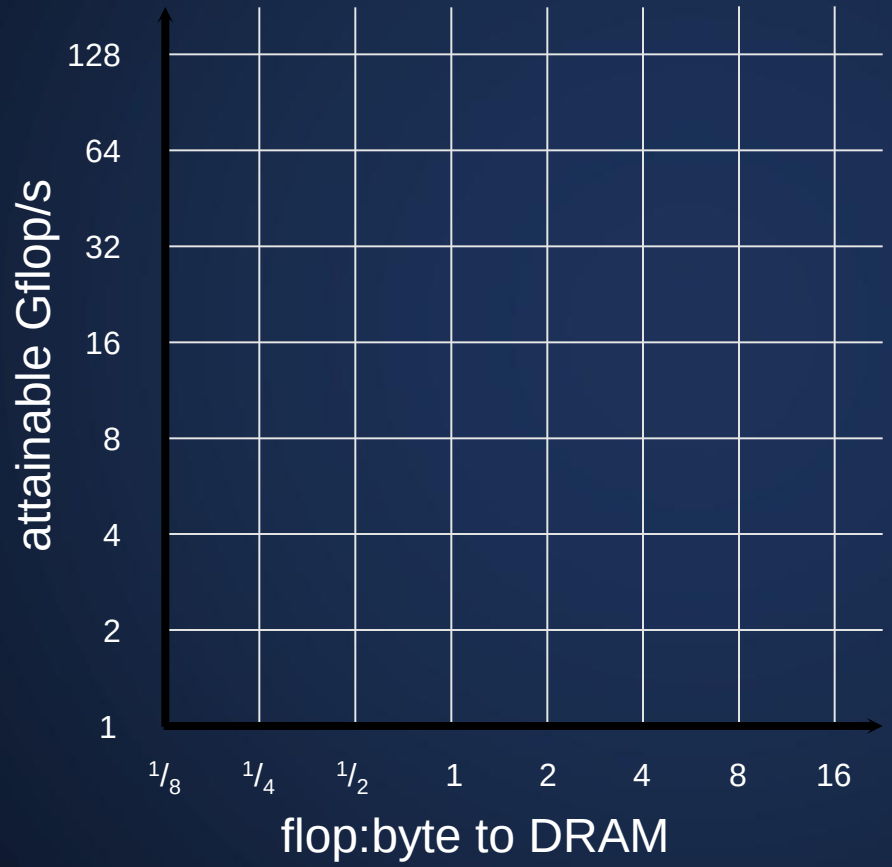
What's in the Roofline Model?

Three pieces: 2 for machine and 1 for application

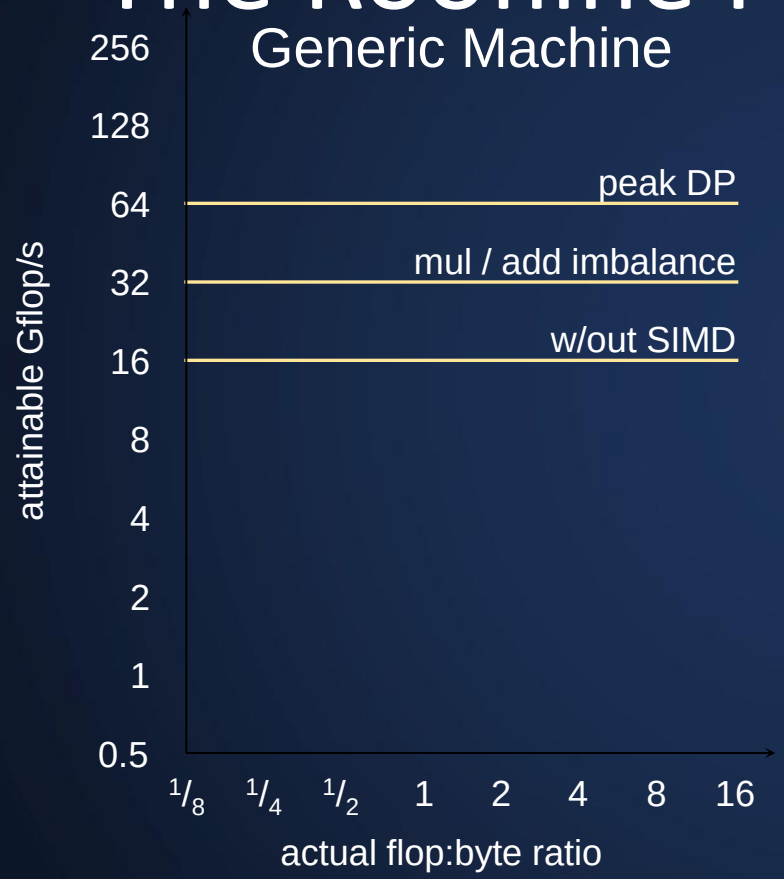
- Arithmetic performance (flops/sec)
  - Clock Speed and Parallelism (SIMD, Multicore)
- Memory bandwidth (bytes /sec)
- Computational (Arithmetic) Intensity
  - Application balances (flops/word or flops/byte)

Application Machine

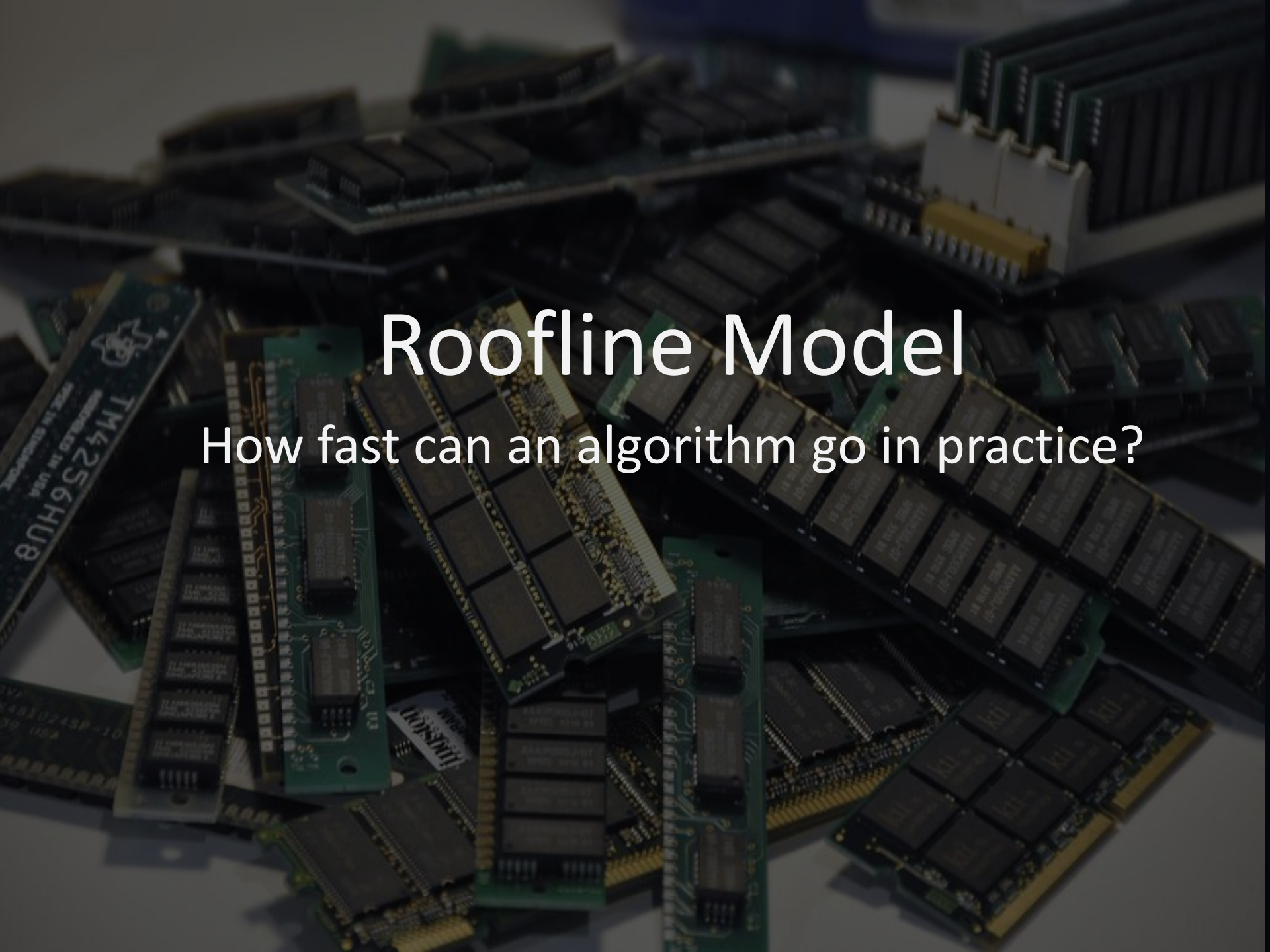
The Roofline Performance Model



The Roofline Performance Model



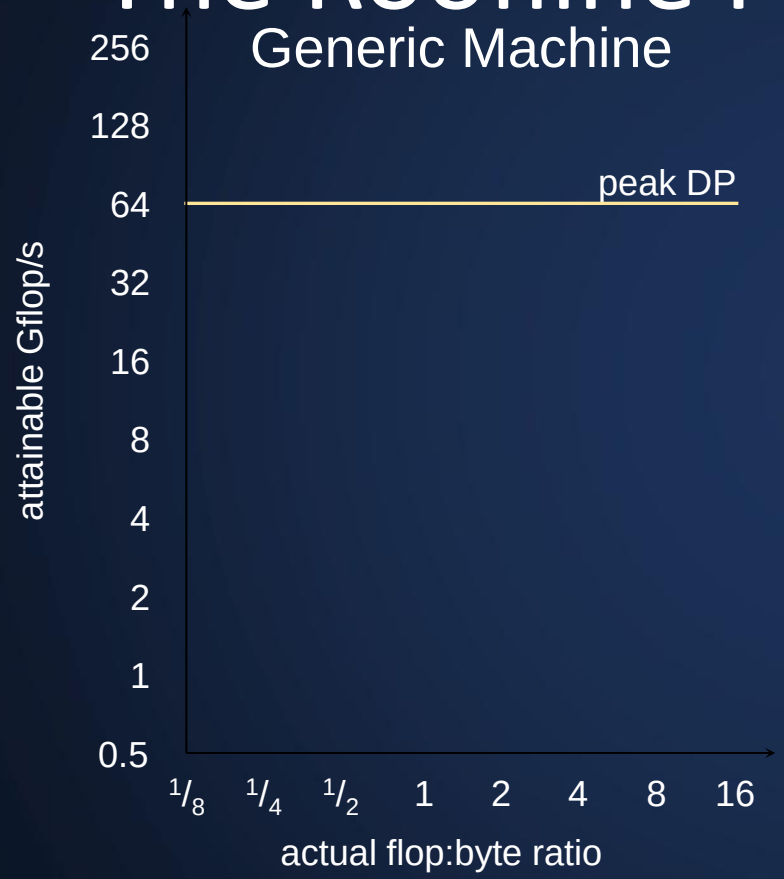
- Top of the roof is the peak compute rate
- No FMA, no SIMD, will lower what is attainable



Roofline Model

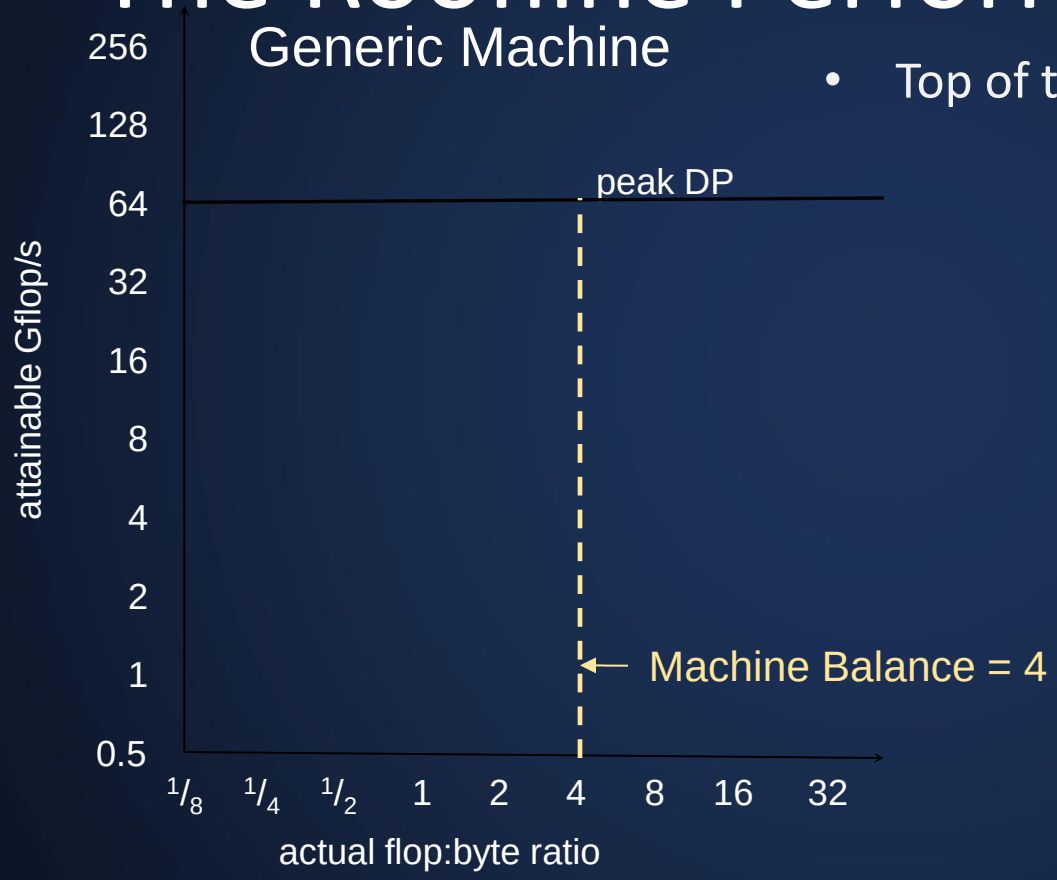
How fast can an algorithm go in practice?

The Roofline Performance Model



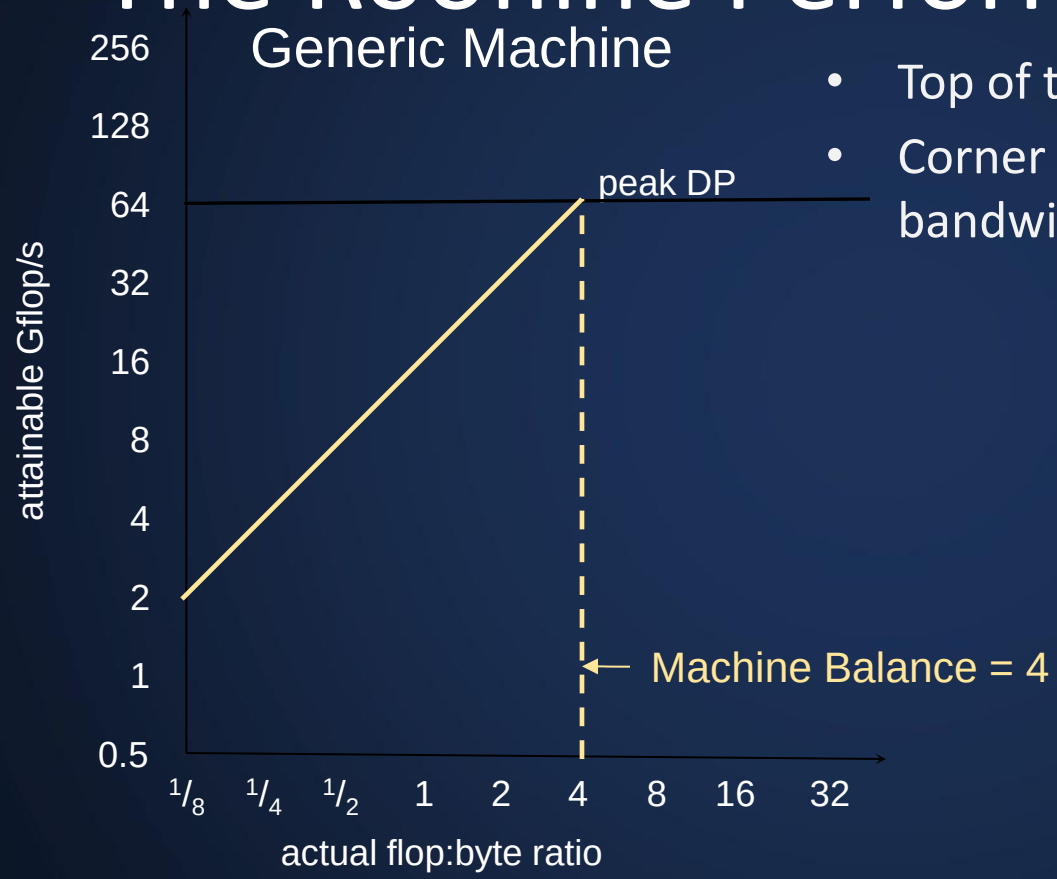
- Top of the roof is the peak compute rate

The Roofline Performance Model



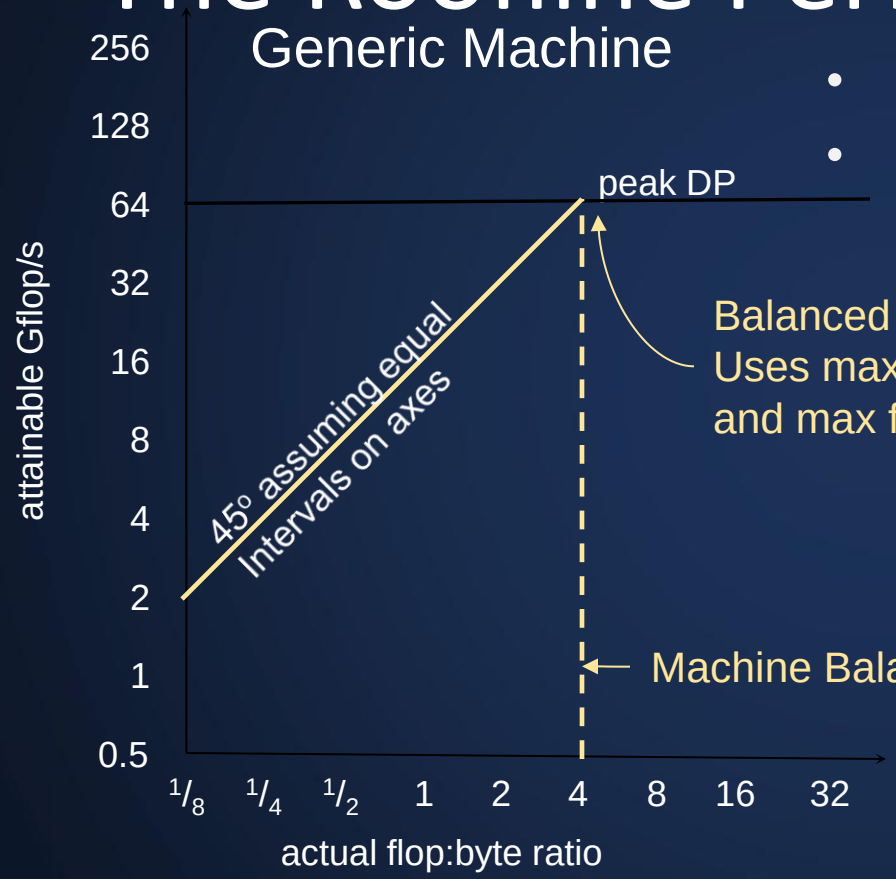
- Top of the roof is the peak compute rate

The Roofline Performance Model



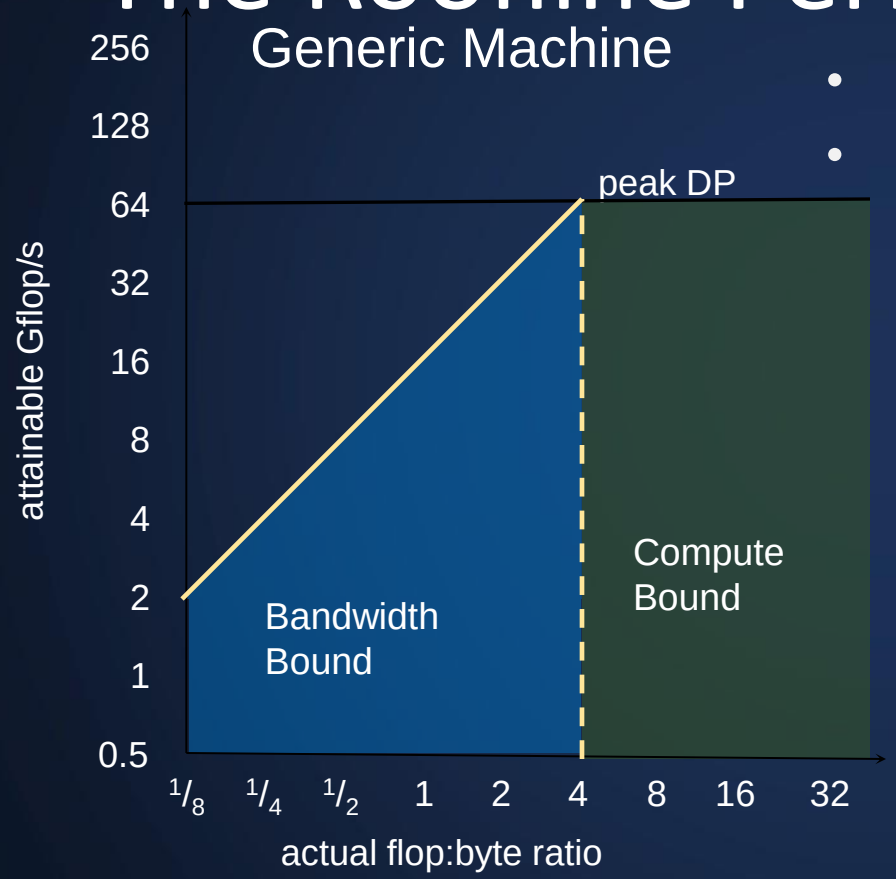
- Top of the roof is the peak compute rate
- Corner of roof depends on memory bandwidth

The Roofline Performance Model



- Top of the roof is the peak compute rate
- Corner of roof depends on memory bandwidth

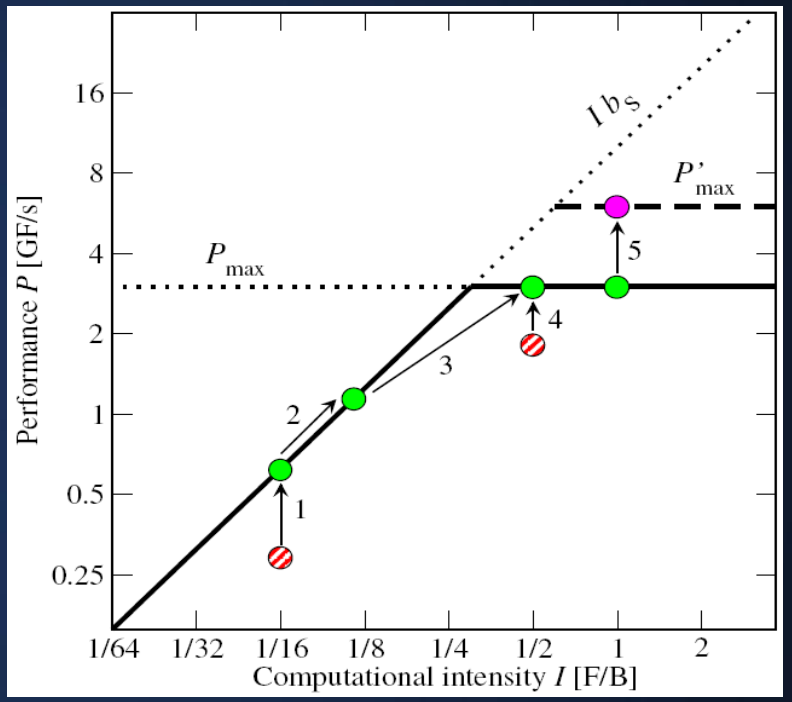
The Roofline Performance Model



- Top of the roof is the peak compute rate
- Corner of roof depends on memory bandwidth

Typical code optimizations in the Roofline Model

1. Hit the BW bottleneck by good serial code
   (e.g., Java/Ninja -> C)
2. Increase intensity to make better use of BW bottleneck
   (e.g., loop blocking as in block matrix multipy)
3. Increase intensity and go from memory-bound to cmpute-bound
   (e.g., temporal blocking)
4. Hit the core bottleneck by good serial code
   (e.g., **-fno-alias**)
5. Shift P_{max} by accessing additional hardware features or using a different algorithm/implementation
   (e.g., accelerators, GPUs)



[Project 1] Memory Hierarchies and Matrix-Matrix Multiplications

Università della Svizzera italiana

Institute of Computing CI

High-Performance Computing, Fall 2024
Lecturer: Dr. A. Eftekhari, Prof. O. Schenk
Assistants: M. Lechekhab, D. Vega
D. Santarsiero, J. Palumbo, I. Ecevit

Project 1 – Performance Characteristics, Memory Hierarchies and Matrix-Matrix Multiplications

Due date: 9 October 2024 at 23:59 (See iCorsi for updates)

The first part of this project will introduce you to using the Rosa cluster and collecting some performance characteristics. The second part of the project will be about the optimization of general matrix multiplication. Both project parts will be single-threaded. Parallel programming will be the subject of the forthcoming projects.

1. Rosa warm-up [5 points]

Review the Rosa documentation and the Slurm tutorial to gain an understanding of its features and usage. Answer briefly the following questions in your report:

1. What is the module system and how do you use it?
2. What is Slurm and its intended function?
3. Write a simple "Hello World" `C/C++` program which prints the host name of the machine on which the program is running.
4. Write a batch script which runs your program on one node. The batch script should be able to target nodes with specific CPUs with different memories. You can obtain the information on available nodes using the command `sinfo`. **Hint**: Slurm hast the option `--constraint` to select some nodes with specific features.
5. Write another batch script which runs your program on two nodes. Check that the Slurm output file (`inputs/slurm-*.out` by default) contains two different host names.

Include the source code, batch scripts, and Slurm output files in your submission (see the notes at the end of the document).

2. Performance characteristics [30 points]

Università della Svizzera italiana

Institute of Computing CI

High-Performance Computing, Fall 2024
Lecturer: Dr. A. Eftekhari, Prof. O. Schenk
Assistants: M. Lechekhab, D. Vega
D. Santarsiero, J. Palumbo, I. Ecevit

Project 1 – Performance Characteristics, Memory Hierarchies and Matrix-Matrix Multiplications

Due date: 9 October 2024 at 23:59 (See iCorsi for updates)

The first part of this project will introduce you to using the Rosa cluster and collecting some performance characteristics. The second part of the project will be about the optimization of general matrix multiplication. Both project parts will be single-threaded. Parallel programming will be the subject of the forthcoming projects.

1. Rosa warm-up [5 points]

Review the Rosa documentation and the Slurm tutorial to gain an understanding of its features and usage. Answer briefly the following questions in your report:

1. What is the module system and how do you use it?
2. What is Slurm and its intended function?
3. Write a simple "Hello World" `C`/`C++` program which prints the host name of the machine on which the program is running.
4. Write a batch script which runs your program on one node. The batch script should be able to target nodes with specific CPUs with different memories. You can obtain the information on available nodes using the command `sinfo`. **Hint**: Slurm hast the option `--constraint` to select some nodes with specific features.
5. Write another batch script which runs your program on two nodes. Check that the Slurm output file (`inputs/slurm-*.out` by default) contains two different host names.

Include the source code, batch scripts, and Slurm output files in your submission (see the notes at the end of the document).

2. Performance characteristics [30 points]

In the realm of HPC, understanding the pproject1.pdfperformance limits of computing systems, particularly in relation to the memory system and floating-point operations, is crucial for designing efficient algorithms and optimizing their implementations on (a particular) hardware. The performance characteristics and concept of lightspeed estimates comes into play when considering the physical constraints that limit data transfer rates and computation speeds within a computing system. These constraints are not only dictated by the raw processing power of the CPU but also by the latency and throughput of the memory hierarchy. It is essential to grasp these concepts in order to appreciate the theoretical and practical limits of computing, including how data locality and memory bandwidth impact the achievable performance. This knowledge is fundamental in pushing the boundaries of what is computationally possible while being mindful of the inherent practical limitations posed by physics and current technology.
However, determining the performance characteristics of computing systems is often a complex and nuanced task that requires a multi-faceted approach. Manufacturer documentation provides a baseline of theoretical capabilities and specifications, but real-world performance can deviate significantly from these idealized scenarios due to a myriad of factors. Web resources, including insights from HPC centers, offer valuable empirical data and analyses that reflect the

performance characteristics. Yet another source of information is experimentation in the form of low-level benchmarking. Thus, a combination of scrutinizing manufacturer documentation, leveraging the wealth of information available from HPC centers and other web resources, and conducting methodical benchmarking experproject1.pdfiments is essential for a comprehensive understanding of computing system performance.
The goal of this task is to collect such performance characteristics and lightspeed estimates. As a refresher (or a concise introduction), we *strongly* encourage you to read the article and appendix [10], and Chapters One and Three of the book [3][1].

2.1. Peak performance

In this task, we ask you to compute the core, CPU, node and cluster peak performance for the Rosa nodes. Please detail your computation and all sources in your report. In the following, we provide some context and give a step-by-step guide to complete and report the task.
In scientific computing, the focus often lies on floating-point (FP) data, typically utilizing *double precision*. The rate at which the CPU's floating-point unit (FPU) can produce results for multiplication and addition operations is quantified in terms of *floating-point operations per second* (Flops/s, or simply FLOPS). The maximum rate at which a system is capable of executing Flops/s is the so-called (theoretical[2]) *peak (FP) performance*. The peak performance P_{core} of a single core is computed as follows

$$P_{\text{core}} = n_{\text{super}} \times n_{\text{FMA}} \times n_{\text{SIMD}} \times f$$

where

- n_{super} is the superscalarity factor: the number of FP operations the FPU of a core can execute in parallel within a single clock cycle.
- n_{FMA} is the fused multiply-add factor: $n_{\text{FMA}} = 2$ if the FPU supports the FMA instruction (enabling it to perform a multiplication and an addition in a single operation), or $n_{\text{FMA}} = 1$ if the FPU does not support FMA (requiring separate instructions for multiplication and addition).
- n_{SIMD} is the SIMD (Single Instruction, Multiple Data) factor: the number of doubles the FPU can process concurrently with a single SIMD instruction (i.e., the width of the SIMD registers in units of doubles).
- f is the (base) clock frequency (modern multi-core CPUs may dynamically increase their clock frequency to make use of Thermal Design Power (TDP) more efficiently if fewer cores are active).

The peak performance P_{CPU} of a CPU is then

$$P_{\text{CPU}} = n_{\text{cores}} \times P_{\text{core}},$$

where n_{cores} is the number of *physical* cores[3]. The peak performance P_{node} of a node with n_{sockets} identical CPUs

$$P_{\text{node}} = n_{\text{sockets}} \times P_{\text{CPU}}.$$

Finally, the peak performance P_{cluster} of a cluster consisting of n_{nodes} identical nodes is

$$P_{\text{cluster}} = n_{\text{nodes}} \times P_{\text{node}}.$$

[1] To gain access, you have to be within the USI network, e.g., by using the VPN.
[2] There is a distinction between theoretical and measured peak performance, but we will ignore it for the time being (see here)
[3] As opposed to *logical* cores often present on modern CPUs with *threading* capabilities, such as Hyper-threading (HT) for Intel CPU or Simultaneous Multithreading (SMT) for AMD.

Let's compute the peak performance of the Euler III (2016-2022) nodes as an example. The system comprised 1215 nodes (i.e., $n_{\text{nodes}} = 1215$), each equipped with a single Intel Xeon E3-1585Lv5 CPU (i.e., $n_{\text{sockets}} = 1$). According to the Euler webpage, which conveniently provides detailed specifications, each of these CPUs has four cores (i.e., $n_{\text{cores}} = 4$). Additionally, the link to Intel's ARK website offers further technical details, revealing that the CPU operates at a base clock frequency of $f = 3.00$ GHz and supports AVX2 SIMD instructions with 256-bit wide vector registers. This setup allows for processing four 64-bit double-precision FP numbers simultaneously, leading to $n_{\text{SIMD}} = 4$.
The remaining factors to consider for calculating peak performance are the superscalarity and FMA factors. However, these details are not as readily available and typically require a deep dive into technical documents provided by the CPU manufacturer or from technology review sites and academic research. In this specific instance, valuable information is found in Figure 2-9 of Intel's Optimization Reference Manual Volume 1 (around p. 77 in the PDF). This figure indicates that Ports 0 and 1 each can perform one vector FMA operation (i.e., $n_{\text{FMA}} = 2$), establishing the superscalarity factor $n_{\text{super}} = 2$. Hence, we can compute the (theoretical) peak performance of the Euler III nodes

$$\begin{aligned} P_{\text{core}} &= 2 \times 2 \times 4 \times 3 \text{ GHz} = 48 \text{ GFlops/s}, \\ P_{\text{CPU}} &= 4 \times P_{\text{core}} = 192 \text{ GFlops/s}, \\ P_{\text{node}} &= 1 \times P_{\text{CPU}} = 192 \text{ GFlops/s}, \\ P_{\text{Euler III}} &= 233'280 \text{ GFlops/s} = 233.28 \text{ TFlops/s} \end{aligned}$$

Here are some useful hints for the task:

- Rosa cluster reveals that each node has two Intel Xeon E5-2650 CPUs and the Intel's ARK website provides general specifications. From Intel's specifications you get that the CPU is part of the Intel Xeon E5 collection, from which you deduce that it is based on fourth-generation core microarchitecture (codenamed "Haswell").
- The cpu-world.com website will provide you with the SIMD factor.
- The uops.info website will provide you with the superscalarity factor. Instead, you could consult Intel's documentation for this factor. However, this approach is slightly more complex and can be avoided.
- This Intel's Optimization Reference Manual will then provide you with the FMA factor.
- Remember that ThroughPut (TP) is defined by how many cycles it takes to execute one instruction (i.e., Cycles Per Instruction (CPI)).

In general (beyond this project and course), the following resources are valuable:

- Intel's software and optimization manuals: The easiest way to find them is to search for "Software Developer's Manual" or "Optimization Reference Manual" in your favorite web search engine.
- AMD's software and optimization manuals: The easiest way to find them is to search for "AMD documentation hub" in your favorite web search engine. Once on the hub, search for "Software optimization guide" specific to the CPU model you are interested.
- Intel Intrinsics Guide
- Agner Fog's software optimization resources: `https://www.agner.org/optimize/`
- Latency and throughput, listing: `https://uops.info/`.
- `https://en.wikichip.org`: Can be a useful resource (despite the ads).

2.2. Memory Hierarchies

2.2.1. Cache and main memory size

The next task is to identify the parameters of the memory hierarchy on a node of the Rosa cluster . Follow the step-by-step guide provided below and detail your results in your project report.
To guide you, we show how this can be achieved for a Rosa login node. A useful tool to determine the CPU architecture is

```
[user@icslogin]$ lscpu
```

From this we obtain the CPU model (Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz) featuring ten cores and some basic information on the memory hierarchy.
To obtain the total available main memory, one option is to look at the

```
[user@icslogin]$ cat /proc/meminfo
```

In summary, we have collected so far the information listed in Table 2. Since this is a multi-core CPU, it is valuable to obtain more information on how the memory hierarchy is organized and shared among the cores. A (possible; there are others) tool for this is provided by `hwloc`. In particular, the following commands will give us the desired information:

```
[user@icslogin]$ module load hwloc
[user@icslogin]$ hwloc-ls
```

From the output, we conclude that each of the ten cores has its own L1 instruction (L1i) and data (L1d) caches as well as an L2 cache. The L3 cache, along with the main memory, are shared among all cores. This observation has been noted in the caption of Table 2 to provide a complete overview. A graphical representation of the memory hierarchy is displayed in Fig. 3. To obtain such a figure, first ask `hwloc-ls` to output in the "fig" format

```
[user@icslogin]$ hwloc-ls --whole-system --no-io -f --of fig XEON\_E5-2650.fig
```

You can open the file using Xfig on your machine, you can convert `XEON_E5-2650.fig` by copying it on your local machine and converting it to a PDF file:

```
$ scp username@rosa.usi.ch:~/XEON_E5-2650.fig /your/local/folder
$ fig2dev XEON_E5-2650.fig XEON_E5-2650.pdf
```

For more information on these commands and their options, please have a look at there man pages (e.g., `man lscpu`).

Main memory	23 GB
L3 cache	25 MB
L2 cache	256 KB
L1 cache	32 KB

Table 1: Memory hierarchy of a Rosa login node with an Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz. Each core has its own L1 and L2 cache, while L3 is shared among all the ten cores.

2.3. Bandwidth: STREAM benchmark

Now that we have an overview of the peak performance, as well as the topology and size of the memory system, another key performance characteristic to consider is the speed of the memory system. The speed of the memory system is quantitatively measured in terms of its bandwidth, which indicates the amount of data that can be transferred within a specific time frame. McCalpin's STREAM benchmark [7] is a widely recognized tool used for measuring the memory bandwidth of a CPU. It consists of four operations or kernel — Copy, Scale, Add, and Triad — that evaluate

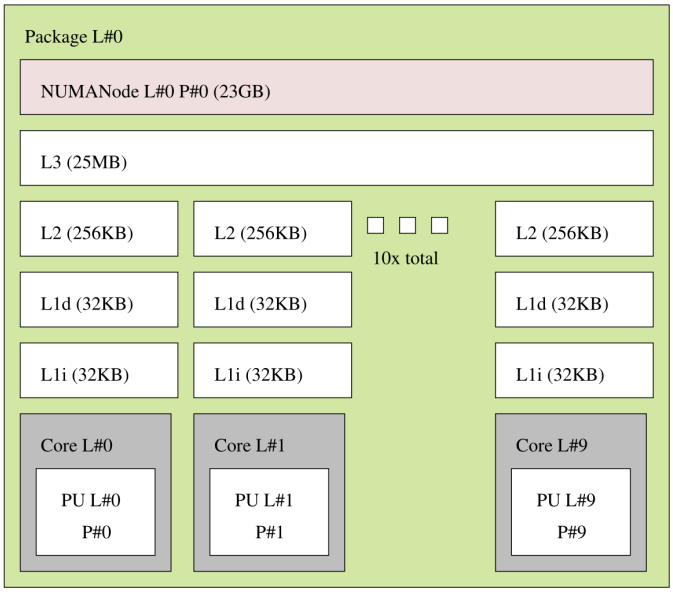


Figure 1: Schematic of a Rosa login node with an Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz.

the sustainable memory bandwidth and the corresponding computation rate for simple vector kernels (see, e.g, [3, Sec. 3.1.2]).
In this task, we ask you to run the STREAM benchmark on the Rosa nodes in order to measure the single-core bandwidth. Find below a step-by-step guide:

1. Download the STREAM Benchmark: it can be obtained from the official website in the source code directory. You will need `stream.c` and `mysecond.c`.

2. In `stream.c`, you will find precise instructions on how to compile and run the benchmark. It is particularly important to tune the sizes of the arrays (used in the Copy, Scale, Add, and Triad kernels) to match the cache sizes of the system of interest: Each array must be at least four times the size of the last cache level. This is set by the `STREAM_ARRAY_SIZE` preprocessor macro and two examples are given in the `stream.c`. For the CPU studied in the example of 2.2.1, we would then

```
module load gcc                                    # load a compiler (gcc/13.2.0)
gcc -O3 -march=native -DSTREAM_TYPE=double -DSTREAM_ARRAY_SIZE=128000000 /
-DNTIMES=20 stream.c -o stream_c.exe                # compile
sbatch --mem-per-cpu=4G --wrap "./stream_c.exe" # submit to queue and run
```

3. The output (`slurm-jobid.out`) produced:

```
-------------------------------------------------------------
Function    Best Rate MB/s  Avg time     Min time     Max time
Copy:           19158.8     0.106950     0.106896     0.107101
Scale:          11228.1     0.182465     0.182399     0.182589
Add:            12278.5     0.250293     0.250193     0.250484
Triad:          12285.6     0.250150     0.250048     0.250300
-------------------------------------------------------------
```

   We observe that the bandwidths resulting from the Scale, Add, and Triad kernels are roughly consistent, while the bandwidth for the Copy kernel appears to be substantially higher. There may be several explanations for this discrepancy, but they are beyond the scope of this project (and course). For a rough estimate, we can assume a maximum bandwidth $b_{\text{STREAM}} = 12$ GB/s.

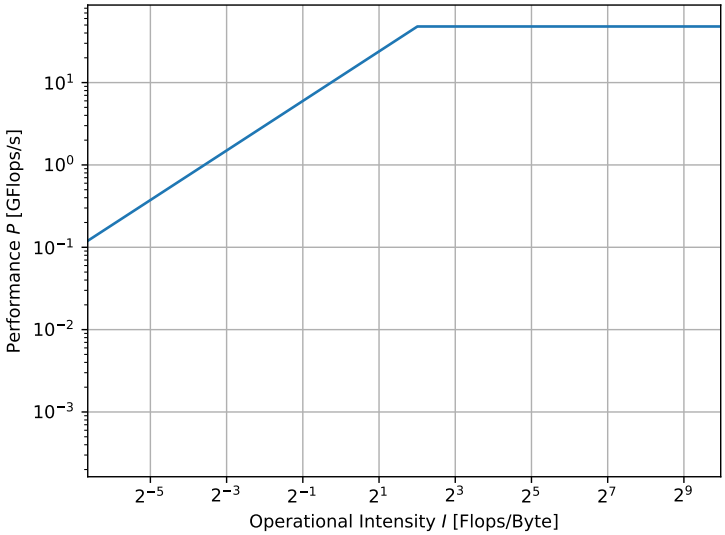


Figure 2: Simple (or naive) roofline model for a single core of a hypothetical CPU.

Of course, you will need to adapt some these steps to the task. In your report, follow a similar style to the above description. Include source code (including build files), batch scripts, and Slurm output files in your submission.
For further context and advice regarding the STREAM benchmark, we recommend looking at the official website.

2.4. Performance model: A simple roofline model

The *roofline* model is a visual representation that serves as a powerful tool for evaluating the performance potential of computing systems, especially in the context of high-performance computing [10]. It plots achievable performance (in GFlops/s) against operational intensity (in Flops/Byte), delineating the maximum performance given by the system's peak performance and the memory bandwidth constraints. The "roofline" itself consists of two main segments: the *memory-bound* region, where performance is limited by the system's memory bandwidth, and the *compute-bound* region, where performance is capped by the peak computational power of the processor. The frontier between both regimes is called the *ridge point*. This model allows developers and researchers to identify whether their algorithms are memory-bound or compute-bound and to understand the performance implications of hardware limitations, guiding optimizations towards achieving maximum efficiency within the given hardware constraints.
A simple (or sometimes called naive) roofline model for a hypothetical system combining Subsections 2.1 and 2.2.1 is displayed in Fig. 2. We observe that the ridge point is roughly at an operational intensity of $I_{\text{ridge}} \approx 4$. Hence, a given kernel or application with an operational intensity below I_{ridge} is memory-bound. Similarly, a given kernel or application with an operational intensity above I_{ridge} is compute-bound.
Your tasks are:

1. With your performance data collected in the previous tasks, create a roofline model for a single core of the Rosa nodes.

2. Visualize your roofline models in a plot similar to Fig. 2.

3. At what operational intensity is a kernel or application memory/compute-bound?

Include plotting scripts in your submission.

3. Matrix multiplication optimization[4] [50 points]

The first part of this project has introduced you how to use the Rosa cluster and you collected some performance characteristics in the second part. Your next task in this project is to write an optimized matrix multiplication function on the Rosa computer. We will give you a generic matrix multiplication code (also called matmul or dgemm), and it will be your job to tune our code to run efficiently on the Rosa processors. Write an optimized single-threaded matrix multiply kernel. This will run on only one core.

3.1. Motivation

Linear algebra is a cornerstone operation in computational science and holds a particularly pivotal role in HPC. In particular, we are going to focus on matrix multiplication. This mathematical operation underpins a vast array of scientific and engineering disciplines, facilitating the modeling of complex systems, simulations of physical phenomena, and the processing of large datasets. In HPC environments, the efficiency and scalability of matrix multiplication directly influence the performance of algorithms (used in computational sciences, engineering, machine learning, and beyond). This functieonality is offered by BLAS (Basic Linear Algebra Subprograms), a specification encompassing a suite of low-level routines designed for executing common linear algebra operations efficiently.
Each manufacturer typically supplies a BLAS library (e.g., Intel's MKL library, AMD's AOCL, ...), carefully hand-optimized for its machines. It includes a large number of routines, not just matrix multiplication, although matrix multiplication is one of the most important, because it is used as a benchmark to compare the speed of computers (e.g., LINPACK Benchmark[5] used in the Top 500[6] list of the fastest supercomputers). In addition to manufacturer supplied libraries, there are also a number of open source projects. For instance, there are the ATLAS[7], GotoBLAS[8], OpenBLAS[9], and BLIS[10] (the 2023 recipient of the James H. Wilkinson Prize for Numerical Software). The list is not meant to be exhaustive.
The performance difference between naive and library implementations can be drastic. Figure 3 shows the performance in GFlops/s of n-by-n matrix multiplication (on a single core of an AMD EPYC 7763 CPU) as a function of matrix size n. It compares the highly optimized Intel MKL and OpenBLAS libraries with a naive implementation. As apparent from the figure, the naive implementation is sub-optimal (to say the least). As it turns out, matrix multiplication, commonly referred to as `DGEMM` (Double precision GEneral Matrix Multiplication), has a growing operational intensity (with the matrix size n). Therefore it can be (asymptotically) pushed into the compute bound regime (e.g., roofline model). However, achieving this requires specialized, hardware-dependent optimization techniques. The goal of this project task is to familiarize you with some basic techniques and offer pointers to literature for more advanced techniques.
However, achieving this requires very specialized, hardware dependent, optimization techniques. It is the goal of this project task to familiarize you with some basic techniques and provide you some pointers to the literature for more advanced techniques in the next paragraph.
The anatomy of a HPC matrix multiplication is described in the article [2]. The BLIS framework is described by Zee et al. [11] and Low et al. [5] propose an analytical performance model to determine certain parameters. The latter article gives a nice overview of the anatomy of a fast `DGEMM`. The recent article by Alaejos et al. [1] provides templates to write micro-kernels using vector intrinsics. Last but not least, we highly recommend the very comprehensive

[4] This task is based on a previous HPC Lab for CSE at ETH Zurich by Prof. Olaf Schenk from the Institute of Computing at the Faculty of Informatics at Universita della Svizzera Italiana (USI), which was itself originally based on a tutorial from Prof. Katherine A. Yelick from the Computer Science Department at the University of Berkeley (`http://www.cs.berkeley.edu/~yelick/`).
[5] `https://www.top500.org/project/linpack/`
[6] `https://www.top500.org/`
[7] `https://math-atlas.sourceforge.net/`
[8] `https://en.wikipedia.org/wiki/GotoBLAS`
[9] `https://www.openblas.net/`
[10] `https://github.com/flame/blis`

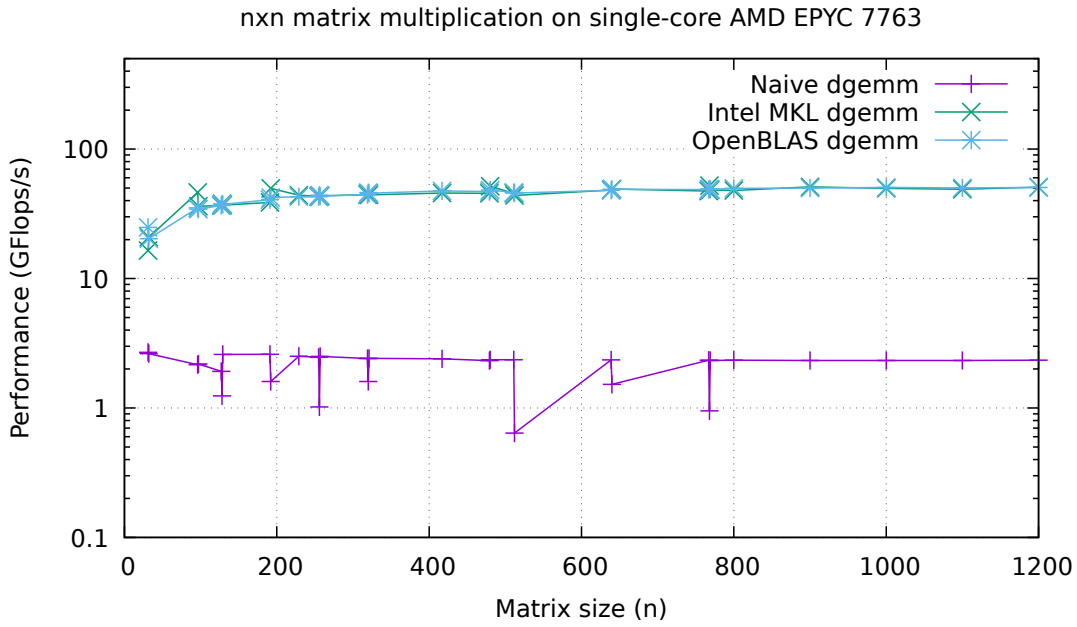


Figure 3: Serial performance of matrix multiplication on the AMD EPYC 7763.

online course "LAFF-On Programming for High Performance" [9]. On a less scientific level, you might also find the newspaper article [6] interesting.

3.2. Importance of data access optimization

Among the various factors that can limit performance in HPC, data access stands out as the most critical. This is due to the inherent imbalance in modern CPUs, where there is a significant discrepancy between their theoretical peak performance and the available memory bandwidth. Therefore, any attempt at optimization should initially focus on minimizing data traffic, or, if that proves to be impractical, strive to make data transfer as efficient as possible. If you haven't already, we strongly recommend to read Chapter 3 of the book [3].

3.2.1. Modeling memory hierarchies

Real memory hierarchies are very complicated, so modeling them carefully enough to predict the performance of an algorithm is hard. To make progress, we will use the following simplified model:

1. There are only **two levels** in the memory hierarchy, **fast** (e.g., registers, caches) and **slow** (e.g., main memory).
2. All the input data starts in the slow level, and the output must eventually be written back to the slow level.
3. The size of the fast memory is M, which is large but much smaller than the slow memory. In particular, the input of large problems will not fit in the fast memory simultaneously.
4. The programmer explicitly controls which data moves between the two memory levels and when.
   Remark: It might seem incorrect to suggest that programmers can control data movement between main memory and cache, as this process is typically automated by the hardware. However, we know how the hardware works: it moves data to the cache precisely when the user first tries to load it into a register to perform arithmetic, and puts it back in main memory when the cache is too full to get the next requested word. This means that we can write programs as if we controlled the cache explicitly by doing arithmetic operations in different

orders. Thus, two programs that do the same arithmetic in different orders may run at very different speeds, because one reduces data movement between the main memory and cache.

5. Arithmetic (and logic) can only be done on data residing in the fast memory and data movement and computation cannot overlap. Each arithmetic operation takes time t_a. Moving a word of data from one memory to the other takes time $t_m \gg t_a$. Hence, if a program performs m memory moves and a arithmetic operations, the total time it takes is $T = a \cdot t_a + m \cdot t_m$, where it may be that $m \cdot t_m \gg a \cdot t_a$.

With this model, let us get a **lower bound** on the speed of any algorithm for a problem that has m_i inputs and m_o outputs, and does a arithmetic operations. The only difference between algorithms that we permit is the order in which the arithmetic operations are executed. According to our model, the run-time taken will be at least

$$T = a \cdot t_a + (m_i + m_o) \cdot t_m,$$

no matter which clever order we do the arithmetic in.
For example, suppose the problem is to take two input arrays of n numbers each, $a[i]$ and $b[i]$ for $i = 1, .., n$, and produce two output arrays s and p, where $s[i] = a[i] + b[i]$ and $p[i] = a[i] \cdot b[i]$. Then the number of arithmetic operations is $a = 2n$, and the number of memory moves is $m_i = 2n$ and $m_o = 2n$. Thus, a lower bound on the run-time for any algorithm for this problem is $T = 2nt_a + 4nt_m$. Let us look at two different algorithms for this problem, and analyze which will be faster. The two algorithms are listed in Table 2. According to our model, we assume that a and b are initially in the slow memory, and $M \ll n$, so a and b cannot fit in the fast memory, and we have indicated when loads from and stores to slow memory occur. The point is that Algorithm 1 loads a and b twice,

Algorithm 1

```
for i = 1, n
  Load a[i], b[i] into fast memory
  s[i] = a[i] + b[i]
  Store s[i] into slow memory
end for

for i = 1, n
  Load a[i], b[i] into fast memory
  p[i] = a[i] * b[i]
  Store p[i] into slow memory
end for
```

Algorithm 2

```
for i = 1, n
  Load a[i], b[i] into fast memory
  s[i] = a[i] + b[i]
  Store s[i] into slow memory
  p[i] = a[i] * b[i]
  Store p[i] into slow memory
end for
```

Table 2: Two algorithms to compute the sum and product of the arrays.

whereas Algorithm 2 only loads them once (remember, there is not enough room in the fast memory to keep all the $a[i]$ and $b[i]$ for a second pass to compute $p[i]$). As per our simple model, Algorithm 1 takes run-time $T_1 = 2nt_a + 6nt_m$, whereas Algorithm 2 takes only $T_2 = 2nt_a + 4nt_m$, the minimum possible. Thus, for $t_a \ll t_m$, Algorithm 1 takes 50% longer to run than Algorithm 2. In the case of matrix multiplication, we will observe an even more dramatic difference.

3.2.2. Minimizing slow memory access in matrix multiplication

We will only consider different ways to implement n-by-n matrix multiplication $C = C + A \cdot B$ using $2n^3$ operations (i.e., we will not consider Strassen's method [8] and variants thereof). Thus, the only difference between algorithms will depend on the number of loads and stores to the slow memory. We also assume that the slow memory is large

enough to contain our three n-by-n matrices A, B and C, but the fast memory is too small for this. Otherwise, if the fast memory were large enough to contain A, B, and C simultaneously, then our algorithm would simply be:

```
Load A, B and C from slow into fast memory

method dgemm_fastmem(A, B, C)

  Compute $C = C + A \cdot B$ entirely in fast memory

end method dgemm_fastmem

Store the result $C$ back to slow memory
```

The number of slow memory accesses for this algorithm is $4n^2$ ($m_i = 3n^2$ loads of A, B, and C into fast memory, and $m_o = n^2$ stores of C to slow memory), yielding a run-time of $T = 2n^3 t_a + 4n^2 t_m$. Clearly, no algorithm doing $2n^3$ arithmetic operations can be faster. At the extreme where the fast memory is very small (M = 1), then there will be at least 1 memory reference per operand for each arithmetic operation involving entries of A and B, for a run-time of at least $T = 2n^3 t_a + (2n^3 + 2n^2)t_m$.

So, when the size of fast memory M satisfies $1 \ll M \ll 3n^2$, what is the fastest algorithm?

This is the practical question for large matrices, since real caches have thousands of entries. As we just saw, the worst case run-time is $2n^3 t_a + (2n^3 + 2n^2)t_m$, and the best we can hope for is $2n^3 t_a + 4n^2 t_m$, which would be almost $n/2$ times faster for $t_m \gg t_a$.
We begin by analyzing the simplest matrix multiplication algorithm, which we repeat below, including descriptions of when data moves between the slow and the fast memories. Remember that A, B, and C all start in the slow memory, and that the result C must be finally stored in the slow memory.

```
Naive matrix multiplication $C = C + A \cdot B$

method dgemm_naive(A, B, C)

  for i = 1, n
    for j = 1, n
      Load c_{i,j} into fast memory
      for k = 1, n
        Load a_{i,k} into fast memory
        Load b_{k,j} into fast memory
        c_{i,j} = c_{i,j} + a_{i,k} * b_{k,j}
      end for
      Store c_{i,j} into slow memory
    end for
  end for

end method dgemm_naive
```

Let m_{naive} denote the number of **slow memory references** in this naive algorithm. Then

$$\begin{aligned} m_{\text{naive}} &= n^3 && \dots \text{ for loading each entry of A } n \text{ times} \\ &+ n^3 && \dots \text{ for loading each entry of B } n \text{ times} \\ &+ 2n^2 && \dots \text{ for loading and storing each entry of C once} \\ &= 2n^3 + 2n^2, \end{aligned}$$

or about as many slow memory references as arithmetic operations. Thus, the run-time is $T_{\text{naive}} = 2n^3 t_a + (2n^3 + 2n^2)t_m$, the worst possible.

Let us analyse an improved algorithm called **blocked matrix multiplication** (sometimes it is called **tiled** or **panelled** instead of **blocked**). The n-by-n matrix A is divided into smaller s-by-s sub-matrices or blocks A_{ij}, where s is a parameter called the block size that we will specify later. We assume s divides n for simplicity. Matrices B and C are similarly partitioned. Then, we can think of A as an n/s-by-n/s block matrix, where each entry $A_{i,j}$ is an s-by-s block. The inner loop $C_{i,j} = C_{i,j} + A_{i,k} \cdot B_{k,j}$ now runs for $k = 1$ to s, and represents an s-by-s matrix multiplication and addition. The algorithm becomes:

```
1  Blocked matrix multiplication $C = C + A \cdot B$
2
3  method dgemm_blocked(A, B, C)
4
5    for i = 1, n/s
6      for j = 1, n/s
7        Load C_{i,j} block into fast memory
8        for k = 1, n/s
9          Load A_{i,k} block into fast memory
10         Load B_{k,j} block into fast memory
11         dgemm_fastmem(A_{i,k}, B_{k,j}, C_{i,j})
12       end for
13       Store C_{i,j} into slow memory
14     end for
15   end for
16
17 end method dgemm_blocked
```

The inner loop `dgemm_fastmem`$(A_{i,k}, B_{k,j}, C_{i,j})$ has all its data residing in the fast memory, and so causes no slow memory traffic at all. Redoing the count of slow memory references yields

$$\begin{aligned} m_{\text{blocked}} = m_{\text{blocked}}(s) &= (n/s)^3 \cdot s^2 && \dots \text{ for loading each block } A_{i,k} \ (n/s)^3 \text{ times} \\ &+ (n/s)^3 \cdot s^2 && \dots \text{ for loading each block } B_{k,j} \ (n/s)^3 \text{ times} \\ &+ 2(n/s)^2 \cdot s^2 && \dots \text{ for loading and storing each block } C_{i,j} \text{ once} \\ &= 2n^3/s + 2n^2. \end{aligned}$$

Comparing $m_{\text{naive}} = 2n^3 + 2n^2$ and $m_{\text{blocked}} = 2n^3/s + 2n^2$, it is obvious that we want to pick s as large as possible to make $m_{\text{blocked}}(s)$ as small as possible. But how big can we pick s? The largest possible value is obviously $s = n$, which corresponds to loading all of A, B and C into the fast memory, which we cannot do. So, s depends on the size M of the fast memory, and the constraint it must satisfy is that the three s-by-s blocks $A_{i,k}, B_{k,j}$, and $C_{i,j}$ must simultaneously fit in the fast memory, which implies $3s^2 \leq M \implies s \leq \sqrt{M/3}$. Therefore, the largest value

$s = \sqrt{M/3}$ yields

$$m_{\text{blocked}}(\sqrt{M/3}) = \sqrt{12}\frac{n^3}{\sqrt{M}} + 2n^2.$$

In other words, for large matrices (large n) we decrease the number of slow memory references, the most expensive operation, by a factor $\mathcal{O}(M)$. This is attractive, because it says that cache (fast memory) helps, and the larger the cache the better.
In summary, the running time for this algorithm is

$$T_{\text{blocked}} = 2n^3 \cdot t_a + (\sqrt{12}\frac{n^3}{\sqrt{M}} \cdot t_m).$$

There is a theorem, which we will not prove, that says that up to constant factors, we cannot do fewer slow memory references than this (while doing the usual $2n^3$ arithmetic operations):

Theorem (Hong + Kung, 1981, 13th Symposium on the Theory of Computing [4]): Any implementation of matrix multiplication using $2n^3$ arithmetic operations performs at least $\mathcal{O}(n^3/\sqrt{M})$ slow memory references.

In practice, this technique is very important to get matrix multiplication to run as fast as possible. But careful attention must also be paid to other details of the instruction set, arithmetic units, and so on. If there are more levels of memory hierarchy (two levels of cache), then one might use this technique recursively, dividing s-by-s blocks into yet smaller blocks to exploit the next level of memory hierarchy. However, this goes beyond the scope of this project and course.

3.3. Optimizing square matrix multiplication

Your task is now to write a blocked matrix multiplication (as outlined above) and optimize it for a single core on the Rosa Cluster.

Skeleton code

Download the skeleton code provided on the iCorsi page. It contains the following:

- `dgemm-naive.c` – For illustrative purposes, a naive implementation of matrix multiplication using three nested loops.
- `dgemm-blas.c` – A wrapper which calls an optimized BLAS implementation of matrix multiplication (OpenBLAS).
- `dgemm-blocked.c` – A skeleton for a blocked implementation of matrix multiplication. It is your task to implement and optimize the `square_dgemm` function in this file.
- `benchmark.c` – A driver program that generates matrices of a number of different sizes and benchmarks the performance. It outputs the performance in GFlops/s and in a percentage of theoretical peak performance attained. You should not need to modify this file, except perhaps to change the peak performance constant `MAX_SPEED` if you wish to test on another computer.
- `Makefile` – A simple makefile to build the executables for the benchmarking of the naive, optimized BLAS and your blocked implementation.
- `run_matrixmult.sh` – A job script that executes all three executables and produces log files (`*.data`) that contain the performance logs. It also plots the data in the performance logs and produces `timing.pdf` showing the results.

Familiarize yourself with the code and the used conventions. In particular, note that we use the *column major order* storage scheme to conform with BLAS' Fortran origins.

Running the code

The skeleton code should run out of the box and a file listing should look as:

```
[user@icslogin]$ cd 3-Optimize-Matrix-Matrix-Mult/
[user@icslogin]$ ls
benchmark.c  dgemm-blas.c  dgemm-blocked.c  dgemm-naive.c  Makefile
run_matrixmult.sh  timing_basic_dgemm.data  timing.gp
```

We will use the GNU Compiler Collection and the OpenBLAS implementation, which can be loaded on Rosa as follows:

```
[user@icslogin]$ module load gcc intel-oneapi-mkl
[user@icslogin]$ make
```

Building the code

```
[user@icslogin]$ module load gcc intel-oneapi-mkl
[user@icslogin]$ make
```

generates the three executables `benchmark-naive`, `benchmark-blas` and `benchmark-blocked`. The easiest way to run the code is to submit a batch job. We have already provided batch files which will launch jobs for each matrix multiply version using one core:

```
[user@icslogin]$ sbatch run_matrixmult.sh
Submitted batch job 49100417
```

or

```
[user@icslogin]$ sbatch --reservation=hpc-wednesday run_matrixmult.sh
Submitted batch job 49100415
```

Our jobs are now submitted to the Rosa cluster 's job queue. We can now check on the status of our submitted jobs using a few different commands.

```
[user@icslogin]$ squeue
   JOBID  USER PART  NAME      ST  START_TIME END_TIME TIME_LEFT  NODES
 1173224  user slim  matrixmu  R   14:47:17   15:17:17 29:50      1
```

When our job is finished, you will find new files in the directory containing the output of the benchmark programs. For example, we will find the files `matrixmult-jobid.out` and `matrixmult-jobid.err`. The first file contains the standard output and the second file contains the standard error. Additionally, the performance data are stored in `*.data` files and `timing.pdf` is a plot of the performance.

Optimizing square matrix multiplication [50 points]

- **Implementation**: Implement blocking for square matrix multiplication in `dgemm_blocked.c`.
- **Optimization**: Optimize your code to maximize its performance. Consider various strategies, such as compiler options, tuning data access patterns (including block size adjustments), using `#pragma` directives, vectorization, loop unrolling, and more.
- **Documentation and Analysis**: Document the used or attempted optimizations in your report, including performance graphs and, if applicable, tables. Provide a detailed description of your experimental setup. Compare your optimized implementation to the OpenBLAS library.
- Employ any advanced techniques from library implementations as described in the references provided in the motivation 3.1, including any references therein or other relevant resource. However, remember to accurately cite all your sources.

4. Quality of the Report [15 Points]

Each project will have 100 points (out of 15 point will be given to the general written quality of the report).

Additional notes and submission details

Submit the source code files (together with your used `Makefile`) in an archive file (tar, zip, etc.) and summarize your results and the observations for all exercises by writing an extended Latex report. Use the Latex template from the webpage and upload the Latex summary as a PDF to iCorsi .

- Your submission should be a gzipped tar archive, formatted like project_number_lastname_firstname.zip or project_number_lastname_firstname.tgz. It should contain:
  - All the source codes of your solutions.
  - Build files and scripts. If you have modified the provided build files or scripts, make sure they still build the sources an run correctly. We will use them to grade your submission.
  - project_number_lastname_firstname.pdf, your write-up with your name.
  - Follow the provided guidelines for the report.
- Submit your .tgz through iCorsi .

Code of Conduct and Policy

- Do not use or otherwise access any on-line source or service other than the iCorsi system for your submission. In particular, you may not consult sites such as GitHub Co-Pilot or ChatGPT.
- You must acknowledge any code you obtain from any source, including examples in the documentation or course material. Use code comments to acknowledge sources.
- Your code must compile with a standard-configuration C/C++ compiler.

References

[1] Guillermo Alaejos, Adrian Castello, Hector Martinez, Pedro Alonso-Jorda, Francisco D. Igual, and Enrique S. Quintana-Orti. Micro-kernels for portable and efficient matrix multiplication in deep learning. *The Journal of Supercomputing*, 79(7):8124–8147, May 2023. doi:10.1007/s11227-022-05003-3.

[2] Kazushige Goto and Robert A. van de Geijn. Anatomy of high-performance matrix multiplication. *ACM Transactions on Mathematical Software*, 34(3):1–25, May 2008. doi:10.1145/1356052.1356053.

[3] Georg Hager and Gerhard Wellein. Introduction to high performance computing for scientists and engineers. *Chapman & Hall/CRC Computational Science*, July 2010. URL: http://dx.doi.org/10.1201/EBK1439811924, doi:10.1201/ebk1439811924.

[4] Jia-Wei Hong and H. T. Kung. I/O complexity: The red-blue pebble game. In *Proceedings of the thirteenth annual ACM symposium on Theory of computing*, STOC '81, pages 326–333, New York, NY, USA, May 1981. Association for Computing Machinery. URL: https://dl.acm.org/doi/10.1145/800076.802486, doi:10.1145/800076.802486.

[5] Tze Meng Low, Francisco D. Igual, Tyler M. Smith, and Enrique S. Quintana-Orti. Analytical Modeling Is Enough for High-Performance BLIS. *ACM Transactions on Mathematical Software*, 43(2):12:1–12:18, August 2016. URL: https://dl.acm.org/doi/10.1145/2925987, doi:10.1145/2925987.

[6] John Markoff. Writing the Fastest Code, by Hand, for Fun: A Human Computer Keeps Speeding Up Chips. *The New York Times*, 2005. [Online; accessed 24-February-2024]. URL: https://www.nytimes.com/2005/11/28/technology/writing-the-fastest-code-by-hand-for-fun-a-human-computer-keeps.html?smid=url-share.

[7] John D. McCalpin. Memory bandwidth and machine balance in current high performance computers. *IEEE computer society technical committee on computer architecture (TCCA) newsletter*, 2, December 1995. URL: https://www.cs.virginia.edu/~mccalpin/papers/balance/.

[8] Volker Strassen. Gaussian elimination is not optimal. *Numerische Mathematik*, 13(4):354–356, aug 1969. doi:10.1007/bf02165411.

[9] Robert van de Geijn, Margaret Myers, and Devangi Parikh. LAFF - On Programming for High Performance. https://www.cs.utexas.edu/users/flame/laff/pfhp/, November 2021. [Online; accessed 24-February-2024].

[10] Samuel Williams, Andrew Waterman, and David Patterson. Roofline: an insightful visual performance model for multicore architectures. *Communications of the ACM*, 52(4):65, April 2009. doi:10.1145/1498765.1498785.

[11] Field G. Van Zee, Tyler M. Smith, Bryan Marker, Tze Meng Low, Robert A. Van De Geijn, Francisco D. Igual, Mikhail Smelyanskiy, Xianyi Zhang, Michael Kistler, Vernon Austel, John A. Gunnels, and Lee Killough. The BLIS framework. *ACM Transactions on Mathematical Software*, 42(2):1–19, jun 2016. doi:10.1145/2755561.

Università della Svizzera italiana | Institute of Computing CI

High-Performance Computing Lab **Institute of Computing**

Student: Dennys Mike Huber Discussed with: Alberto Finardi & Leon Ackermann

Solution for Project 1

In this project you will practice memory access optimization, performance-oriented programming, and OpenMP parallelizaton on the Rosa Cluster.

1. Rosa Warm-Up *(5 Points)*

Question 1 What is the module system and how do you use it?

The module system allows the user to switch between different versions of pre-installed programs and libraries. Each module comes with its own bundle of programs and libraries, which can be loaded using the `module` command as follows:

```
module load [module-name]
```

Internally the `module` command interprets pre-configured `modulefiles`, which typically instruct the `module` command to adjust or set shell environment variables such as `PATH`, `MANPATH`, etc, giving the user an easy way to access and switch between different compilers, loader, libraries and utilities. In order to discover, which modules are available on the current system, simply run the following command. [1, 7]

```
module avail
```

Further details about a specific module can be displayed by using the following command.

```
module show [module-name]
```

Additionally the following commands can be used to manipulate the current set of loaded modules.

```
module list [module-name] # lists all currently loaded modules
module switch [module-name] # Switches loaded module1 with module2.
module rm [module-name] # unloads module
module purge  # unloads all modules
```

Question 2 What is Slurm and its intended function?

The Simple Linux Utility for Resource Management (Slurm) is a cluster management and job scheduling system with the intend to be fault-tolerant and highly scalable. It is widely used on supercomputers in order to simplify the job execution of parallel jobs by hiding several complexities, including load management, job scheduling etc. Slurm comes with three core functionalities:

1. Allocation of access to resources to users.
2. Framework for starting, executing and monitoring of jobs.
3. Arbitration of contended resources.

Slurm is aware of all the available computing resources of the system, which includes number of cores and nodes, memory of each core, sockets, number of GPUs and number of hyperthreads. After successfully submitting a job, Slurm will place the job in a queue based on the requested resources and priorities. If the resources become available, the job will be executed according to the scheduling policy and can then be monitored using various Slurm utilities.[1, 8]

Question 3 Write a simple "Hello World" C/C++ program which prints the host name of the machine on which the program is running.

The program prints the environment variable `HOSTNAME` of the node it is running on.

```cpp
#include <cstdlib>
#include <iostream>

int main() {

  const char *hostname = std::getenv("HOSTNAME");

  if (hostname) {
    std::cout << "This program is running on " << hostname << "." << std::endl;
  } else {
    std::cout
        << "Environment variable for HOSTNAME was not found on your maschine."
        << std::endl;
  }

  return 0;
}
```

Listing 1: Hello World

Question 4 Write a batch script which runs your program on one node. The batch script should be able to target nodes with specific CPUs with different memories. You can obtain the information on available nodes using the command sinfo.

In order to figure out which specific features are available on the Rosa cluster, after consulting the manual for `sinfo` [9] one can run the following command.

```
# Command
sinfo -o "%N %f"
# Output
NODELIST AVAIL_FEATURES
icsnode[01-39,41-42] (null)
```

Listing 2: Detecting features on Rosa

Here `-o` specifies the output format and `%N %f` specifies the output format to be a string and display the list of nodes and the list of features associate to it. Unfortunatly this cluster does not have any features, as seen by the output of listing 2 therefore the `--constraint` option cannot be applied. The batch file is given in Listing 3, notice that the `--nodes` and the `--ntask` are set to one.

```
#!/bin/bash
#SBATCH --job-name=slurm_job_one       # Job name    (default: sbatch)
#SBATCH --output=slurm_job_one-%j.out # Output file (default: slurm-%j.out)
#SBATCH --error=slurm_job_one-%j.err  # Error file  (default: slurm-%j.out)
#SBATCH --nodes=1                     # Number of nodes
#SBATCH --ntasks=1                    # Number of tasks
#SBATCH --cpus-per-task=1             # Number of CPUs per task
#SBATCH --time=00:01:00               # Wall clock time limit

# load some modules & list loaded modules
module load gcc
module list

```

```
# print CPU model
lscpu | grep "Model name"

# run (srun: run job on cluster with provided resources/allocation)
srun ./hello_world
```

Listing 3: single node sbatch script

Question 5 Write another batch script which runs your program on two nodes.

In comparison to the previous script the important changes are that `--ntasks` and `--nodes` are now set to two. Resulting in the hello_world program being run twice. Depending on the load on Rosa both instances of the program can be executed either on two separate nodes or on the same node. When I ran this sbatch script I encounter the latter scenario.

```
#!/bin/bash
#SBATCH --job-name=slurm_job_two       # Job name    (default: sbatch)
#SBATCH --output=slurm_job_two-%j.out # Output file (default: slurm-%j.out)
#SBATCH --error=slurm_job_two-%j.err  # Error file  (default: slurm-%j.out)
#SBATCH --nodes=2                     # Number of nodes
#SBATCH --ntasks=2                    # Number of tasks
#SBATCH --cpus-per-task=1             # Number of CPUs per task
#SBATCH --time=00:01:00               # Wall clock time limit

srun ./hello_world
```

Listing 4: two nodes and tasks sbatch script

2. Performance Characteristics *(30 Points)*

2.1. Peak Performance

As stated on the Compute Resources Webpage of the Institute of Computing [1], Rosa is a high performance computing cluster made up of 42 Nodes, each with two Intel XEON E5-2650 v3 CPUs with 10 cores each. To calculate the Peak performance of the cluster, we begin by calculating the peak performance of a single core P_{core} given by the equation 1.

$$\begin{aligned} P_{\text{core}} &= n_{\text{super}} \times n_{\text{FMA}} \times n_{\text{SIMD}} \times f \\ &= 2 \times 2 \times 4 \times 2.3 \text{ GHz} \\ &= 36.8 \text{ GFlops/s} \end{aligned} \tag{1}$$

By consulting Intel's ARK website [4] about the XEON E5-2650 v3 CPUs, we can find out further details about the CPU, firstly that it operates on a base frequency f of 2.3 GHz, secondly that it is part of the *Haswell* microprocessor architecture family and thirdly that the CPU supports Intel's AVX2 SIMD instruction set extension. The AVX2 SIMD instructions [3] uses a 256-bit wide vector registers and because we consider 64-bit double precision float operations to compute the peak performance, each core is able to perform 4 operations simultaneously. Hence leading n_{SIMD} to be 4. Furthermore, by looking at the website uops.info [10] and filtering the table using the keywords `Haswell` and `FMA` we can see that the throughput (TP) is 0.5, this means that a core can execute two float point operations per cycle, resulting in $n_{\text{super}} = 2$. In addition the port column shows 1*p01, indicating that one vector FMA operation can be executed on each port 0 and 1. Hence we can set n_{FMA} to 2 as well.

After successfully calculating the peak performance of a single core, the result can now be used to compute the peak performance of a CPU, given by equation 2.

$$\begin{aligned} P_{\text{CPU}} &= n_{cores} \times P_{core} \\ &= 10 \times 36.8 \text{ GFlops/s} \\ &= 368 \text{ GFlops/s} \end{aligned} \tag{2}$$

The number of physical cores n_{cores} can be found on the previously mentioned Intel ARK website [4]. Now we can compute the peak performance of a single Node on the Rosa cluster given by

$$\begin{aligned} P_{\text{node}} &= n_{\text{sockets}} \times P_{\text{CPU}} \\ &= 2 \times 368 \text{ GFlops/s} \\ &= 736 \text{ GFlops/s.} \end{aligned} \tag{3}$$

n_{sockets} can be set to 2, because for each we Node we have two sockets as stated on the USI's Institute for Computing website [1]. Now we are finally ready to compute the peak performance of the 42 node cluster giving us the final peak performance of

$$\begin{aligned} P_{\text{Rosa}} &= n_{\text{nodes}} \times P_{\text{node}} \\ &= 42 \times 736 \text{ GFlops/s} \\ &= 30912 \text{ GFlops/s} = 30.912 \text{ TFlops/s.} \end{aligned} \tag{4}$$

2.2. Memory Hierarchies

The memory hierarchy of a single node on the Rosa cluster can be seen in Figure 1. It shows two identical CPU's with 10 cores and the corresponding Level of caches including their size.

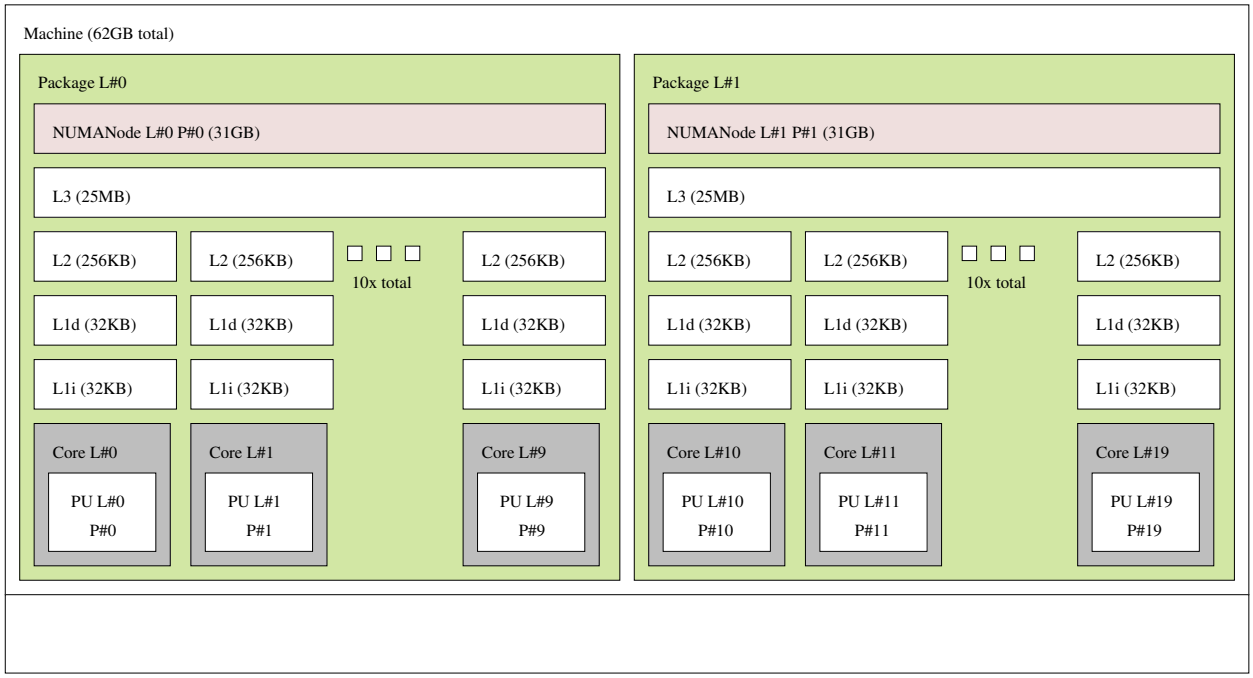


Figure 1: 2 x Intel Xeon E5-2650 v3 @ 2.30GHz, 20 (2 x 10) cores

For single core the memory hierarchy is summarized in Table 1.

Main Memory	31 GB
L3 cache	25 MB
L2 cache	256 KB
L1 cache	32 KB

Table 1: Memory hierarchy of a Rosa node with an Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz

2.3. Bandwidth: STREAM Benchmark

After illustrating the topology of single node, we are now interested in the speed of the memory system. This speed is measured in terms of bandwidth, indicating the how much data can be transfer within a specific time period. The STREAM is a common benchmark to assess the bandwidth. It uses four operations: Copy, Scale, Add and Triad to evaluate the bandwidth and the rate of computing. In order to properly setup benchmark we need to compute the number of element, which is, according to instructions given on the STREAM benchmark website [6], four times the sum of all last-level cache. Because we have the same CPU as the login-node and therefore the same L3 cache size, we can take the same arraysize as in the exercise, which is 128×10^6 elements.

```
-------------------------------------------------------------
Function    Best Rate MB/s  Avg time     Min time     Max time
Copy:           19199.2     0.106757     0.106671     0.106884
Scale:          11241.1     0.182266     0.182188     0.182441
Add:            12299.4     0.249855     0.249768     0.250008
Triad:          12294.1     0.249960     0.249875     0.250129
-------------------------------------------------------------
```

Listing 5: Output of STREAM benchmark

The Copy function is ignored and significantly higher due to several factors, which is beyond the scope of this report. The observed best rate for Scale, Add and Triad are approximately similar, therefore we can take as a rough estimate the maximum bandwidth as

$$b_{\text{STREAM}} \approx 12\text{GB/s}. \tag{5}$$

2.4. A simple roofline model

As a final step the values computed in previous section are now coming together under one roof. Using the peak performance of a single core, as seen in equation 1 and the maxium bandwidth, calculated in 5, we can construct a simple roofline model. The performance potential of Rosa is shown in 2 with the ridge point being located at $I_{\text{ridge}} \approx 3$, separating the Bandwidth bound region to the left and the compute bound region to the right. This roofline model can now provided us a reference guide when improving the performance of our code on Rosa.

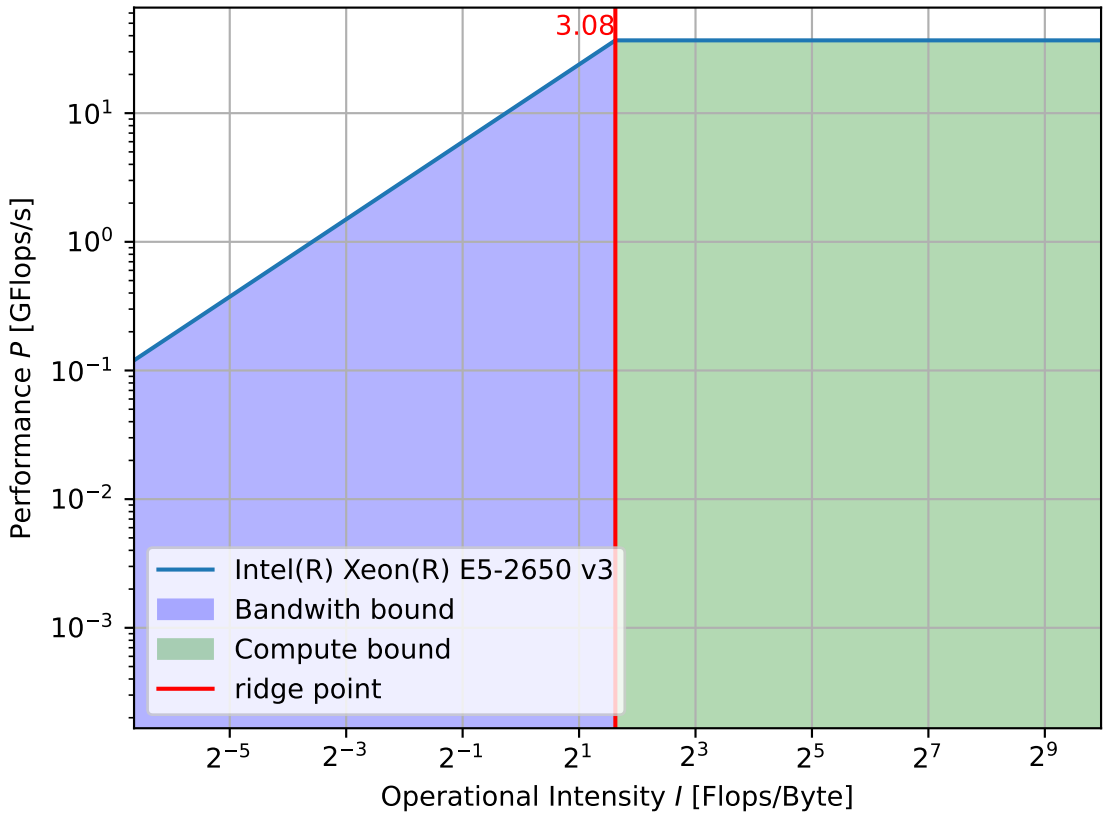


Figure 2: Naive roofline model for Intel Xeon E5-2650 v3 @ 2.30GHz

3. Optimize Square Matrix-Matrix Multiplication *(50 Points)*

3.1. Implementation

As given in the exercise, the optimal block size is

$$s = \sqrt{\frac{M}{3}} \tag{6}$$

where M is the size of the fast memory. In our case this is the size of the L1 cache, which is 32KB as seen in table 1. In our dgemm-blocked implementation we use the double type, which has size 64 bit or 8 byte. Therefore we can convert our L1 cache size into bytes and divide it by the size of the double type in order to get M.

$$M = \frac{32000}{8} = 4000 \text{ doubles} \tag{7}$$

After finding M, we can now calculate s:

$$s = \sqrt{\frac{4000}{3}} = 36.5148 \approx 36 \text{ blocks} \tag{8}$$

We can now insert this result into the blocked dgemm code given in Listing 6. Shifting our attention to the actual implementation of Blocked dgemm. The goal is to improve cache utilization by optimizing the memory access pattern and reduce cache misses.
The two outer loops split the matrix up into smaller blocks of the previously calculated block sizes s. The next two inner loops go through the individual entry of each block and make sure if the final block cannot be completely filled due to the matrix dimension, that there's not an out-of-bounds error. The innermost loop iterates over the shared dimension. Finally the computation is then performed according to the column-major format. Notice that an additional variable sum is introduced, which reduces the times we have to write to the C matrix.

```
1  const char *dgemm_desc = "Blocked dgemm.";
2
3  // Optimize for single core
4  #define BLOCK_SIZE 36
5
6  /* This routine performs a dgemm operation
7   *
8   *  C := C + A * B
9   *
10  * where A, B, and C are lda-by-lda matrices stored in column-major format.
11  * On exit, A and B maintain their input values.
12  */
13 void square_dgemm(int n, double *A, double *B, double *C) {
14   double sum;
15   for (int bi = 0; bi < n; bi += BLOCK_SIZE) {
16     for (int bj = 0; bj < n; bj += BLOCK_SIZE) {
17       for (int i = bi; i < bi + BLOCK_SIZE && i < n; i++) {
18         for (int j = bj; j < bj + BLOCK_SIZE && j < n; j++) {
19           sum = C[i + j * n];
20           for (int k = 0; k < n; k++) {
21             sum += A[i + k * n] * B[k + j * n];
22           }
23           C[i + j * n] = sum;
24         }
25       }
26     }
27   }
28 }
29
```

Listing 6: Dgemm blocked

This initial dgemm block implementation shows a significant improvement compared to the naive dgemm implementation as depicted in Figure 3

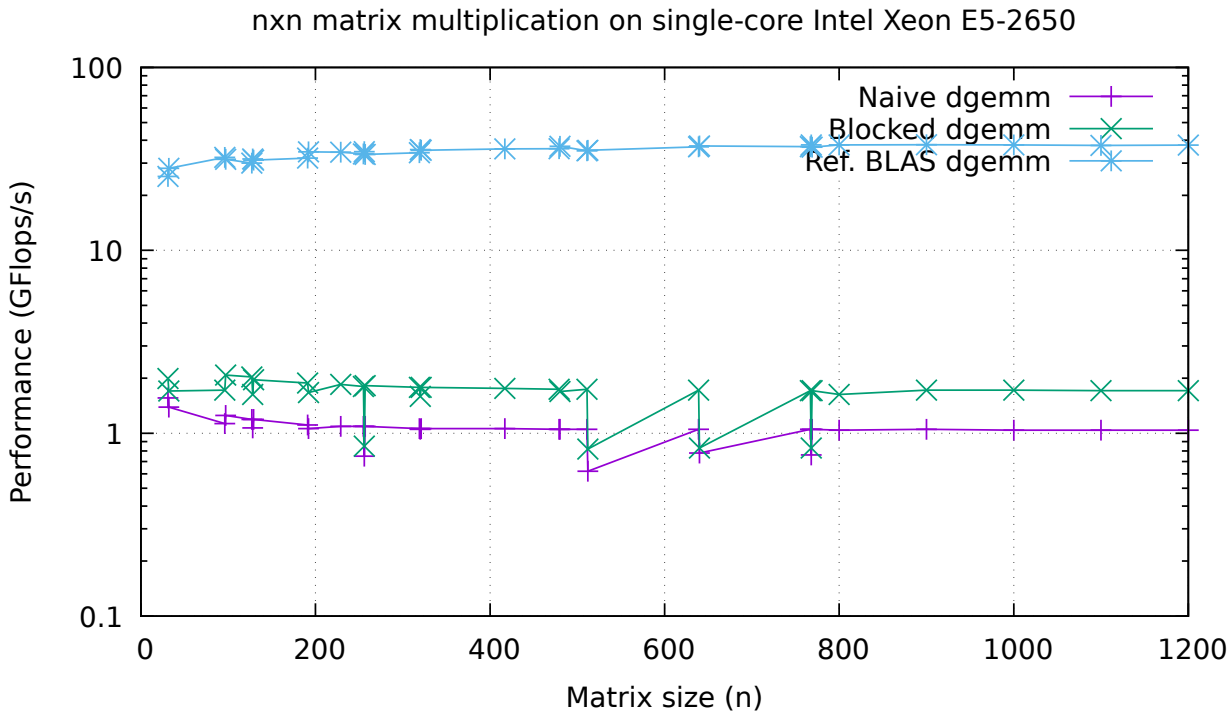


Figure 3: Timing of initial block matrix implementation

3.2. Unrolling Loops and Pragma statements

As an attempt of improving the performance further, the innermost loop was replaced by the unrolled loop seen in Listing 7. This loop is unrolled by a factor of 4, because this is equivalent to the number of doubles the FPU can process concurrently with a single SIMD instruction.

```
int k;
for (k = 0; k <= n - 4; k += 4) {
  sum += A[i + k * n] * B[k + j * n];
  sum += A[i + (k + 1) * n] * B[(k + 1) + j * n];
  sum += A[i + (k + 2) * n] * B[(k + 2) + j * n];
  sum += A[i + (k + 3) * n] * B[(k + 3) + j * n];
}
// For the final block if the number of elements is not divisible by 4
for (; k < n; k++) {
  sum += A[i + k * n] * B[k + j * n];
}
```

Listing 7: Unrolled loop

As a second approach the `#pragma gcc ivdep` [5] directive was introduced in order to force the compiler to unconditionally vectorize the innermost loop. Both approach did not show any improvement in performance compared to the initial block matrix implementation in Figure 3. Hence these approaches were subsequently removed from my final blocked dgemm implementation, but the output file for this option can be found in the submission. These approaches were most likely unsuccessful due to the compiler being able of doing these optimization on its own. This theory is reinforced by having a look at the provided compiler flags in the `Makefile` this includes flags such as `-funroll-loops` and `-march=native`.

3.3. Compiler Flags

By introducing further compiler flags, it is possible to increase the performance even more. Especially the addition of `-ffast-math` [2] resulted in an significant uptick in performance, see Figure 4. It enables a set of flag that enables aggressive floating-point optimizations, but be aware `-ffast-math` is not always safe.

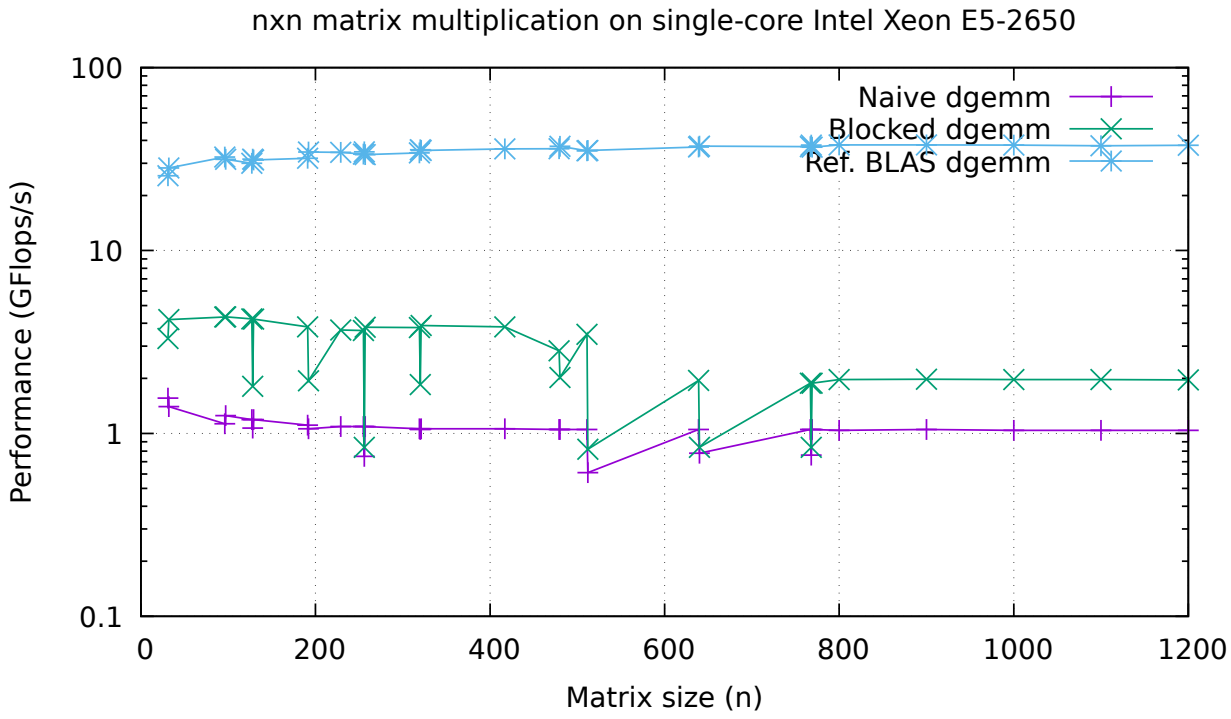


Figure 4: Timing of dgemm using ffast-math

References

[1] *Compute Resources – Institute of Computing USI*. en-US. URL: https://www.ci.inf.usi.ch/research/resources/ (visited on 10/05/2024).

[2] *FloatingPointMath - GCC Wiki*. URL: https://gcc.gnu.org/wiki/FloatingPointMath (visited on 10/07/2024).

[3] *Intel® Intrinsics Guide*. en. URL: https://www.intel.com/content/www/us/en/docs/intrinsics-guide/index.html (visited on 10/09/2024).

[4] *Intel® Xeon® Processor E5-2650 v3 (25M Cache, 2.30 GHz) Product Specifications*. en. URL: https://www.intel.com/content/www/us/en/products/sku/81705/intel-xeon-processor-e5-2650-v3-25m-cache-2-30-ghz.html (visited on 09/29/2024).

[5] *Loop-Specific Pragmas - Using the GNU Compiler Collection (GCC)*. URL: https://gcc.gnu.org/onlinedocs/gcc-4.9.0/gcc/Loop-Specific-Pragmas.html (visited on 10/08/2024).

[6] *MEMORY BANDWIDTH: STREAM BENCHMARK PERFORMANCE RESULTS*. URL: https://www.cs.virginia.edu/stream/ (visited on 10/08/2024).

[7] *module(1): command interface to Modules package - Linux man page*. URL: https://linux.die.net/man/1/module (visited on 10/05/2024).

[8] *Slurm Workload Manager - Overview*. URL: https://slurm.schedmd.com/overview.html (visited on 10/05/2024).

[9] *Slurm Workload Manager - sinfo*. URL: https://slurm.schedmd.com/sinfo.html (visited on 10/09/2024).

[10] *uops.info - Table*. URL: https://uops.info/table.html (visited on 09/29/2024).