Implementing BLAS Operation using OpenMP

Amit Singh Cem Bassoy

March 24, 2021

Contents

Chapter 1

Introduction

Linear algebra is a vital tool in the toolbox for various applications, from solving a simple equation to the art of Deep Learning algorithm or Genomics. The impact can felt across modern-day inventions or day to day life. Many tried to optimize these routines by hand-coding them in the assembly or the compiler intrinsics to squeeze every bit of performance out of the CPU; some chip manufacturers provide library specific to their chip. A few high-quality libraries, such as **Intel's MKL**, **OpenBLAS**, **Flame's Blis**, **Eigen**, and more, each one has one common problem, they are architecture-specific. They need to be hand-tuned for each different architecture.

There three ways to implement the routines and each one has its shortcomings:

1. Hand code them in assembly and try to optimize them for each architecture.
2. Use the compiler intrinsics.
3. Simply code them and let the compiler optimize.

1.1 Hand-tuned Assembly

Hand-coded assembly gives high performance due to more control over registers, caches, and instructions used, but that comes with its issues, which we exchange for performance:

- Need a deep understanding of the architecture
- Maintenance of the code
- It violates the DRY principle because we need to implement the same algorithm for a different architecture
- Development time is high

- Debugging is hard
- Unreadable code
- Sometimes lead to micro-optimization or worst performance than the compiler

Intel's MKL, **OpenBLAS** and **Flame's Blis** comes under this category

1.2 Compiler Intrinsics

The compiler intrinsics is one layer above the assembly, and we lose control over which register to use. If an intrinsic has multiple representations, then we do not know which instruction will emit. There are the following issues with this approach:

- Need the knowledge of intrinsic
- Maintenance of the code much better than assembly but not great
- DRY principle achieved with little abstraction
- Development time is better than assembly but not great
- Debugging is still hard
- Unreadable code if not careful
- May lead to micro-optimization or worst performance than the compiler

Eigen uses the compiler intrinsics, and they fixed and avoided some of the above problems with the right abstraction.

1.3 Compiler Dependent Optimization

The compiler has many tools for optimizing code: loop-unrolling, auto-vectorization, inlining functions, etc. We will rely on code vectorization heavily, but auto-vectorization may or may not be applied if the compiler can infer enough information from the code. To avoid unreliable auto-vectorization, we will use **OpenMP** for explicit vectorization, an open standard and supported by powerful compilers. The main issues are:

- Performance depends on the compiler
- No control over vector instructions or registers
- Auto-vectorization may fail

Boost.uBLAS depends on the auto-vectorization, which does not guarantee the code will vectorize.

1.4 BLAS Routines

There are four BLAS routines that we will implement using **OpenMP**, which uses explicit vectorization and parallelization using threads. Each routine has its chapter, and there we go much deeper with performance metrics.

1. Vector-Vector Inner Product (**?dot**)
2. Vector-Vector Outer Product (**?ger**)
3. Matrix-Vector Product or Vector-Matrix Product (**?gemv**)
4. Matrix-Matrix Product (**?gemm**)

1.5 Machine Model

The machine model that we will follow is similar to the model defined in the [Low et al., 2016], which takes modern hardware into mind. Such as vector registers and a memory hierarchy with multiple levels of set-associative data caches. However, we will ignore vector registers because we let the **OpenMP** handle the registers, and we do not have any control over them. However, we will add multiple cores where each core has at least one cache level that not shared among the other cores. The only parameter we need to put all our energy in is the cache hierarchy and how we can optimize the cache misses.

All the data caches are set-associative and we can characterize them based on the four parameter defined bellow:
C_{L_i}: cache line of the i^{th} level
W_{L_i}: associative degree of the i^{th} level
N_{L_i}: Number of sets in the i^{th} level
S_{L_i}: size of the i^{th} level in Bytes

$$S_{L_i} = C_{L_i} W_{L_i} N_{L_i} \tag{1.1}$$

Let the S_{data} be the width of the type in Bytes.

We are assuming that the cache replacement policy for all cache levels is **LRU**, which also assumed in the [Low et al., 2016] and the cache line is same for all the cache levels. For most of the case, we will try to avoid the associative so that we could derive a simple equation containing the cache size only from the equation 1.1.

1.6 Performance Metrics

1.6.1 FLOPS

It represents the number of floating-point operations that a processor can perform per second. The higher the Flops are, the faster it achieves the floating-point specific operations, but we should not depend on the flops all the time because it might be deceiving. Moreover, it does not paint the whole picture.

$$FLOPS = \frac{Number\ of\ Operation}{Time\ taken} \tag{1.2}$$

1.6.2 Speedup

The speedup tells us how much performance we were able to get when compared to the existing implementation. If it is more significant than one, then reference implementation performs better than the existing implementation; otherwise, if it is less than one, reference implementation performs worse than the existing implementation.

$$Speedup = \frac{Flops_{reference}}{Flops_{existing}} \tag{1.3}$$

1.6.3 Speed-down

The speed down is the inverse of the speedup, and if it is below one, then we performing better than the existing implementation; otherwise, we are performing worse.

$$Speeddown = \frac{Flops_{existing}}{Flops_{reference}} \tag{1.4}$$

1.6.4 Peak Utilization

This tells us how much CPU we are utilizing for floating-point operations when the CPU can compute X amount of floating-point operations.

$$Peak\ Utilization = \frac{Flops}{Peak\ Performance} \times 100 \tag{1.5}$$

1.7 System Information

Processor	2.3 GHz 8-Core Intel Core i9
Architecture	x86
L1 Cache	8-way, 32KiB
L2 Cache	4-way, 256KiB
L3 Cache	16-way, 16MiB
Cache Line	64B
Single-Precision(FP32)	32 FLOPs per cycle per core
Double-Precision(FP64)	16 FLOPs per cycle per core

Peak Utilization found using the formula defined on the [FLOPS, 2021]

$$Peak\ Utilization = Frequency \times Cores \times \frac{FLOPS}{Cycle} \tag{1.6}$$

$$Peak\ Utilization_{FP32} = 588.8\ GFLOPS$$

$$Peak\ Utilization_{FP64} = 294.4\ GFLOPS$$

1.8 Compiler Information

Compiler	Apple clang version 12.0.0 (clang-1200.0.32.29)
Compiler Flags	-march=native -ffast-math -Xclang -fopenmp -O3
C++ Standard	20

1.9 library Version

Intel MKL	2020.0.1
Eigen	3.3.9
OpenBLAS	0.3.13
BLIS	0.8.0

Chapter 2

Inner Product

Chapter 3

Outer Product

Chapter 4

Matrix-Vector Product

Chapter 5

Matrix-Matrix Product

Bibliography

FLOPS. Flops — Wikipedia, the free encyclopedia. = https://en.wikipedia.org/wiki/FLOPS,, 2021. [Online; accessed 24-March-2021].

Tze Meng Low, Francisco D. Igual, Tyler M. Smith, and Enrique S. Quintana-Ortí. Analytical modeling is enough for high-performance BLIS. *ACM Transactions on Mathematical Software*, 43(2):12:1–12:18, August 2016. URL `http://doi.acm.org/10.1145/2925987`.