Impact of AVX-512 Instructions on Graph Partitioning Problems.

Md Maruf Hossain
University of North Carolina at Charlotte
Charlotte, NC, USA
mhossa10@uncc.edu

Erik Saule
University of North Carolina at Charlotte
Charlotte, NC, USA
esaule@uncc.edu

ABSTRACT

Graph analysis now percolates society with applications ranging from advertising and transportation to medical research. The structure of graphs is becoming more complex every day while they are getting larger. The increasing size of graph networks has made many of the classical algorithms reasonably slow. Fortunately, CPU architectures have evolved to adjust to new and more complex problems in terms of core-level parallelism and vector-level parallelism (SIMD-level).

In this paper, we are exploring how the modern vector architecture of CPUs can help with community detection, partitioning, and coloring kernels by studying two representatives algorithms. We consider the Intel SkylakeX and Cascade Lake architectures, which support gather and scatter instructions on 512-bit vectors.

The existing vectorized graph algorithms of classic graph problems, such as BFS and PageRank, do not apply well to community detection; we show the support of gather and scatter are necessary. In particular for the implementation of the reduce-scatter patterns. We evaluate the performances achieved on the two architectures and conclude that good hardware support for scatter instructions is necessary to fully leverage the vector processing for graph partitioning problems.

ACM Reference Format:
Md Maruf Hossain and Erik Saule. 2020. Impact of AVX-512 Instructions on Graph Partitioning Problems.. In *ICPP '21: The 50th International Conference on Parallel Processing, August 9th, 2021, Chicago, Illinois, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Graphs are at the center of most modern applications today: city and road analysis [24], social media analysis [10, 12, 25], biological data processing and medical research [12, 14], academic networks [30], intelligence [16]. And with the advent of the big data era, graph size has grown exponentially in recent years. We are particularly interested here in partitioning algorithms at large: coloring [6, 18, 21], clustering [28], partitioning [15], community detection [2, 25]. Recent interest in fast graph algorithms has met with a new look at how computer architectures can leverage. GPUs have been understandably popular because of the high flop rate, high memory bandwidth, and high power efficiency for graph problems [4]. CPU architectures have reacted by increasing core count but also by increasing SIMD width in a move to catch up in terms of performance and energy efficiency. In particular, modern Intel processors support AVX-512. These SIMD operations bring the expectation to provide higher energy efficiency than increasing the number of cores.

In this paper, we consider the use of these new instructions to solve graph problems in the class of partitioning. We pick two graph partitioning algorithms, namely a speculative parallel greedy algorithm for graph coloring and the Louvain method for modularity optimization, as representative of graph partitioning algorithms. Section 3 describes these two problems. And we will study their performance on two different processors architecture; Intel Cascade Lake and SkylakeX.

With support for scatter operations, we designed, in Section 4, a strategy called ONPL, for One Neighbor Per Lane. Scatter operations enable us to write to the color of groups of neighbors at once. The operation in the Louvain Method adds some affinity values to the neighboring communities. Because the same community may appear multiple times, we call this operation a reduce-scatter, and we provide two implementations of this operation for different use cases.

Then, we show that the vectorization of these algorithms on x86-64 processors is impractical if they do not support scatter operations. Indeed the only feasible strategy in such a case is to use the different lanes of the vector to process different vertices at the same time. While this strategy applies to classic problems like BFS or SpMV, it requires reordering the graph so that no two vertices in a block of 16 vertices are neighbors for partitioning problems. This strategy only makes sense for the Louvain Method. The derived algorithm, presented in Section 5, is OVPL for One Vertex Per Lane.

Section 6 presents the experimental settings, the code base used as baselines, and the set of graphs to be analyzed. Section 7 presents experimental results which show that ONPL can outperform the scalar implementation for graph coloring for some graphs. The Louvain Method is more computationally expensive. And using *ONPL* and *OVPL* in NetworKit leads to performance improvement on both architectures.

2 NOTATIONS

A graph is denoted by $G = (V, E)$ where V and E represent the vertex and edge set respectively. Edges are represented by (u, v) pair and are associated with an edge weight $\omega : E \rightarrow \mathbb{R}^+$. We use ζ to represent the community set and communities are represented by distinct integers. We use $N(u)$ to represent the neighbor set of a vertex $u \in V$. The *volume* of a node and a community are defined as $vol(u) = \sum_{\{u,v\}:v \in N(u)} \omega(u, v) + 2 \times \omega(u, u)$ and $vol(\zeta) = \sum_{u \in \zeta} vol(u)$ respectively.

Chicago '21, August 9th, 2021, Chicago, Illinois, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 GRAPH PARTITIONING PROBLEMS

Graph partitioning problems are seen here as a large class of graph algorithms that encompass graph coloring algorithms [6, 18, 21], partitioning to minimize edge cuts [15], modularity optimizing community detection algorithms [2, 25], overlapping community detection algorithms [32], label propagation, and certainly many others. All these algorithms have a similar structure in that each vertex is associated with a group of vertices (or multiple groups), and when considering the neighbors of a vertex, the group the neighbor belongs to is the key information rather than the neighbor itself.

We picked two classical partitioning algorithms to represent this class, namely Greedy Graph Coloring (for graph coloring) and Louvain Method (for non-overlapping modularity optimizing community detection).

3.1 Speculative Parallel Greedy Graph Coloring

The distance-1 graph coloring algorithm assigns colors to the vertices of the graph so that no adjacent vertices have the same color. Minimize the number of colors is an NP-hard problem [9], and that is why various heuristic algorithms have proposed for the problem. In particular, a greedy algorithm can obtain near-optimal solutions [21]. The classic parallel algorithm for graph coloring is a speculative parallel greedy algorithm [3, 27] and presented in Algorithms 1, 2, 3.

Algorithm 1 Iterative Parallel Graph Coloring

Input: $G = (V, E)$

1: $C(v) \leftarrow 0$, for all $v \in V$
2: CONF $\leftarrow V$
3: **while** CONF $\neq 0$ **do**
4: AssignColors(G, C, CONF)
5: CONF $\leftarrow$ DetectConflicts(G, C, CONF)
6: **end while**
7: **return** C

Algorithm 2 AssignColors

Input: $G = (V, E), C$, CONF

1: Allocate private FORBIDDEN with size max degree
2: **for** $v \in$ CONF in parallel **do**
3: FORBIDDEN $\leftarrow false$
4: FORBIDDEN($C(u)$) $\leftarrow true$ for $u \in adj(v)$
5: $C(v) \leftarrow \min\{i > 0 | \text{FORBIDDEN}(i) = false\}$
6: **end for**
7: **return** C

Algorithm 1 represents an iterative parallel graph coloring. It takes a graph G with vertex set V and edge set E as an input. It first initializes the set of colors C for all vertices by 0 and a set of conflicts **CONF** by all vertices. It will iteratively color the vertices in **CONF** using a speculative greedy algorithm. And then check whether two neighboring vertices use the same color in which case they are in conflict and need to be colored again.

Algorithm 3 DetectConflicts

Input: $G = (V, E), C$, CONF

1: NEWCONF $\leftarrow 0$
2: **for** $v \in$CONF in parallel **do**
3: **for** $u \in adj(v)$ **do**
4: **if** $C(u) = C(v)$ and $u < v$ **then**
5: ATOMIC NEWCONF $\leftarrow$ NEWCONF $\cup\, v$
6: **end if**
7: **end for**
8: **end for**
9: **return** NEWCONF

Algorithm 2 is the algorithm that will be vectorized and handles the assignment of the color to vertices. It takes graph G, a set of color C and a set of conflicts **CONF** as input. It traverses all the conflict vertices and finds out all the forbidden colors **FORBIDDEN** for the particular vertex. To do that it iterates all its neighbors and track down their colors. Line 4 of Algorithm 2 represents this operation. After collecting all the forbidden colors, it assigns to the vertex the first color that is not in the **FORBIDDEN** set.

Algorithm 3 detects conflicts that could arise during parallel speculative coloring. It takes a graph G, a set of color C, and a previous conflict set **CONF** as input. It defines a new empty conflict set **NEWCONF**. It considers all the previous conflict set of vertices in parallel and for each visits the neighbors to detect if the edge has both ends with the same color. In that case, one of the two vertices is added to the new conflict set atomically.

3.2 Parallel Louvain Method

The *modularity* is defined as the fraction of edges that fall within the partitions minus the expected fraction that would be within the partition if the edges are distributed randomly. This definition enables to greedily optimize modularity by considering moving a vertex to one of its neighbor community. Indeed, if a node $u \in C$ moves to the neighboring community D, then the modularity gain is $\Delta mod(u, C \rightarrow D) = \frac{\omega(u, D/\{u\}) - \omega(u, C/\{u\})}{\omega(E)} + \frac{(vol(C/\{u\}) - vol(D/\{u\})) * vol(u)}{2 * \omega(E)^2}$

The Louvain Method, first proposed by Blondel *et al.* [2], is one of the most popular methods to extract communities from a large network. It is a greedy multilevel algorithm that uses modularity as the objective function [29]. It alternates between two phases, the *Move Phase*, and the *Coarsening Phase*. In the *Move Phase*, nodes are repeatedly moved to adjacent communities to maximize modularity. This process repeats until the communities are stable. Then, the graph goes through a *Coarsening Phase* where each community collapse into a single vertex. The coarsened graph is then recursively processed with the same two phases. In that sense, the Louvain Method is representative of multi-level partitioning algorithms, such as [15].

The *Move Phase* (Algorithm 4) considers all the vertices in the network. For each vertex $u \in V$, for each neighbor $v \in N(u)$, it calculates the modularity difference between having u in its

Algorithm 4 Louvain Method: Move Phase

Input: graph $G = (V, E, \omega)$, communities $\zeta : V \rightarrow \mathbb{N}$
Result: communities $\zeta : V \rightarrow \mathbb{N}$

```
repeat
  for u ∈ V do
    δ ← max_{v∈N(u)}{Δmod(u, ζ(u) → ζ(v))}
    if δ > 0 then
      C ← ζ(arg max_{v∈N(u)}{Δmod(u, ζ(u) → ζ(v))})
      ζ(u) ← C
    end if
  end for
until ζ stable
return ζ
```

current community and moving it to the community of v. The decision of highest modularity gain is retained, and it is enacted if the modularity gain δ is positive. The algorithm repeats until no vertex changes community.

For each vertex u, the *move phase* is split into two parts. First, calculate the affinity(*measures of similarity between pairs of vertices*) of each neighboring community $\zeta(v)$ by adding edge weight $\omega(u, v)$ of the each neighbor v of u. Second, assign the node to the community of highest affinity.

The affinity calculation of a vertex is the computationally expensive part of the algorithm. It is the part that we vectorize in this paper. We do not describe the *Coarsening Phase* since we will not make any changes to it. In this work, we only investigate the performances of the *Move Phase* of the *Louvain* method.

Many parallel methods exist to detect communities in massive networks. The most recent effort is included in NetworKit [31], GRAPPOLO [20] and studied in [13, 29]. GRAPPOLO uses a different and more complex algorithm than NetworKit. For simplicity, we present the Parallel Louvain Method (PLM), used by NetworKit.

PLM [29] is a shared-memory parallelization of the Louvain Method [2]. The algorithm performs the move phase in parallel by giving each thread different vertices to compute the affinity and their assignment to communities. It then coarsens the graph and recursively performs its optimizations. The runtime of PLM is mostly dictated by the first `move` phase; the process of converging the communities on the original graph, before any coarsening in done [29]. Trying to move vertices in parallel is not a race condition free process. Indeed, the algorithm may attempt to move two adjacent vertices simultaneously. PLM is optimistic and assumes that only a few benign race conditions will happen in practice. However, race conditions may cause the process not to converge; PLM stops the move phase after 25 iterations, whether communities have converged or not.

In practice, Parallel Community Detection codes have limited multi-core scalability [29]; in particular because of the noted convergence issues. Since using multiple core reaps little benefit, this paper focuses on using each core more efficiently by leveraging vector SIMD operations. And we consider improving multi-core scalability orthogonal to this work.

4 ONPL: ONE NEIGHBOR PER LANE

The first strategy that we investigate, One Neighbor Per Lane (ONPL), uses the entire vector to process different neighbors of the same vertex.

4.1 Speculative Greedy Graph Coloring

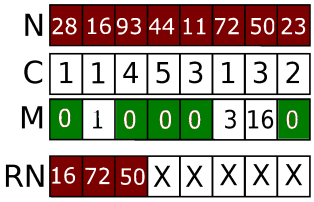
For graph coloring, the conflict detection method naturally vectorizes. Vectorization will be useful when marking which colors can not be used for a vertex. One can vectorize the loop that considers all the neighbors of a vertex. The operation boils done to loading 16 neighbors at a time with a `load` instruction; load the colors of these neighbors using a `gather` instruction. Then marking the used colors using `scatter` instruction. Identifying the first available color and identifying conflicting coloring vectorize naturally.

4.2 Louvain Method

Vectorized affinity calculation is complex because if two neighbors in the same community appear in the same vector, their contribution to the affinity of that community compounds. It will lead to conflicts during affinity calculation that requires resolving. We present the *One Neighbor Per Lane(*ONPL) vectorized Louvain method for community detection using intrinsic notations.

4.2.1 Affinity Calculation. In `AVX-512`, the registers are 512 bits large so that it enables the ability to load 16 neighbors of a vertex at a time to process. Computing the affinity values requires a sequence of load, gather, addition, and scatter operations. Vectorized affinity will work well if all the vertices have their neighbors in different communities. Otherwise, blindly scattering causes some of the updates to be discarded, leading to incorrect affinity values. It requires summing the edge weights of every community before accessing the current affinity of adjacent communities. This operation is essentially a *reduce and scatter*. Unfortunately, no instruction directly does this operation. But the `AVX-512F` and `AVX-512CD` instruction sets enable two different ways to handle this. *ONPL* uses either one of them, depending on circumstances and these two instruction sets are sufficient to implement the algorithm.

Consider the extreme case where all the communities in the vector are different. It is typical at the beginning of the execution of the community detection code. In such a case, the addition and scattering can occur independently without requiring any reduction. If we know that all the lanes are independent, then no two lanes will write to the same location. Fortunately, the AVX-512CD instruction set provides `_mm512_conflict_epi32` instruction that tests each 32-bit element of an array `A` for equality with all other elements in `A` closer to the least significant bit. Each element's comparison forms a zero extended bit vector in `dst`. This instruction(`_mm512_conflict_epi32`) is the basis of the *conflict detection* method for reduce and scatter as it enables the extraction of different sets of communities and neighbors that can safely process at the same time. Figure 1(a) represents the process. Here, N is the list of neighbors of a vertex, and C is the corresponding list of communities. Instruction(`_mm512_conflict_epi32`) is applied on C to calculate the mask M. Figure 1(b) shows the code snippet to calculate the mask M. There are two techniques to handle the conflicted case: the first iteratively performs the vector operation on



(a) Conflict Detection.

```
const   __m512i set0 = _mm512_set1_epi32(0x00000000);
/// Load at most 16 neighbor vertices.
__m512i N = _mm512_loadu_si512((__m512i *) &pnt_outEdges[i]);
/// Gather community of the neighbor vertices.
__m512i C = _mm512_mask_i32gather_epi32(set0, self_loop_mask, N, &zeta[0], 4);
/// Detect conflict of the community
__m512i C_conflict = _mm512_conflict_epi32(C);
/// Calculate mask M by comparing C_conflict with set0
const __mmask16 M = _mm512_mask_cmpeq_epi32_mask(self_loop_mask, C_conflict, set0);
```

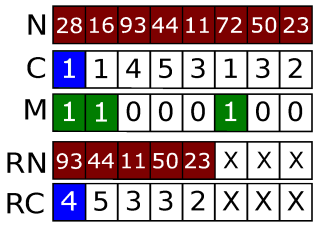
(b) Code snippet to calculate mask M. *pnt_outEdges* represents the list of out edges, *self_loop_mask* is the mask to prevent the self-loop and *zeta* represents the list of community.

Figure 1: Perform reduce scatter using conflict detection. The neighbors (N) are in their Communities (C). A mask (M) is derived from C to denote the entries that will be processed (in green). Some neighbors will remain (RN) to be processed.

the non-conflicted sets and performs as many iterations of vector operations as there are non-conflicted sets; the second one applies vector operation on a non-conflicted set of neighbors only once and performs the remaining entries using purely scalar operations. Indeed, in practice, this conflict detection method uses many instructions. And it only useful if many communities can process at once. The vector will process one entry at a time with expensive vector operations if adjacent vertices belong to the same community. One can avoid the problem by performing vector operation only on the first set of independent communities and use the scalar operations afterward in the conflict detection method.

Another extreme case comes when all the communities in the vector are identical. This case arises when the process has mostly converged. In this case, an *in-vector reduction* is preferable. This method (sketched in Figure 2(a)) masks out all the entries of the vector besides the one mapping to a particular community. Figure 2(b) shows the code snippet to calculate the mask *M*. Then the edge weight mapping to this community is reduced with a masked reduction instruction `_mm512_mask_reduce_add_ps` and is finally added back to the affinity of that community. In Figure 2(a), *RN* represents the remaining vertices that are not processed yet, and *RC* is the list of their corresponding communities. Similar to the conflict detection method, there are two ways to proceed. Successive communities can use mask and reduce; however, this can lead to an issue for vertices that sit at the border of many communities causing potentially a large vector overhead. In practice, ONPL only processes vector operations for the first community of the vector and defaults to scalar implementation for the remaining communities.

The calculation of modularity from the affinity and the assignment of vertices to the community is done with simple vector processing and does not pose particular challenges.



(a) In-vector Reduction.

```
/// Load at most 16 neighbor vertices.
__m512i N = _mm512_loadu_si512((__m512i *) &pnt_outEdges[i]);
/// Gather community of the neighbor vertices.
__m512i C_vec = _mm512_mask_i32gather_epi32(set0, self_loop_mask, N, &zeta[0], 4);
/// Count the number of valid neighbors(u!=v)
sint vertex_cnt = _mm_popcnt_u32((unsigned)self_loop_mask);
/// It will compress communities to left that is not processed yet
C = _mm512_mask_compress_epi32(set0, self_loop_mask, C_vec);
/// Create a mask for Communities that is not processed
__mmask16 comm_mask = pow(2, vertex_cnt) - 1;
index * comm_not_processed = (index *)&C;
/// Handle the first community
__m512i first_comm = _mm512_set1_epi32(comm_not_processed[0]);
/// Calculate Mask for the first community
const __mmask16 M = _mm512_mask_cmpeq_epi32_mask(comm_mask, first_comm, C);
```

(b) Code snippet to calculate mask M. *pnt_outEdges* represents the list of out edges, *self_loop_mask* is the mask to prevent the self-loop and *zeta* represents the list of community.

Figure 2: Perform reduce scatter by compressing the communities. The neighbors (N) are in their Communities (C). A mask (M) is derived from C to denote the entries that will be processed (in green). Some neighbors(RN) and communities(RC) will remain to be processed.

5 OVPL: ONE VERTEX PER LANE

In the One Vertex Per Lane (OVPL) method, each SIMD lane processes different vertices. Initially, vertices of the graph are group into multiple blocks where the size of blocks is the multiple of the vector lanes. We have to restructure the network for the efficiency and convergence of the algorithm.

Because two vertices in a block will be processed simultaneously, OVPL requires two vertices in the same block not to be neighbors. Reordering the graph to have that property requires solving a graph coloring problem. Therefore it makes no sense to deploy OVPL for graph coloring. We only consider OVPL for the community detection problem.

5.1 Preprocessing

Vertices that are part of the same block will always be processed simultaneously. This property might induce race conditions that can prevent convergence. If the adjacent vertices are processed simultaneously, the affinity calculation performs on the changing information. The simplest case is a graph with two vertices that swaps their community infinitely, but the issue also appears on numerous complex networks.

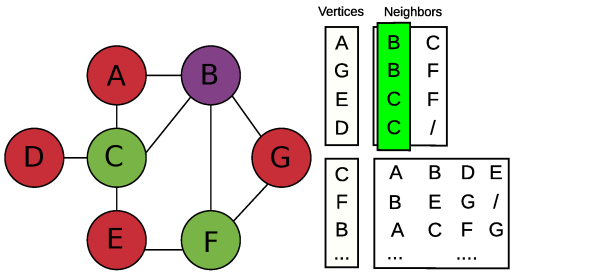
To prevent this from happening, we first solve a graph coloring problem: we allocate a color to each vertex so that no two adjacent vertices have the same color. We then group the vertices where each group holds vertices with the same color. That will make sure that no vertices are adjacent in a group. While finding the coloring with a

minimal number of colors is an NP-Complete problem [9], we do not require such a high-quality solution. We use the speculative parallel greedy graph coloring algorithm [3] we described in Section 3.1.

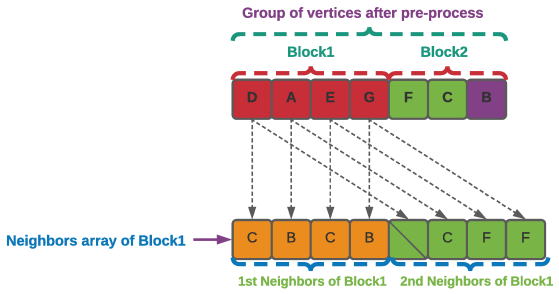
After grouping the vertices, we sort the vertices in each group by non-increasing degrees. Sorting will help to minimize wasted computation during execution.

Finally, we split each group of non-adjacent vertices into small blocks of equal size equal to a multiple of the number of lanes. We reformat the vertices of each block to enable vectorization by interleaving the representation of the different vertices. That also reduces unaligned memory accesses. The format is similar to sliced ELLPACK [22]. A contiguous memory of size $max\_deg\_of\_block \times block\_size$ holds each block of vertices. The index from $(i-1) \times block\_size$ to $i \times block\_size$ will represent the i^{th} neighbor of the vertices of each block. Edge weights also follow a similar representation.

Figure 3 shows a sample graph and its block structure. In the example, we assume the vector length is 4 for readability (instead of 16). So, the initial block will hold vertices that are not adjacent by selecting the same color. But in the second group, there are no four vertices with the same color; that is why it contains vertices of different colors to fill the vector.



(a) Sample Graph (b) Abstract Memory Representation



(c) Physical Memory Representation

Figure 3: OVPL reorders the graph using a graph coloring methods and structure it in blocks of vertices so that the neighbors of the vertices of a block can be loaded in a vector (sketched in green) simultaneously.

5.2 Moving a Block of Vertices

Rather than moving a vertex to its most preferable community, OVPL *moves* a block of vertices at once. It calculates the *affinity* of all the vertices of a block concurrently. Therefore OVPL has a much higher memory utilization than PLM because it keeps *block_size* affinity structures in memory.

OVPL computes the affinity of each vertex of the block one neighbor at a time. OVPL first loads the first neighbor of each vertex of the block at once and gathers the community of the first neighbors. Then it gathers the affinity of the neighbor communities from the different affinity arrays. OVPL adds the edge weights to the obtained affinity and scatters the updated values back to the appropriate locations. Note that because of this, it was not possible to perform this vectorization on x86 processors before scatter was introduced with AVX-512.

This process repeats until all the neighbors of all the vertices of the block are processed, i.e., until the maximum degree of the block. However, some vertices may have a lower degree, so OVPL needs to check the existence of the neighbor. This check increases the number of instructions and causes the algorithm to use masked vector instructions. OVPL does not perform that check before the minimum degree of the block neighbors has been considered. The difference between the maximum and minimum degree in each block leads to wasted SIMD lanes. Preprocessing sorted the different color groups per degree to minimize the degree difference. Also representing the blocks by interleaving the vertices, enables access to the graph to aligned loads.

The assignment of vertices to new communities is done without particular optimization using a natural way of performing this task.

6 EXPERIMENTAL SETTINGS

Hardware Platform and Operating System. We used two different machines for the two architectures we study in this paper. We refer to the first machine as `SkylakeX`. It is a node with two Intel Xeon Gold 6154 processors (SkylakeX architecture, 18 cores per processor, no hyperthreading, 25MB L3 Cache) and 388 GB of DDR4 memory. The second machine is `Cascade Lake`, which is equipped with two Intel Xeon Gold model 6248R (Cascade Lake architecture, 24 cores per processor, no hyperthreading, 36MB L3 Cache) and 384GB GB of DDR4 memory. Both processors support Intel `AVX-512F` and `AVX-512CD` instruction sets with among others. Both machines use Linux `3.10.0`.

Software Environment. All the codes are compiled by the Intel C++ compiler `icpc` version `16.0.0.109`. Codes also compile with optimization flag `-O3` and `xCORE-AVX512` flags, so the compiler generates a binary optimized for the architecture. We pick existing established code bases for both algorithms to confirm we start from implementations of reasonable good qualities.

We build graph coloring and community detection experiments on top of Kokkos [7] and *NetworKit* [31], respectively. We intended to compare to the original PLM implementation from [29]. During our experiments, we realized that PLM suffered from various memory management issues like large buffers were allocated and deallocated for each vertex traversed. We created a Modified PLM implementation (MPLM) that preallocates memory per thread. And

then reuse the same buffer for the computation rather than deallocating and reallocating memory over and over. After confirming that MPLM is an improvement on PLM (See section 7.2.1), we will perform all other comparisons with MPLM.

Graphs. We perform our experiments on real-world data sets to avoid the bias introduced by random graph generator. We select graphs from the *Stanford Large Network Dataset Collection* (SNAP) [17] and *DIMACS* [1, 26] data sets that are well known for graph algorithm research. Graphs are from different categories like Social networks, clustering instances, sparse matrices, internet topology networks, citation networks. We expect that the coverage in the type of graphs enables deriving conclusions that are more general and bias-free than picking all graphs from a single category. Table 1 presents the list of undirected graphs that we use in the experiments. The table also includes basic statistics such as the number of nodes (V), edges (E) of the graph, the maximum degree of the graph (Δ), and average degree (δ).

Table 1: List of graphs used in the experiment

Graph	Nodes (V)	Edges (E)	Δ	δ
333SP	3,712,815	11,108,633	28	5
AS365	3,799,275	11,368,076	14	5
M6	3,501,776	10,501,936	10	5
NACA0015	1,039,183	3,114,818	10	5
NLR	4,163,763	12,487,976	20	5
Oregon-2	11,806	32,730	2,432	5
asia	11,950,757	12,711,603	9	2
belgium	1,441,295	1,549,970	10	2
delaunay_n24	16,777,216	50,331,601	26	5
europe	50,912,018	54,054,660	13	2
germany	11,548,845	12,369,181	13	2
in-2004	1,382,908	13,591,473	21,869	19
kkt_power	2,063,494	6,482,320	95	6
loc-Gowalla	196,591	950,327	14,730	9
luxembourg	114,599	119,666	6	2
netherlands	2,216,688	2,441,238	7	2
nlpkkt200	16,240,000	215,992,816	27	26
roadNet-PA	1,088,092	1,541,898	9	2
uk-2002	18,520,486	261,787,258	194,955	28

Collection of Result Sets. All the variants are run 25 times for each graph. The reported values of time and modularity are average of the 25 runs. For runtime, we only measure the time taken by the community detection(*Move-Phase*) and graph coloring algorithm itself, not the time spent reading the graph from the file system. We computed the 95% *confidence interval* [8] for the results of all the experiments. Once we realized the confidence intervals were very narrow and that the visible differences in the plots were statistically significant, we choose not to report them to improve figures readability.

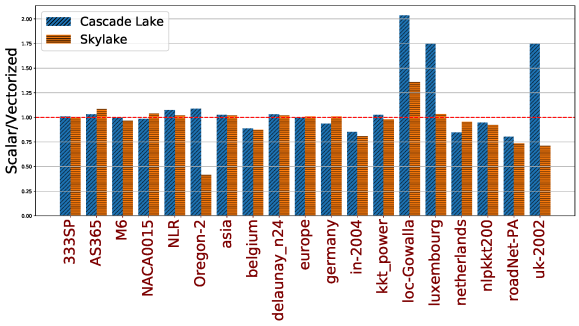


Figure 4: Impact of vectorization of Graph Coloring on both architectures. Y-axis represents the normalized version of the runtime comparison between scalar and vectorized. Scalar/Vectorized = 2.5 means vectorized version is 2 times faster than scalar.

7 PERFORMANCE RESULTS

7.1 Speculative Greedy Graph Coloring

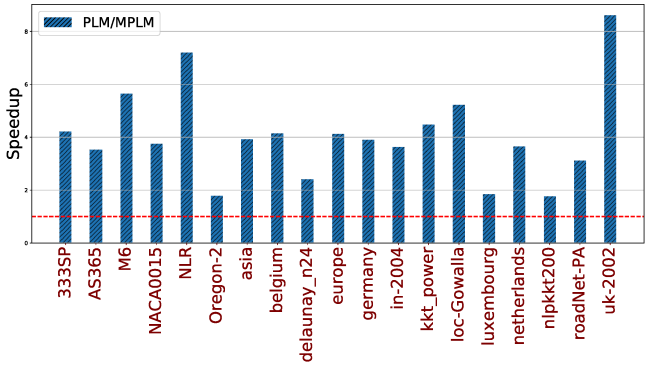
The performance of the ONPL vectorization on graph coloring is displayed in Figure 4 for the *Cascade Lake* and *SkylakeX* architectures. Vectorized speculative graph coloring on both processors shows moderate performance enhancement for some graphs over the scalar version. Vectorized graph coloring on the Cascade Lake and SkylakeX outperform the scalar version by at most factors of 2 and 1.4. Speculative parallel graph coloring has two main parts. One is the assignment of color, and another is conflict detection. We only apply vectorization on the color assignment portion. Graph coloring has a limited opportunity for vectorization that is why it shows a moderate performance for most of the graphs.

7.2 Louvain Method on NetworKit

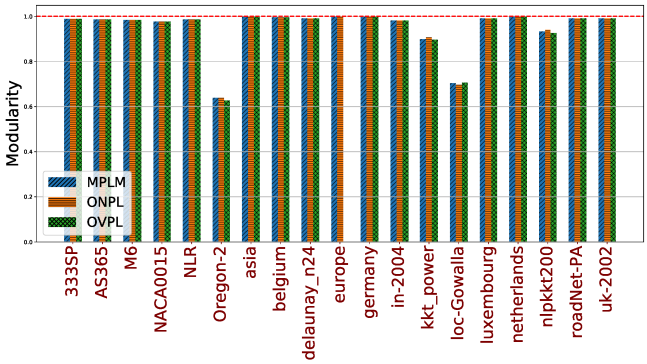
7.2.1 Modified Parallel Louvain Method (MPLM). We noticed some performance deficiencies in PLM, like threads reallocation of the memory needed for the affinity computation for each vertex that it encounters. To be able to study the impact of vector processing, we needed to make sure that the performance difference was rooted in vectorization rather than in memory management. The Modified Parallel Louvain Method (MPLM) is the code that contains various performance fixes for PLM.

Figure 5(a) presents the improvement of MPLM compared to PLM for 48 threads on *Cascade Lake* for all studied graphs. Similar results observe on SkylakeX (not shown for brevity). We will use MPLM as the comparison point to see the impact of vector processing in community detection codes.

7.2.2 Modularity. Since the algorithm has significant race conditions, any change of timings could affect the quality of the communities detected. Modularity is one of the standard metrics to evaluate the quality of the communities and is the metric optimized by MPLM. Figure 5(b) shows the modularity of the implementations of MPLM, ONPL, and OVPL on the *Cascade Lake* architecture using 48 threads. All methods achieve almost the same modularity which confirms the quality of the vectorized communities has not been significantly impacted.



(a) PLM vs MPLM speedup on the Cascade Lake(48 threads).



(b) Modularity of MPLM, ONPL, and OVPL on Cascade Lake(48 threads).

Figure 5: Performance and quality of the Modified PLM (MPLM) over PLM.

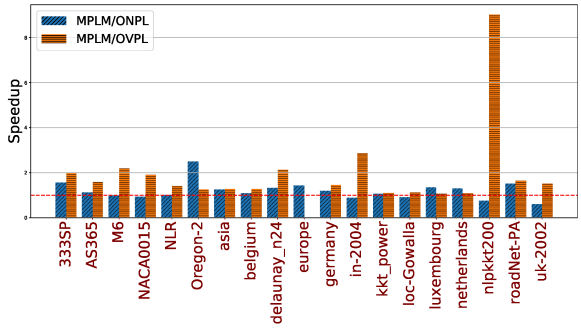


Figure 6: Speedup of ONPL and OVPL over MPLM on the Cascade Lake (48 threads)

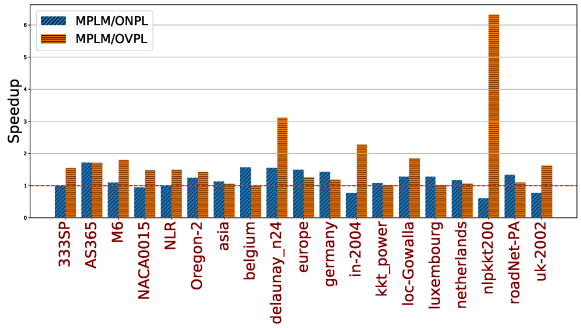


Figure 7: Speedup of ONPL and OVPL over MPLM on the SkyLakeX (36 threads)

7.2.3 ONPL. is a vectorized algorithm with the same memory consumption as the scalar algorithm MPLM and similar memory access patterns. Figure 6 shows the performance of ONPL compared to MPLM on the Cascade Lake for 48 threads: ONPL shows performance improvement for most of the selected graphs and at most a factor of 2.5 performance gain compared to *MPLM*. Figure 7 shows the *ONPL* performance in the NetworKit on the *SkylakeX* architecture. ONPL performs better than its scalar counterpart for almost all the graphs. The best performance of *ONPL* is recorded on the *SkylakeX* processor is around a factor of 1.8 compared to *MPLM*.

7.2.4 OVPL. is an algorithm that consumes a lot more memory than the scalar algorithm due to having to store community affinity information for an entire block of vertices. Figure 6 presents the results of OVPL on the *Cascade Lake* architecture relative to the scalar implementation. For the graphs that were completed (some graphs ran out of memory), the performance derived is much better than the scalar implementation. Figure 7 shows the performance of OVPL on SkylakeX. We can see a factor of 9.0 and 6.5 performance gain for *OVPL* on the *Cascade Lake* and *SkylakeX* processors respectively compared to *MPLM*.

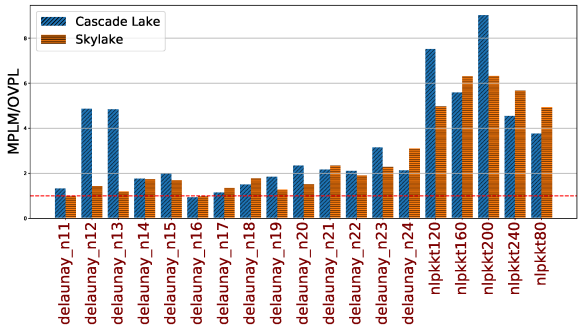


Figure 8: Speedup of *OVPL* over *MPLM* for the selected graphs where many vertices have degrees close to the average on both architectures.

From the algorithm perspective, *OVPL* performs vectorization on a block of vertices, more specifically proper vectorization applies on the iteration only the minimum degree of vertices from the block. The rest of the iterations need more branching and also some lanes of the vectorization always remain unused. Our experimental

results also reflect the scenario. Figure 8 shows only the performance of the selected graphs where most of the vertices have the same degree or very small variations. It shows a great performance gain. Graphs like `Delaunay`(average Degree 5) triangulations of random points or sparse matrix `nlpkkt`(average Degree 26) have most vertices with degrees close to the average. Every vertex in *OVPL*'s block is in sorted order and properly distributed by their degree, which also brings great load balancing.

8 RELATED WORK

Label propagation is one of the most popular community detection algorithm proposed by Raghavan *et al.* [25]. The algorithm iteratively refines labeling of vertices to communities by finding for each vertex the label that most frequently appears in its neighborhood and migrating the vertex to that label. PLM [29] is the shared-memory parallelization of the Louvain Method [2] we use as a reference. Halappanavar *et al.* [13] presented community detection for static and dynamic networks using Grappolo.

Cheong *et al.* [5] proposed a parallel Louvain method for GPUs using three levels of parallelism for the single and multi-GPU architectures. Later, Naim and Manne *et al.* [23] proposed a highly scalable GPU algorithm for the Louvain method, which parallelizes the access to individual edges. There are other recent works like Sanders *et al.* [19] proposed Louvain method for the python; the main objective of their work is the simplicity to implement the algorithm in python language. Gheibi *et al.* [11] proposed a cache efficient Louvain method for Intel Knight Landing(KNL) and Haswell architecture.

Both GPUs and CPUs are SIMD systems, at least in spirit. Taking the analogy of a GPU warp as a core and a thread inside a warp as a lane, algorithms for GPUs can be re-envisioned as vectorized CPU algorithm. At a very high level, the distinction between vertex-based algorithms (such as OVPL) and edge-based algorithms (such as ONPL) appears in GPUs. However, there are still many differences between the architectures which cause engineering and algorithmic decisions for CPU and GPU systems very different.

9 CONCLUSION

We considered the impact of AVX-512 instructions on graph partitioning problems. We investigated, in particular, the *Cascade Lake* and the *SkylakeX* architectures and how to use them to perform speculative greedy graph coloring and the Louvain method. *OVPL* proved to be efficient for graphs with balance and high average degree. The vectorization strategy that processes multiple neighbors of a single vertex at once also shows great performance. That strategy is only possible thanks to scatter instructions and other various new instructions in `AVX-512` that are critical to partitioning problems. The reduce and scatter pattern is critical in implementing these vectorizations. These had to be implemented by intrinsic operations in our software environment. In future works, we want to investigate compiler techniques to enable us to deploy these techniques on more graph partitioning kernels without requiring expert programmers.

ACKNOWLEDGEMENT

This work is supported by grant from the National Science Foundation CCF-1652442 and was made possible by a computing allocation given by TACC through XSEDE.

REFERENCES

[1] David A Bader, Henning Meyerhenke, Peter Sanders, and Dorothea Wagner. 2013. *Graph partitioning and graph clustering*. Vol. 588. American Mathematical Society Providence, RI.
[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008.
[3] Ümit V. Çatalyürek, John Feo, Assefaw H. Gebremedhin, Mahantesh Halappanavar, and Alex Pothen. 2012. Graph Coloring Algorithms for Multi-core and Massively Multithreaded Architectures. *Parallel Comput.* 38, 10-11 (Oct-Nov 2012), 576–594.
[4] Chun Yew Cheong, Huynh Phung Huynh, David Lo, and Rick Siow Mong Goh. 2013. Hierarchical Parallel Algorithm for Modularity-Based Community Detection Using GPUs. In *Proc EuroPar*. 775–787.
[5] Chun Yew Cheong, Huynh Phung Huynh, David Lo, and Rick Siow Mong Goh. 2013. Hierarchical parallel algorithm for modularity-based community detection using GPUs. In *Proc. EuroPar*. 775–787.
[6] Mehmet Deveci, Erik G Boman, Karen D Devine, and Sivasankaran Rajamanickam. 2016. Parallel graph coloring for manycore architectures. In *Proc. IPDPS*. 892–901.
[7] H Carter Edwards, Christian R Trott, and Daniel Sunderland. 2014. Kokkos: Enabling manycore performance portability through polymorphic memory access patterns. *JPDC* 74, 12 (2014), 3202–3216.
[8] Bradley Efron and Robert Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science* (1986), 54–75.
[9] Michael R. Garey and David S. Johnson. 1979. *Computers and Intractability*. Freeman.
[10] Ullas Gargi, Wenjun Lu, Vahab S Mirrokni, and Sangho Yoon. 2011. Large-Scale Community Detection on YouTube for Topic Discovery and Exploration.. In *ICWSM*.
[11] Sanaz Gheibi, Tania Banerjee, Sanjay Ranka, and Sartaj Sahni. 2020. Cache Efficient Louvain with Local RCM. In *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 1–6.
[12] Michelle Girvan and Mark EJ Newman. 2002. Community structure in social and biological networks. *PNAS* 99, 12 (2002), 7821–7826.
[13] Mahantesh Halappanavar, Hao Lu, Ananth Kalyanaraman, and Antonino Tumeo. 2017. Scalable static and dynamic community detection using grappolo. In *Proc. IEEE HPEC*. 1–6.
[14] Pall F Jonsson, Tamara Cavanna, Daniel Zicha, and Paul A Bates. 2006. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC bioinformatics* 7, 1 (2006), 2.
[15] George Karypis and Vipin Kumar. 1999. A Fast and Highly Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM SISC* 20, 1 (1999), 359–392.
[16] V. Krebs. 2002. Mapping Networks of Terrorist Cells. *Connections* 24 (2002). Issue 3.
[17] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.
[18] Gary C Lewandowski. 1995. Practical implementations and applications of graph coloring. (1995).
[19] Tze Meng Low, Daniele G Spampinato, Scott McMillan, and Michel Pelletier. 2020. Linear Algebraic Louvain Method in Python. In *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 223–226.
[20] Hao Lu, Mahantesh Halappanavar, and Ananth Kalyanaraman. 2015. Parallel heuristics for scalable community detection. *Parallel Comput.* 47 (2015), 19–37.
[21] David W Matula, George Marble, and Joel D Isaacson. 1972. Graph coloring algorithms. In *Graph theory and computing*. Elsevier, 109–122.
[22] Alexander Monakov, Anton Lokhmotov, and Arutyun Avetisyan. 2010. Automatically tuning sparse matrix-vector multiplication for GPU architectures. In *Proc. HiPEAC*. 111–125.
[23] M. Naim, F. Manne, M. Halappanavar, and A. Tumeo. 2017. Community Detection on the GPU. In *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 625–634.
[24] S. Porta, V. Latora, F. Wang, E. Strano, A. Cardillo, S. Scellato, V. Iacoviello, and Messora R. 2009. Street centrality and densities of retail and services in Bologna, Italy. *Environment and Planning B: Planning and Design* 36, 3 (2009), 450–465.
[25] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E* 76, 3 (2007), 036106.
[26] Peter Sanders, Christian Schulz, and Dorothea Wagner. 2014. Benchmarking for graph clustering and partitioning. (2014).
[27] Erik Saule and Ümit V Çatalyürek. 2012. An early evaluation of the scalability of graph algorithms on the intel mic architecture. In *2012 IEEE 26th International*

Parallel and Distributed Processing Symposium Workshops & PhD Forum. IEEE, 1629–1639.

[28] Satu Elisa Schaeffer. 2007. Graph clustering. *Computer science review* 1, 1 (2007), 27–64.

[29] Christian Staudt and Henning Meyerhenke. 2016. Engineering Parallel Algorithms for Community Detection in Massive Networks. *IEEE TPDS* 27 (2016), 171–184.

[30] Christian Staudt, Andrea Schumm, Henning Meyerhenke, Robert Gorke, and Dorothea Wagner. 2012. Static and dynamic aspects of scientific collaboration networks. In *Proc. of ASONAM*. 522–526.

[31] Christian L Staudt, Aleksejs Sazonovs, and Henning Meyerhenke. 2016. NetworKit: A tool suite for large-scale complex network analysis. *Network Science* 4, 4 (2016), 508–530.

[32] Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM CSUR* 45, 4 (2013), 43.