# Performance Model of Iterated SpMV for Distributed System.

Md Maruf Hossain
Dept. Computer Science
*University of North Carolina at Charlotte*
Charlotte, USA
mhossa10@uncc.edu

Erik Saule
Dept. Computer Science
*University of North Carolina at Charlotte*
Charlotte, USA
esaule@uncc.edu

***Abstract*—Many applications rely on basic sparse linear algebra operations from numerical solvers to graph analysis algorithms. Yet, the performance of these operations is still reasonably unknown. Users and practitioners rely on the rule of thumb understanding of what typically works best for some application domain.**
**This paper aims at providing an overall framework for the distributed system to think about the performance of sparse applications. We use the sparse matrix-vector(SpMV) multiplication as the representative of the experiments. We model the performance of multiple SpMV implementations on distributed systems. We model the performance of different modes of execution of SpMV using linear and polynomial regression models for a distributed system. The models enable us to predict how to partition and represent the sparse matrix to optimize the performance of iterated SpMV on a cluster with 225 cores.**

***Index Terms*—SpMV, MPI, Graph Partitioning**

## I. Introduction

Sparse matrix-vector multiplication(*SpMV*) is one of the fundamental operations in sparse linear algebra. It is critical to solving linear systems and is widely used in engineering and scientific applications [1]–[3]. Distributed memory systems have entered a new age with the popularization of departmental clusters to support these scientific and engineering applications. Many different approaches have been proposed to improve *SpMV* performance on clusters. But choosing the right approach still mostly relies on the rule of thumbs. We posit that building models to predict the performance of different configurations is the only way to make better-informed decisions.

In this paper, we propose a linear and polynomial *Support Vector Regression*(*SVR*) [4] model to predict and analyze the run time of *SpMV* on the distributed system for different ways to execute the operation. Our system will analyze the predicted run time and return the best possible algorithm for the system. The performance of the *SpMV* mainly depends on the size and structure of the matrices and the architecture of the system.

To perform sparse matrix-vector multiplication(*SpMV*) on distributed systems, the most common mechanism is to partition the matrix into multiple parts and perform *SpMV* on each part individually in the different processors. Good partitioning can ensure better load balance and limits the volume of communication between the underlying *MPI process*. Many partitioning algorithms have been proposed to ensure good load balance and to minimize *MPI communications* [5]–[7].

In this paper, we explore two partitioning modes (Uniform 2D-Partitioning and 1D Row Partitioning) and the performance of the different *SpMV* representation based on these partitioning mechanisms on distributed systems. We develop a linear and a polynomial support vector regression (*SVR*) performance models for SpMV operations on the distributed system for these different techniques. The models are accurate enough to predict the best configuration to execute SpMV given a matrix.

## II. SpMV on AVX-512 Architecture (Skylake, Cascade Lake)

SpMV on the distributed system initially can be divided into two parts,

- MPI Communication: latency to share the vector and gather the results from the MPI process.
- Core SpMV Calculation: latency to calculate part of SpMV for a single MPI process.

### A. MPI communication

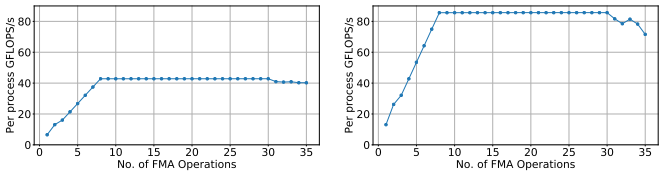
### B. Core SpMV Calculation

We can represent the basic SpMV by $y = y + Val * x$, which contains two floating point operation(multiply and addition). So, we can say we need to calculate NNZ(number of non zeros) times FMA(fused multiply addition) to perform SpMV. In our experiment we divide the core SpMV calculation into two major parts,

- Run time for FMA: $L_{FMA} \times NNZ$, where $L_{FMA}$ is the latency of a single FMA and $NNZ$ is number of nonzeros.
- Read-write latency: $L_{RW}$

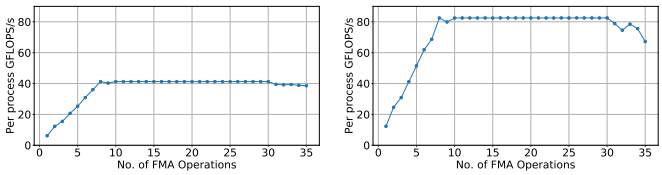
*1) Roofline Model for FMA:* The throughput of Skylake and Cascade Lake are same as 2 instructions per cycle and the latency of FMA is 4 cycles for both of them. So there is a potential of pipelinling($2 \times 4 = 8$) to get the optimal results. Now, both of the architecture has 512-bit register that can give the ability of the vectorization. For 64-bits floating point operation it can give vector width 8 and for 32-bits it can give at max vector width 16. To find the peak performance of the FMA and to avoid the read-write latency we need to setup the

benchmark that datasets can be contained in the register. Now, both of the architecture have 32 registers. We can populate the pipeline by using sufficient ampunt of work. By varying the number of fused-multiply-addition calculation, we can find out the performance limitation. From the informations of the processors, we can say that 4 cycles required for FMA and the throughput of the FMA is 2 instruction per cycle, that means we should at least use $4 \times 2 = 8$ instruction at a time to populate the pipeline.



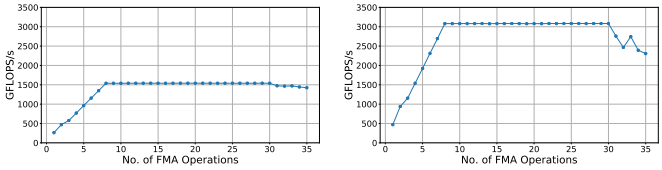

(a) Vector Width: 4 (b) Vector Width: 8

Fig. 1. Skylake: (MPI)Roofline model for bandwidth for FMA operation



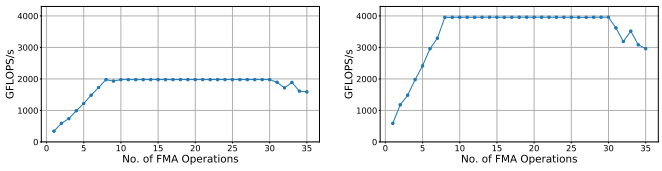

(a) Vector Width: 4 (b) Vector Width: 8

Fig. 2. Cascade Lake: (MPI)Roofline model for bandwidth for FMA operation




(a) Vector Width: 4 (b) Vector Width: 8

Fig. 3. Skylake: (Shared memory parallel)Roofline model for bandwidth for FMA operation




(a) Vector Width: 4 (b) Vector Width: 8

Fig. 4. Cascade lake: (Shared memory parallel)Roofline model for bandwidth for FMA operation

We can estimate the theoritical peak performance a single FMA by the following equation,

$$P = Base\_Clock\_Frequency \times Vector\_Width \times \frac{FLOPs}{Instruction}. \quad (1)$$

## REFERENCES


[1] David F Gleich. Pagerank beyond the web. *siam REVIEW*, 57(3):321–363, 2015.

[2] Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

[3] Tomas Dytrych, Pieter Maris, Kristina D Launey, Jerry P Draayer, James P Vary, Daniel Langr, Erik Saule, MA Caprio, U Catalyurek, and Masha Sosonkina. Efficacy of the su (3) scheme for ab initio large-scale calculations beyond the lightest nuclei. *Computer Physics Communications*, 207:202–210, 2016.

[4] Mariette Awad and Rahul Khanna. Support vector regression. In *Efficient learning machines*, pages 67–80. Springer, 2015.

[5] Mehmet Deveci, Kamer Kaya, Bora Uçar, and Ümit V Çatalyürek. Hypergraph partitioning for multiple communication cost metrics: Model and methods. *JPDC*, 77:69–83, 2015.

[6] George Karypis and Vipin Kumar. Multilevel graph partitioning schemes. In *ICPP (3)*, pages 113–122, 1995.

[7] Kamer Kaya, Bora Uçar, and Ümit V Çatalyürek. Analysis of partitioning models and metrics in parallel sparse matrix-vector multiplication. In *Proc. PPAM*, pages 174–184. Springer, 2013.