Reordering Library

Deliverable No: D2.3
Deliverable Title: Reordering Library
Deliverable Publish Date: 30 April 2023

Project Title: SparCity: An Optimization and Co-design Framework for Sparse Computation
Call ID: H2020-JTI-EuroHPC-2019-1
Project No: 956213
Project Duration: 36 months
Project Start Date: 1 April 2021
Contact: sparcity-project-group@ku.edu.tr

List of partners:

Participant no.	Participant organisation name	Short name	Country
1 (Coordinator)	Koç University	KU	Turkey
2	Sabancı University	SU	Turkey
3	Simula Research Laboratory AS	Simula	Norway
4	Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa	INESC-ID	Portugal
5	Ludwig-Maximilians-Universität München	LMU	Germany
6	Graphcore AS (until M21)	Graphcore	Norway

CONTENTS

1 INTRODUCTION

The SparCity project is funded by EuroHPC JU (the European High Performance Computing Joint Undertaking) under the 2019 call of Extreme Scale Computing and Data Driven Technologies for research and innovation actions. SparCity aims to create a supercomputing framework that will provide efficient algorithms and coherent tools specifically designed for maximizing the performance and energy efficiency of sparse computations on emerging High Performance Computing (HPC) systems, while also opening up new usage areas for sparse computations in data analytics and deep learning.

Sparse computations are commonly found at the heart of many important applications, but at the same time it is challenging to achieve high performance when performing the sparse computations. SparCity delivers a coherent collection of innovative algorithms and tools for enabling high efficiency of sparse computations on emerging hardware platforms. More specifically, the objectives of the project are:

- to develop a comprehensive application and data characterization mechanism for sparse computation based on the state-of-the-art analytical and machine-learning-based performance and energy models,
- to develop advanced node-level static and dynamic code optimizations designed for massive and heterogeneous parallel architectures with complex memory hierarchy for sparse computation,
- to devise topology-aware partitioning algorithms and communication optimizations to boost the efficiency of system-level parallelism,
- to create digital SuperTwins of supercomputers to evaluate and simulate what-if hardware scenarios,
- to demonstrate the effectiveness and usability of the SparCity framework by enhancing the computing scale and energy efficiency of challenging real-life applications.
- to deliver a robust, well-supported and documented SparCity framework into the hands of computational scientists, data analysts, and deep learning end-users from industry and academia.

1.1 OBJECTIVES OF THIS DELIVERABLE

The objective of this deliverable is to provide a technical overview of the reordering library produced within WP2, as well as the analysis performed using this library.

1.2 WORK PERFORMED

The following software was created to facilitate sparse matrix reordering:

1. Source code for Libmtx 0.4.0, which can be used to perform matrix reordering with Reverse Cuthill-McKee, Nested Dissection and Graph Partitioning based on METIS.
2. Source code for a utility that can be used to perform matrix reordering with Hypergraph Partitioning based on PaToH.
3. Source code for SparseBase 0.3.1, which can be used to perform matrix reordering with Approximate Minimum Degree and Gray order.

4. Source code for SpMV with two different kernels for matrices in compressed sparse row (CSR) format.
5. Source code for sparse Cholesky factorization and for counting fill-in.
6. Source code for computing matrix features.

To explain the performance behaviors, we devise a set of metrics and order-dependent matrix features and attempt to correlate them with the reordering performance. These features include objectives that matrix reordering algorithms attempt to minimise, such as matrix bandwidth, profile and others, as well as the amount of load imbalance in parallel SpMV.

1.3 DEVIATIONS AND COUNTER MEASURES

There are no noteworthy deviations from the original work plan.

1.4 RESOURCES

The software for the reordering libray, along with our experimental results, can be found at: Trotter, James D. (2023). Software for "Bringing Order to Sparsity: A Sparse Matrix Reordering Study on Multicore CPUs" (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.7837264

2 SPARSE MATRIX REORDERINGS

Sparse matrices arise from a wide variety of problems in scientific computing, graph theory, finance, and deep learning. Sparse matrix reordering is an optimization technique used to improve the efficiency of operations on sparse matrices by rearranging their rows and columns. Matrix reordering serves many purposes. It can be used to achieve lower work and storage requirements, improve data locality and cache reuse, expose additional parallelism or improve the effectiveness of other optimization techniques. Sparse direct solvers rely heavily on appropriate orderings to reduce fill-in during factorization, whereas iterative solvers can benefit from reordering through improved data locality.

Various reordering algorithms[1] have been proposed over the years. In the case of sparse direct solvers, it is well known that the right ordering can drastically reduce the number of operations required to perform factorisation.[2] For sparse matrix-vector multiplication (SpMV), one of the most frequently encountered sparse matrix operations, there are some examples of reordering being used to significantly improve performance (e.g., by a factor of 3.6×[3]). However, reordering faces several challenges, including the difficulty of finding an optimal ordering, matrices already having an efficient ordering, or the new ordering causing performance degradation by introducing load imbalance in parallel computations. Additionally, orderings that benefit one architecture may not be useful or even harmful for others.

To demonstrate this with a concrete example, Figure 1 displays a few matrices with their original sparsity patterns, along with their patterns after applying three frequently used reorderings. Additionally, the figure indicates the speedup (or slowdown) of SpMV over the unordered matrix in two different platforms. The figure highlights three main observations: (i) different reorderings lead to a diverse distribution of matrix nonzeros, which consequently results in significant performance improvement or degradation depending on the algorithm-matrix pair; (ii) although a reordering algorithm can enhance the performance of one matrix, it may reduce the performance of another; (iii) the efficacy of a reordering algorithm depends on the architecture.

[1] E. Cuthill and J. McKee. "Reducing the Bandwidth of Sparse Symmetric Matrices". *Proceedings of the 1969 24th National Conference*. Association for Computing Machinery, 1969, pp. 157–172. DOI: 10.1145/800195.805928; Wai-Hung Liu and Andrew H Sherman. "Comparative analysis of the Cuthill-McKee and the reverse Cuthill-McKee ordering algorithms for sparse matrices". *SIAM Journal on Numerical Analysis* 13.2 (1976), pp. 198–213; Patrick R. Amestoy, Timothy A. Davis, and Iain S. Duff. "Algorithm 837: AMD, an Approximate Minimum Degree Ordering Algorithm". *ACM Trans. Math. Softw.* 30.3 (2004), pp. 381–388. ISSN: 0098-3500. DOI: 10.1145/1024074.1024081; Alan George. "Nested Dissection of a Regular Finite Element Mesh". *SIAM Journal on Numerical Analysis* 10.2 (1973), pp. 345–363. DOI: 10.1137/0710032; Alan George and Joseph WH Liu. "The evolution of the minimum degree ordering algorithm". *SIAM Review* 31.1 (1989), pp. 1–19; J. R. Gilbert and R. E. Tarjan. "The Analysis of a Nested Dissection Algorithm". *Numer. Math.* 50.4 (1987), pp. 377–404. ISSN: 0029-599X. DOI: 10.1007/BF01396660; Haoran Zhao et al. "Exploring Better Speculation and Data Locality in Sparse Matrix-Vector Multiplication on Intel Xeon". *2020 IEEE 38th International Conference on Computer Design (ICCD)*. 2020, pp. 601–609. DOI: 10.1109/ICCD50377.2020.00105.

[2] Patrick R. Amestoy et al. "Analysis and Comparison of Two General Sparse Solvers for Distributed Memory Computers". *ACM Trans. Math. Softw.* 27.4 (2001), pp. 388–421. ISSN: 0098-3500. DOI: 10.1145/504210.504212.

[3] Zhao et al., "Exploring Better Speculation and Data Locality in Sparse Matrix-Vector Multiplication on Intel Xeon".

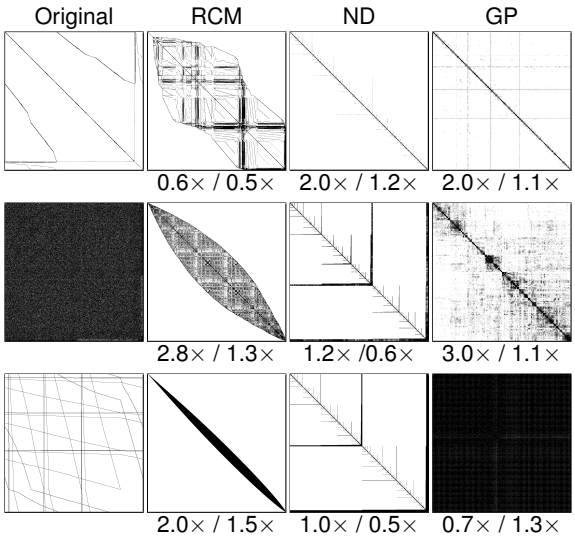

Figure 1 *Matrices reordered with Reverse Cuthill-McKee (RCM), Nested Dissection (ND) and graph partitioning (GP). The numbers below represent speedup (or slowdown) of SpMV on 64-core AMD Epyc Milan and 36-core Intel Ice Lake CPUs, respectively.*

3 REORDERING ALGORITHMS

For describing reordering algorithms, we need to define some key terminology. Many reordering strategies arise from linear solvers where certain features of sparse matrices are desirable. The *bandwidth* of a sparse matrix is the width of the diagonal band that contains its nonzero elements, whereas the *profile* is a sum of distances from the leftmost element to the diagonal of each row (see Section 3.3). In sparse matrix factorization, *fill-in* refers to the appearance of additional nonzero elements in the factorization compared to the original matrix.

3.1 TAXONOMY OF REORDERING ALGORITHMS

To provide an overview, we categorise reordering algorithms into 1) bandwidth-reducing orderings, 2) fill-reducing orderings 3) graph and hypergraph partitioning-based orderings, and 4) other orderings.

3.1.1 BANDWIDTH-REDUCING ORDERINGS.

Well-known examples of such orderings include the Cuthill-McKee (CM) algorithm[4] and the method described by Gibbs et al..[5] The CM ordering attempts to reduce the matrix bandwidth through a breadth-first search of the undirected graph corresponding to a symmetric sparse matrix. The vertices of the graph, which correspond to rows and columns of the matrix, are ordered by choosing a starting vertex (e.g., by finding a pseudo-peripheral vertex[6]) and then traversing the graph in breadth-first search order, where the vertices at each level are sorted in ascending order by degree. In the end, after traversing the entire graph, the ordering may be

[4]Cuthill and McKee, "Reducing the Bandwidth of Sparse Symmetric Matrices".

[5]Norman E. Gibbs, William G. Poole, and Paul K. Stockmeyer. "An Algorithm for Reducing the Bandwidth and Profile of a Sparse Matrix". *SIAM Journal on Numerical Analysis* 13.2 (1976), pp. 236–250. ISSN: 00361429.

[6]Alan George and Joseph W. H. Liu. "An Implementation of a Pseudoperipheral Node Finder". *ACM Transactions on Mathematical Software* 5.3 (1979), pp. 284–295. DOI: 10.1145/355841.355845.

reversed to obtain the more commonly used Reverse Cuthill-McKee (RCM)[7] ordering.

3.1.2 FILL-REDUCING ORDERINGS.

Minimum degree orderings[8] arise in the context of reducing fill-in during sparse Cholesky factorisation. The elimination graph of a sparse symmetric matrix consists of a vertex for every row, as well as edges between any pair of vertices a and b for which row a has a nonzero above the diagonal in column b. At each step of the factorisation, one row is eliminated by removing a vertex and its edges in the elimination graph, and replacing it with a clique consisting of the former neighbours of the vertex. The new edges that are created by forming such a clique lead to fill-in of the Cholesky factor. The minimum degree algorithm is a graph-based heuristic to find node orderings with low amounts of fill by always selecting a vertex of least degree.

Another commonly used fill-reducing ordering is Nested dissection (ND),[9] which is based on computing a vertex separator for the undirected graph of a symmetric sparse matrix. The two subgraphs that arise from removing the separator are ordered first, while rows and columns corresponding to the separator are moved to the end of the matrix. This process is applied recursively for the two subgraphs. The underlying motivation for the ND ordering is that it incurs low fill-in for sparse Cholesky factorization if the separators are small. Since the method relies on graph partitioning, it can be grouped under graph partitioning-based orderings as well.

3.1.3 (HYPER)GRAPH PARTITIONING-BASED ORDERINGS.

Graph partitioning can be used to define an ordering by directly partitioning a matrix into a given number of parts, then grouping rows and columns by their assigned parts. This approach is frequently used in a distributed-memory setting to perform work division of sparse matrix operations, and the same idea can be applied to the shared-memory case.

METIS[10] is a well-known graph partitioning tool that can be used to partition large irregular graphs. It is based on the multilevel paradigm which consists of the graph coarsening, initial partitioning, and uncoarsening phases. The aim of the partitioning is to minimize a partitioning objective, while obeying a load balancing criteria.

Hypergraph partitioning may similarly be used for reordering. PaToH[11] is a commonly-used hypergraph partitioning tool which is known to reflect the actual communication volume requirement of parallel SpMV. Hypergraphs are the generalization of graphs, in which the hyperedges (nets) can be incident to any number of vertices instead of exactly two vertices in simple graphs. The hypergraph partitioning problem is the task of dividing a hypergraph into roughly balanced parts such that the cutsize is minimized. Other reorderings based on hypergraph partitioning include the separated block diagonal form proposed by Yzelman and Bisseling.[12]

[7] Liu and Sherman, "Comparative analysis of the Cuthill-McKee and the reverse Cuthill-McKee ordering algorithms for sparse matrices".

[8] Alan George and David R. McIntyre. "On the Application of the Minimum Degree Algorithm to Finite Element Systems". *SIAM Journal on Numerical Analysis* 15.1 (1978), pp. 90–112. ISSN: 00361429. URL: http://www.jstor.org/stable/2156565; George and Liu, "The evolution of the minimum degree ordering algorithm"; Amestoy, Davis, and Duff, "Algorithm 837: AMD, an Approximate Minimum Degree Ordering Algorithm".

[9] George, "Nested Dissection of a Regular Finite Element Mesh"; Gilbert and Tarjan, "The Analysis of a Nested Dissection Algorithm".

[10] George Karypis and Vipin Kumar. "A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs". *SIAM Journal on Scientific Computing* 20.1 (1998), pp. 359–392. DOI: 10.1137/S1064827595287997.

[11] U.V. Catalyurek and C. Aykanat. "Hypergraph-partitioning-based decomposition for parallel sparse-matrix vector multiplication". *IEEE Transactions on Parallel and Distributed Systems* 10.7 (1999), pp. 673–693. DOI: 10.1109/71.780863.

[12] A. N. Yzelman and Rob H. Bisseling. "Cache-Oblivious Sparse Matrix–Vector Multiplication by Using Sparse Matrix Partitioning Methods". *SIAM Journal on Scientific Computing* 31.4 (2009), pp. 3128–3154. DOI: 10.1137/

3.1.4 OTHER ORDERINGS

A number of alternative matrix orderings have been proposed with the goal of improving data locality in SpMV, including approaches based on the travelling salespesrson problem[13] and space-filling curves.[14] One particular method, which we call Gray ordering,[15] is motivated by microarchitectural concerns to reduce branch mispredictions and improve data locality for SpMV. First, to improve branch prediction, rows with similar nonzero density are grouped together (density reordering). Second, to improve locality, a bitmap-based reordering is applied, where each row is segmented into multiple sections of nonzeros (to construct the row bitmaps), which are then labeled and ordered based on the Gray code.[16] In general, the matrix is first split into dense and sparse submatrices according to the number of nonzeros in each row, while the density and bitmap reorderings are applied depending on the characteristics of each submatrix.

3.2 SPMV KERNELS

As the primary evaluation criterion of reordering algorithms, we assess the performance of sparse matrix-vector multiplication (SpMV) with shared-memory parallel kernels based on the popular compressed sparse row (CSR) format.

A sparse matrix A with M rows and N columns is defined by its K nonzeros, which we denote a_{i_k,j_k} for $k = 1, 2, \ldots, K$, where i_k and j_k are the row and column offsets of the kth nonzero, respectively. Any sparse matrix storage format must somehow store the row and column offsets as well as the nonzero values. The well-known CSR format groups nonzeros by rows in ascending order, and then compresses the row offsets to form row pointers, $r_1, r_2, \ldots, r_{M+1}$, such that r_i and $r_{i+1} - 1$ indicate the location of the first and last nonzeros of row i, respectively. Thus, multiplying A by a vector x to obtain another vector y amounts to computing the sum $y_i \leftarrow y_i + \sum_{k=r_i}^{r_{i+1}-1} a_{i_k,j_k} x_{j_k}$, for every row $i = 1, 2, \ldots, M$.

The standard method for performing the above SpMV computation in parallel is by partitioning the rows into equal-sized, contiguous blocks and assigning one block to each thread. (This is easily achieved in OpenMP with a single `#pragma omp for` directive.) We refer to this as the *1D algorithm*. Although this scheme is simple and works well in some cases, it suffers from load imbalance for many realistic sparse matrices due to the nonzeros being unevenly divided among threads.

As a result, we also consider a second SpMV kernel which is still based on the CSR storage format, but offers a more balanced workload among threads. Rather than partitioning the rows, we instead perform an equal partitioning of the matrix nonzeros. This produces a balanced workload among threads, at least in terms of nonzeros. Loop scheduling must now be performed manually, since the built-in loop scheduling directives of OpenMP are no longer sufficient. In addition, this approach requires each thread to handle its first and final row specially to avoid

080733243.

[13] Ali Pinar and Michael T. Heath. "Improving Performance of Sparse Matrix-Vector Multiplication". *Proceedings of the 1999 ACM/IEEE Conference on Supercomputing*. Portland, Oregon, USA: Association for Computing Machinery, 1999. DOI: 10.1145/331532.331562; D.B. Heras et al. "Modeling and improving locality for the sparse-matrix–vector product on cache memories". *Future Generation Computer Systems* 18.1 (2001), pp. 55–67. ISSN: 0167-739X. DOI: 10.1016/S0167-739X(00)00075-3.

[14] Leonid Oliker et al. "Effects of Ordering Strategies and Programming Paradigms on Sparse Matrix Computations". *SIAM Review* 44.3 (2002), pp. 373–393. DOI: 10.1137/S00361445003820.

[15] Zhao et al., "Exploring Better Speculation and Data Locality in Sparse Matrix-Vector Multiplication on Intel Xeon".

[16] Sardar Anisul Haque and Shahadat Hossain. "A Note on the Performance of Sparse Matrix-vector Multiplication with Column Reordering". *2009 International Conference on Computing, Engineering and Information*. 2009, pp. 23–26. DOI: 10.1109/ICC.2009.40.

Table 1 *Sparse matrix reordering algorithms used in this study*

Name	Reordering Algorithm	Description
RCM[20]	Reverse Cuthill–McKee	bandwidth reduction via breadth-first graph traversal
AMD[21]	Approximate minimum degree	local greedy strategy to reduce fill by selecting sparsest pivot row
ND[22]	Nested dissection	recursive divide-and-conquer using vertex separators to reduce fill
GP[23]	Graph partititoning	METIS multi-level recursive graph partitioning with edge-cut objective
HP[24]	Hypergraph partititoning	column-net hypergraph partitioning with PaToH using cut-net metric
Gray[25]	Gray code ordering	splitting of sparse and dense rows and Gray code ordering

race conditions when updating the output vector y. We call this the *2D algorithm*.

Our 2D algorithm is closely related to and may be considered a simplified version of the merge-based SpMV kernel of Merrill and Garland.[17] Both of these kernels entail a small preprocessing cost to find the balanced partitioning. Even for the more elaborate merge-based kernel, this cost is small enough to keep 2D algorithms competitive. Furthermore, for a given matrix and architecture, this represents a one time cost and can thus easily be amortized over multiple SpMV iterations. Therefore, we ignore this cost in our measurements.

In addition to the reordering code, we provide a custom implementation of the 1D and 2D algorithms for testing. Parallelization is achieved via OpenMP using `static` scheduling. Since our experimental platforms are large NUMA machines, we use the first-touch policy to ensure that the data is placed close to the core using it.

3.3 MATRIX FEATURES FOR REORDERING

We define four matrix features that depend heavily on matrix ordering. These metrics are *bandwidth*, *profile*, *off-diagonal nonzero count*, and *load imbalance factor*, and they are later used to explain the effect of reordering particularly with respect to SpMV performance.

The *bandwidth* and *profile* give an indication of whether nonzeros are clustered near the main diagonal, which in turn may lead to better data locality for SpMV.[18] For an N-by-N sparse matrix A, the bandwidth is the largest distance of any nonzero to the main diagonal, $\max_{a_{i,j}\neq 0} |i-j|$, whereas the profile[19] is a sum over every row of the distance from the leftmost entry to the diagonal, $\sum_{i=1}^{N} i - \min\{j \mid a_{i,j} \neq 0\}$.

Assuming that the matrix is partitioned into N-by-N equal-sized blocks, we count the total number of nonzeros that do not fall into any diagonal blocks. We call this the *off-diagonal nonzero count*, and it is essentially the same as the edge-cut metric which is minimised by graph partitioning, if one assumes that rows are divided equally among the threads, as in the 1D SpMV algorithm.

Finally, to capture effects of load imbalance in shared-memory parallel SpMV, we also consider a *load imbalance factor*. This is defined as the ratio of the maximum number of nonzeros assigned to a single thread to the average number of nonzeros per thread. Thus, if every thread has the same number of nonzeros, the imbalance factor is 1. On the other hand, a thread having twice the number of nonzeros compared to the average yields an imbalance factor of 2.

[17] Duane Merrill and Michael Garland. "Merge-Based Sparse Matrix-Vector Multiplication (SpMV) Using the CSR Storage Format". *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*. Association for Computing Machinery, 2016. DOI: 10.1145/2851141.2851190.

[18] O. Temam and W. Jalby. "Characterizing the behavior of sparse algorithms on caches". *Supercomputing '92: Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*. 1992, pp. 578–587. DOI: 10.1109/SUPERC.1992.236646.

[19] Gibbs, Poole, and Stockmeyer, "An Algorithm for Reducing the Bandwidth and Profile of a Sparse Matrix".

3.4 REORDERING ALGORITHMS IMPLEMENTED

For this study, we have chosen a diverse set of reordering algorithms with relatively different objectives for evaluation. Table 1 gives an overview of our chosen algorithms. We employ the Reverse Cuthill-McKee (RCM) algorithm,[26] which is obtained by simply reversing the usual Cuthill-McKee ordering and is known to work better in practice when used for factorising symmetric, positive definite matrices.

From the fill-in reducing group, variations of the minimum degree reordering algorithm are often used in practice, such as multiple minimum degree (MMD)[27] and approximate minimum degree (AMD).[28] We chose the latter that is based on merely approximating the degree of a vertex, which reduces the runtime complexity. For the AMD and ND orderings, we use reordering routines from SuiteSparse[29] and METIS[30] libraries, respectively.

Next, we include a graph partitioning-based reordering, which will be referred to as *GP*. For GP, we also use METIS, which offers two options for the partitioning objective. The first one is edge-cut, i.e., the number of edges connecting vertices in different partitions, and the second is total communication volume. With respect to partitioning a sparse matrix, the load balance criteria can be chosen to balance the number of rows or the number of nonzeros within the parts. The latter is achieved by weighting each vertex in the graph by the number of nonzeros in the corresponding row. For this study, we use the edge-cut objective and an unweighted graph which implies balancing the number of rows in each part. Moreover, the number of parts to be created by the partitioner is chosen to match the number of CPU cores for the hardware used in this study (see Table 2) by partitioning into 16, 32, 48, 64, 72 or 128 parts.

We also include a hypergraph partitioning reordering, which will be referred to as *HP*. In HP, we employ PaToH with the column-net model in which the rows and columns of the coefficient matrix are represented with vertices and nets, respectively. PaToH includes two metrics that can be used as the partitioning objective, namely cut-net and connectivity metrics. In the column-net model, the former metric corresponds to minimizing the number of nonzero column segments, while the latter corresponds to minimizing the off-diagonal nonzero count. In HP, we adopt the 128-way partitioning of matrices using PaToH with cut-net metric and the same balancing criteria as in GP.

The final algorithm is the Gray code ordering using the parameters suggested by Zhao er al.,[31] meaning that 16 bits are used for bitmap-ordering and rows with more than 20 nonzeros are considered to be dense.

[26] Liu and Sherman, "Comparative analysis of the Cuthill-McKee and the reverse Cuthill-McKee ordering algorithms for sparse matrices".
[27] George and Liu, "The evolution of the minimum degree ordering algorithm".
[28] Amestoy, Davis, and Duff, "Algorithm 837: AMD, an Approximate Minimum Degree Ordering Algorithm".
[29] Ibid.
[30] Karypis and Kumar, "A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs".
[31] Zhao et al., "Exploring Better Speculation and Data Locality in Sparse Matrix-Vector Multiplication on Intel Xeon".

Table 2 *Hardware used in our experiments.*

	Skylake	Ice Lake	Naples	Rome	Milan A	Milan B	TX2	Hi1620
CPUs	Intel Xeon Gold 6130	Intel Xeon Plat-inum 8360Y	AMD Epyc 7601	AMD Epyc 7302P	AMD Epyc 7413	AMD Epyc 7763	Cavium TX2 CN9980	HiSilicon Kunpeng 920-6426
Instr. set	x86-64	x86-64	x86-64	x86-64	x86-64	x86-64	ARMv8.1	ARMv8.2
Microarch.	Skylake	Ice Lake	Zen	Zen 2	Zen 3	Zen 3	Vulcan	TaiShan v110
Sockets	2	2	2	1	2	2	2	2
Cores	2×16	2×36	2×32	1×16	2×24	2×64	2×32	2×64
Freq. [GHz]	1.9–3.6	2.4–3.5	2.7–3.2	1.5–3.3	2.5–3.5	2.5–3.5	2.0–2.5	2.6
L1I/core [KiB]	32	32	64	32	32	32	32	64
L1D/core [KiB]	32	48	32	32	32	32	32	64
L2/core [KiB]	1024	1280	512	512	512	512	256	512
L3/socket [MiB]	22	54	64	16	128	256	32	64
Bandwidth [GB/s]	256	409.6	342	204.8	409.6	409.6	342	342

4 EXPERIMENTAL EVALUATION

4.1 EXPERIMENTAL SETUP

The hardware used in our experiments is shown in Table 2. All codes are compiled with GCC 11.2.0 with the `-O3` and `-march=native` options on each node, and the test systems are running Ubuntu 18.04.6.

Our evaluation relies on the SuiteSparse Matrix Collection.[32] We apply the six reorderings (see Table 1) to 490 matrices that are square, non-complex and have between 1 million and 1 billion nonzeros. On converting the matrices to CSR format, column offsets are stored as 32-bit integers and nonzero values as double precision floating point numbers. In the case of symmetric matrices, whenever an offdiagonal nonzero is encountered, two nonzeros are inserted into the CSR representation, one in the upper and another in the lower triangle of the matrix.

Each SpMV run is repeated 100 times, and we take the maximum performance among the runs. This represents the peak performance of a system with a warm cache and is less susceptible to noise than the average. Note that for smaller matrices used in our evaluation, some or all of the data may fit in last-level cache. For example, the AMD Epyc 7763 has the largest last-level cache at a total of 512 MiB. Only 77 matrices have more than 45 million nonzeros, which is the minimum size needed to exceed the capacity of the last-level cache if matrices are stored in CSR format.

4.2 REORDERING FOR 1D SPMV

Prior to reordering, we observe that the performance of the 1D SpMV algorithm varies greatly from one matrix to another, as expected. As an example, the typical range for the 128-core Milan B is about 50–120 Gflop/s with a median value of about 80 Gflop/s. Our first experiment measures

[32] Timothy A. Davis and Yifan Hu. "The University of Florida Sparse Matrix Collection". *ACM Trans. Math. Softw.* 38.1 (2011). ISSN: 0098-3500. DOI: 10.1145/2049662.2049663.

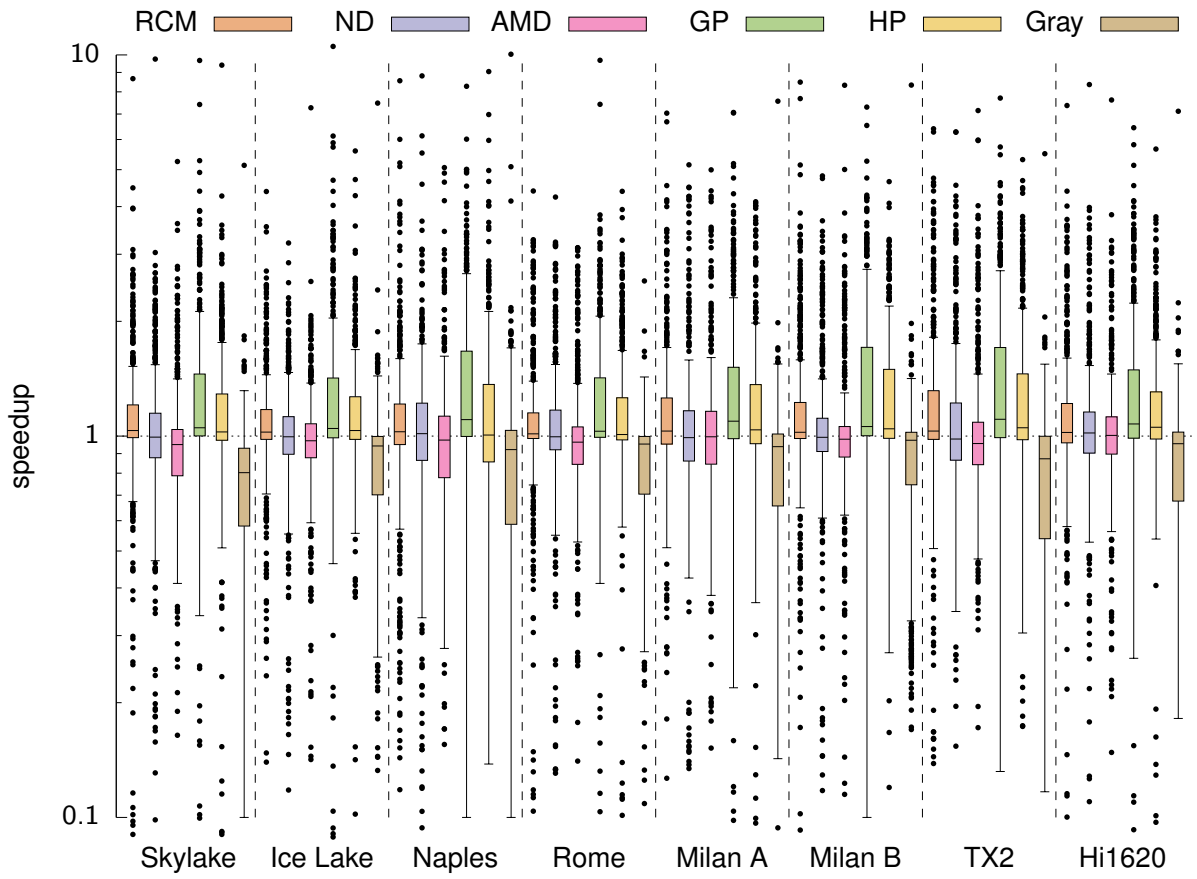

Figure 2 *Speedup of sparse matrix-vector multiplication using 1D algorithm after reordering. For each box, the middle line represents the median and endpoints represent the lower and upper quartiles.*

the SpMV speedup using the 1D algorithm for all reorderings compared to the original ordering, using all 490 matrices in the test set. Figure 2 illustrates the results for all architectures. Note that some outliers are not shown to save space.

The distribution of speedups vary considerably between different reordering techniques. On the other hand, the overall picture is roughly the same regardless of the hardware used. Every ordering has outliers with extreme speedup or slowdown. The greatest slowdown is a factor of 0.05×, whereas the largest speedup is about 40×. If we disregard outliers and consider only the portion between the lower and upper quartiles (i.e., the coloured boxes, which comprises half of the matrices and thus the most typical case), then the speedup ranges from about 0.5 to 1.5×.

With respect to the various orderings, the median speedups of RCM, GP and HP are greater than 1, meaning that SpMV performance improves for more than 50 % of the matrices. Additionally, GP is best with speedup for about 75 % of matrices on every CPU and more matrices achieving higher speedups. This is closely followed by HP, which shows slightly smaller speedups in general and performs noticeably worse on Naples in particular. Next, the median speedups of ND and AMD are close to 1 or slightly less than 1, respectively. These methods are thus equally likely to yield a speedup as they are to result in a slowdown. Finally, the Gray ordering stands out by resulting in slowdowns in most cases. On Skylake in particular, 75 % of matrices reordered with Gray experience a slowdown of 0.9× or worse.

We also compute the geometric mean over the speedups in order to provide a better overview. Results are given in Table 3. They clearly shows that graph partitioning provides far better SpMV

Table 3 *Geometric mean of the speedups of the different reorderings and architectures compared to the original order for all 490 matrices in the 1D algorithm.*

1D	RCM	AMD	ND	GP	HP	Gray	Mean
Skylake	1.054	0.933	0.990	1.189	1.099	0.700	0.981
Ice Lake	1.039	0.939	0.981	1.183	1.100	0.744	0.987
Naples	1.025	0.939	0.978	1.206	1.083	0.757	0.988
Rome	1.032	0.946	0.989	1.197	1.089	0.767	0.994
Milan A	1.039	0.956	0.992	1.198	1.096	0.771	1.000
Milan B	1.048	0.963	0.999	1.212	1.110	0.778	1.009
TX2	1.060	0.969	1.007	1.224	1.123	0.768	1.015
Hi1620	1.061	0.973	1.007	1.228	1.128	0.772	1.017
Mean	1.045	0.952	0.993	1.205	1.103	0.757	0.999

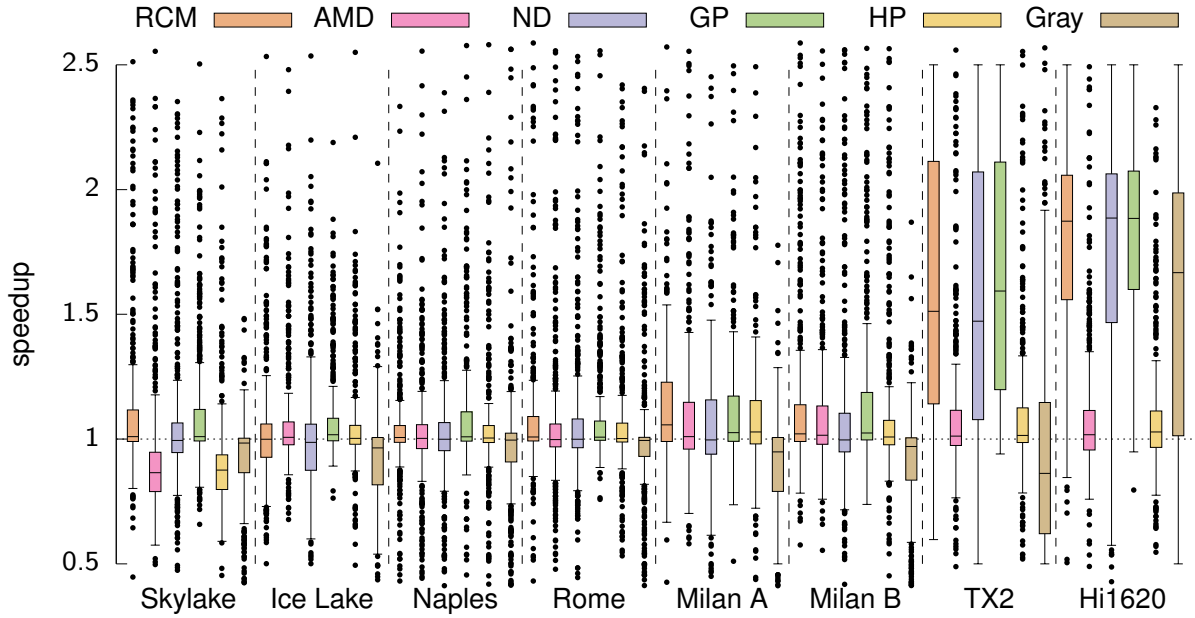

Figure 3 *Speedup of the nonzero-balanced CSR SpMV kernel (2D algorithm) after reordering.*

performance than the alternatives.

4.3 REORDERING FOR 2D SPMV

As discussed in Section 3.2, the 1D algorithm does not ensure load balance. We thus repeat the above experiment for the 2D SpMV algorithm. Results are shown in Figure 3.

Compared to the 1D algorithm, there are fewer and less extreme outliers and the impact of reordering is less pronounced for most architectures. Moreover, the difference between reordering strategies is smaller. On the other hand, the ARM-based CPUs, TX2 and Hi1620, benefit immensely, especially from RCM, ND, and GP. Hi1620 also benefits from Gray ordering in this case. However, we note that the initial performance of both 1D and 2D algorithms on the ARM CPUs is quite low, with median values of 20–30 Gflop/s. We suspect that further tuning and improved compiler support would improve instruction-level parallelism and alleviate performance bottlenecks of the ARM CPUs.

While the 2D algorithm creates a perfect load balance with respect to the number of nonzeros

Table 4 *Geometric mean of the speedups of the different reorderings and architectures compared to the original order for all 490 matrices in the 2D algorithm.*

2D	RCM	AMD	ND	GP	HP	Gray	Mean
Skylake	1.081	0.909	1.034	1.090	0.892	0.906	0.982
Ice Lake	1.043	0.979	1.007	1.088	0.959	0.899	0.994
Naples	1.026	1.005	1.018	1.112	1.003	0.909	1.010
Rome	1.038	1.017	1.028	1.106	1.015	0.920	1.019
Milan A	1.058	1.034	1.036	1.111	1.029	0.908	1.027
Milan B	1.074	1.050	1.045	1.123	1.040	0.899	1.036
TX2	1.134	1.059	1.100	1.186	1.049	0.893	1.066
Hi1620	1.197	1.062	1.160	1.250	1.051	0.944	1.106
Mean	1.080	1.013	1.052	1.132	1.003	0.910	1.029

per thread, execution time may not be fully balanced if cache locality differs in different parts of the matrix and thus in different threads. For that reason, the 2D algorithm can be slower than the 1D algorithm in rare cases. Nonetheless, the 2D algorithm typically observes a considerable speedup over the 1D algorithm for many matrices. For example, for 25 % of the matrices on the Rome processor, a speedup of more than 1.1× or more is observed when comparing the 2D and 1D algorithms with the same ordering, and the largest speedup for an individual matrix is about 10×. Since all other CPUs have more cores, most of their speedups are even higher.

The geometric mean of the speedups is shown in Table 4. The numbers show that most algorithms improve while the speedups of GP and HP are reduced compared to 1D. GP is still superior, but HP drops from second place to second to last place, with only Gray giving weaker results.

To understand this result, remember that RCM, ND, AMD and also Grey do not provide load balancing. Once this drawback is removed, it becomes clear that RCM and ND provide good cache reuse. Thus, these reorderings are much stronger when using a 2D algorithm.

4.4 IN-DEPTH PERFORMANCE ANALYSIS

The previously presented results show that the obtainable speedups greatly vary across different matrices, algorithms and platforms. To better capture the dynamics of this complex interaction, we now present an in-depth analysis that aims to uncover what are the most common execution scenarios that result in performance improvement (or degradation), and what are the major causes for that behaviour.

We identified 6 classes of common execution scenarios, and Figure 4 analyzes SpMV performance with respect to representative matrices from each class, 3 platforms from different vendors (AMD, Intel and ARM), both 1D and 2D SpMV algorithms and all 6 reordering schemes. *Classes 1, 2 and 3* depict different scenarios where reordering improves performance, in *Class 4* no significant performance difference is observed, while reordering in *Classes 5 and 6* degrades performance when compared to the original matrix.

Class 1 covers cases where the original and reordered matrices are load balanced, and significant speedups are observed for both 1D and 2D algorithms. This is seen in Figure 4, where the original and almost all reordered `333SP` matrices from *Class 1* have an imbalance factor (IF) of 1.0 for the 1D runs. (Due to its nature, load balancing with the 2D algorithm is guaranteed, i.e., its imbalance factor is always 1.0 and it does not depend on matrix features.), suggesting that *Class 1* reorderings do not have significant impact on load balancing. As such, the speedups obtained for both 1D and 2D runs mainly demonstrate the ability of the reordering schemes to provide better

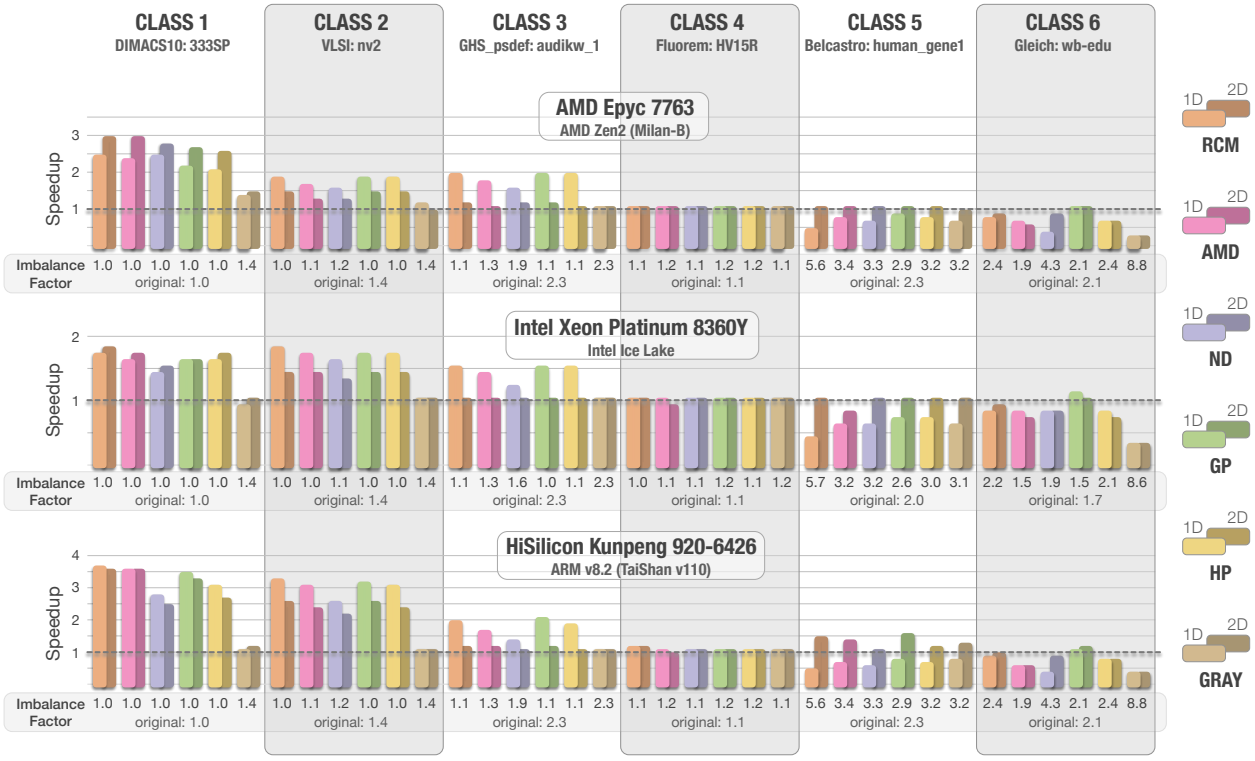

Figure 4 *Performance analysis of matrix classes for different reordering schemes, SpMV algorithms and platforms.*

data locality and cache reuse. A notable exception in Figure 4 is Gray reordering, which induces some load imbalance (IF=1.4), resulting in marginal improvements (AMD and Intel platforms) or even performance degradation (ARM).

A similar scenario can be observed for *Class 2* (see `nv2` in Figure 4), where reordering additionally provides better load balancing (notice the reduction in IF from the original 1.4). Since, in *Class 2*, the speedups are still observed for both 1D and 2D runs, it showcases the ability of reordering schemes to provide better data locality and load balancing. In contrast, reordering of *Class 3* matrices (see `audikw_1` in Figure 4) can mainly improve the load balancing, since speedups are only observed for the 1D runs (no performance changes for 2D).

In *Class 4*, original and reordered matrices deliver similar performance for both 1D and 2D runs. As shown for `HV15R` in Figure 4, reordering does not significantly impact load balancing, thus suggesting that both original and reordered matrices are equally capable of exploiting the data locality (e.g., data fits in cache either way). On the other hand, *Class 5* shows that the reordered matrices can provoke load imbalance in 1D execution, thus resulting in performance degradation, which does not occur in the inherently load-balanced 2D runs. Finally, *Class 6* depicts the case where different reordering schemes can diversely impact SpMV performance, indicating a need for new approaches to efficient matrix reordering.

As shown in Figure 4, it is worth noting that the matrices from different classes maintain very similar behaviour across all three platforms. However, the range of attainable speedups is highly affected by the platform specifics, e.g., the highest range is offered in the ARM system, followed by the AMD and Intel platforms.

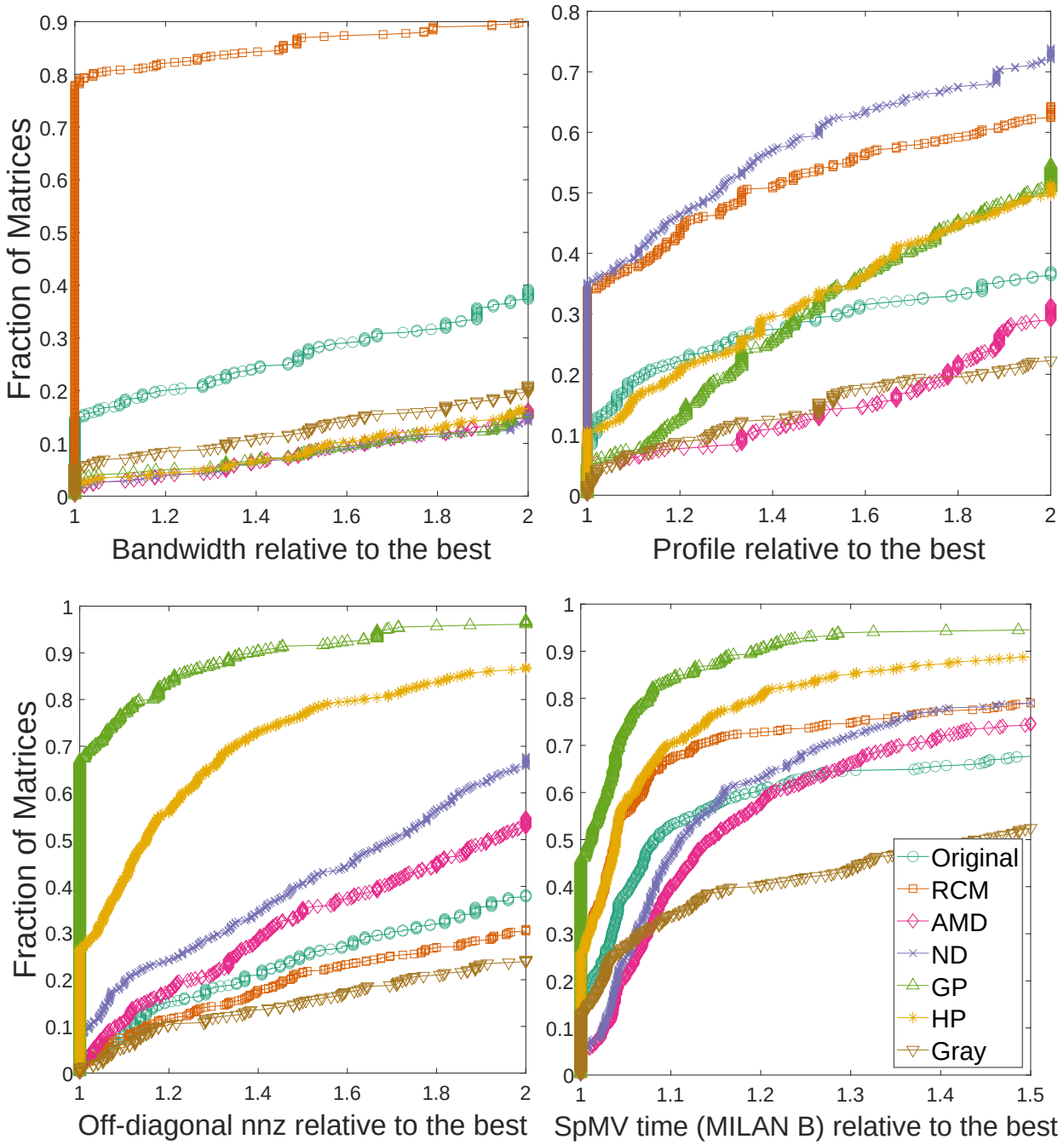

Figure 5 *Performance profiles comparing bandwidth, profile, off-diagonal nonzero count and SpMV runtime.*

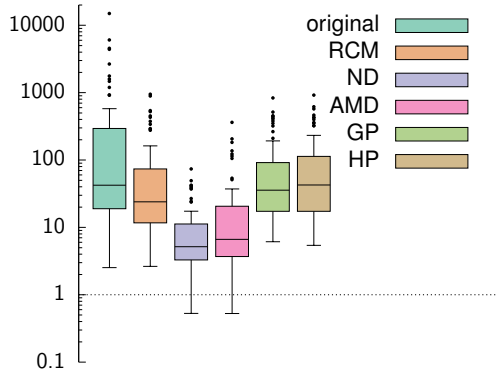

Figure 6 *Nonzero ratio in Cholesky factor* L *to nonzeros in* $A = LL^T$ *for different orderings. The middle line in each box is the median, whereas the endpoints of each box correspond to the lower and upper quartiles.*

4.5 MATRIX FEATURES AND METRIC ANALYSIS

Figure 5 depicts performance profile[33] plots to compare different methods in terms of bandwidth, profile, off-diagonal nonzero count, and SpMV runtime relative to the best performing one for each instance. A point (x, y) on a profile means that the respective model is within x factor of the best result for a fraction y of the instances. For example, the point (1.10, 0.80) on the curve of GP means that for 80 % of the matrices, SpMV with GP is at most 10 % slower than the best SpMV time obtained by any method. Therefore, a curve closer to the top left corner is interpreted as better.

In Figure 5, in terms of reducing the bandwidth, it is seen that RCM is the clear winner while being the best method for almost 80 % of the matrices. The success of RCM in reducing bandwidth is expected since it is primarily exploited for that purpose. What might be more surprising is that all the other methods are worse than the original ordering in that regard. As for reducing the matrix profile, we see ND as the best method closely followed by RCM. Regarding the nonzero count in off-diagonal blocks, GP is the winner with the best performance for nearly 65 % of the instances. This is expected since GP with edge-cut objective is directly aims to minimize the off-diagonal nonzero count. The second method in that regard is HP, which can also be expected since cut-net objective aims to minimize off-diagonal nonzero segments, thus having an indirect yet strong relation with this metric.

Finally, considering the performance profile for SpMV runtimes in Figure 5, it most closely resembles the performance profile for the off-diagonal nonzero count. This suggests that the off-diagonal nonzero count is a much more important feature than profile and bandwidth with respect to SpMV performance. Here, we again see GP and HP as our first and second most effective methods. One important leap belongs to RCM, which is seen as the third best method in reducing SpMV time, while being not that successful in reducing the off-diagonal nonzero count. This might be explained by the superior success of RCM in reducing bandwidth, which might increase cache reuse and hence serve the SpMV effectiveness indirectly. Meanwhile, we expect the success of ND for reducing profile to be more effective in its performance for reducing fill-in, as we will observe in Section 4.6.

It is clear that SpMV performance benefits most from the reordering methods that reduce the number of nonzeros in off-diagonal blocks most.

[33] Elizabeth D Dolan and Jorge J Moré. "Benchmarking optimization software with performance profiles". *Mathematical programming* 91.2 (2002), pp. 201–213.

Table 5 *Time (in seconds) to reorder a matrix on Ice Lake. For comparison, execution time of a single CSR SpMV iteration using 72 threads is also shown.*

Matrix Name	RCM	ND	AMD	Gray	METIS	PaToH	SpMV
delaunay_n24	5.3	210	18.6	2.9	10.9	125	0.010
europe_osm	15.4	437	18.9	7.5	31.5	151	0.013
Flan_1565	1.1	18.3	2.5	0.2	4.7	120	0.004
HV15R	6.5	118	8.3	0.3	31.5	429	0.011
indochina-2004	30.9	163	80.3	2.0	29.6	219	0.072
kmer_V1r	117	2 915	5 047	64.2	1 333	7 865	0.084
kron_g500-logn21	59.4	183	366	1.3	229	22 382	0.010
mycielskian19	24.4	131	80.1	3.1	739	177	0.132
nlpkkt240	14.3	869	71.7	4.9	53.8	1 040	0.035
vas_stokes_4M	4.2	146	17.0	1.2	22.6	243	0.010

4.6 FILL-IN FOR SPARSE CHOLESKY FACTORISATION

We computed the fill-in created by sparse Cholesky factorisation using the row counting algorithm of Gilbert et al.[34] for symmetric, positive definite matrices in SuiteSparse with different orderings. The Gray code ordering is not included, since it does not preserve symmetry and therefore cannot be used for this factorisation. Figure 6 compares matrix orderings with respect to the ratio of nonzeros in the Cholesky factor L to nonzeros in $A = LL^T$ for 78 of the largest matrices.

As expected, the fill-reducing orderings, AMD and ND, produce the least fill-in. While RCM, GP and HP are considerably less effective, they still typically produce better results than the original ordering. With respect to the features discussed in Section 4.5, other matrix features may be needed to explain the reduction of fill-in due to reordering.

4.7 REORDERING OVERHEAD

To compare the cost of different reorderings, Table 5 shows the reordering time on Ice Lake (see Table 2) for a few representative matrices. Roughly speaking, Gray ordering is always fastest and RCM is usually the second fastest, whereas HP and ND are typically the slowest. Although we believe the implementations of the chosen reordering algorithms to be reasonably efficient, we note that they are currently all serial and there may be room for further performance optimizations.

Depending on the matrix and ordering algorithm, the time required for reordering ranges from a few hundred to several million SpMV operations with the unreordered matrix using 72 cores on the same machine. To achieve overall savings from reordering, the number of SpMV operations performed must exceed the ratio of the reordering time to the difference between the unreordered and reordered SpMV. For example, reordering *europe_osm* with RCM takes 15.4 seconds and improves SpMV performance by 22 % on Ice Lake. Since a single SpMV iteration before reordering takes 0.013 seconds, then approximately $15.4/(0.013 \times (1 - 1/1.22)) \approx 6\,569$ SpMV iterations are needed to save time overall. The reordering cost is high, but in the context of scientific computing it can often be amortised over thousands or millions of SpMV iterations with the same matrix in the course of a simulation.

[34] John R. Gilbert, Esmond G. Ng, and Barry W. Peyton. "An Efficient Algorithm to Compute Row and Column Counts for Sparse Cholesky Factorization". *SIAM Journal on Matrix Analysis and Applications* 15.4 (1994), pp. 1075–1091. DOI: 10.1137/S0895479892236921.

5 CONCLUSIONS

We have presented our reordering library that includes the six most prominent reordering algorithms. Through extensive experiments, we show the impact of these reordering and identify graph partitioning as the most effective one for SpMV reorderings. This library constitutes an important building block of the SparCity sparse matrix toolset.

REFERENCES

Amestoy, Patrick R., Timothy A. Davis, and Iain S. Duff. "Algorithm 837: AMD, an Approximate Minimum Degree Ordering Algorithm". *ACM Trans. Math. Softw.* 30.3 (2004), pp. 381–388. ISSN: 0098-3500. DOI: 10.1145/1024074.1024081.

Amestoy, Patrick R. et al. "Analysis and Comparison of Two General Sparse Solvers for Distributed Memory Computers". *ACM Trans. Math. Softw.* 27.4 (2001), pp. 388–421. ISSN: 0098-3500. DOI: 10.1145/504210.504212.

Catalyurek, U.V. and C. Aykanat. "Hypergraph-partitioning-based decomposition for parallel sparse-matrix vector multiplication". *IEEE Transactions on Parallel and Distributed Systems* 10.7 (1999), pp. 673–693. DOI: 10.1109/71.780863.

Cuthill, E. and J. McKee. "Reducing the Bandwidth of Sparse Symmetric Matrices". *Proceedings of the 1969 24th National Conference*. Association for Computing Machinery, 1969, pp. 157–172. DOI: 10.1145/800195.805928.

Davis, Timothy A. and Yifan Hu. "The University of Florida Sparse Matrix Collection". *ACM Trans. Math. Softw.* 38.1 (2011). ISSN: 0098-3500. DOI: 10.1145/2049662.2049663.

Dolan, Elizabeth D and Jorge J Moré. "Benchmarking optimization software with performance profiles". *Mathematical programming* 91.2 (2002), pp. 201–213.

George, Alan. "Nested Dissection of a Regular Finite Element Mesh". *SIAM Journal on Numerical Analysis* 10.2 (1973), pp. 345–363. DOI: 10.1137/0710032.

George, Alan and Joseph W. H. Liu. "An Implementation of a Pseudoperipheral Node Finder". *ACM Transactions on Mathematical Software* 5.3 (1979), pp. 284–295. DOI: 10.1145/355841.355845.

— "The evolution of the minimum degree ordering algorithm". *SIAM Review* 31.1 (1989), pp. 1–19.

George, Alan and David R. McIntyre. "On the Application of the Minimum Degree Algorithm to Finite Element Systems". *SIAM Journal on Numerical Analysis* 15.1 (1978), pp. 90–112. ISSN: 00361429. URL: http://www.jstor.org/stable/2156565.

Gibbs, Norman E., William G. Poole, and Paul K. Stockmeyer. "An Algorithm for Reducing the Bandwidth and Profile of a Sparse Matrix". *SIAM Journal on Numerical Analysis* 13.2 (1976), pp. 236–250. ISSN: 00361429.

Gilbert, J. R. and R. E. Tarjan. "The Analysis of a Nested Dissection Algorithm". *Numer. Math.* 50.4 (1987), pp. 377–404. ISSN: 0029-599X. DOI: 10.1007/BF01396660.

Gilbert, John R., Esmond G. Ng, and Barry W. Peyton. "An Efficient Algorithm to Compute Row and Column Counts for Sparse Cholesky Factorization". *SIAM Journal on Matrix Analysis and Applications* 15.4 (1994), pp. 1075–1091. DOI: 10.1137/S0895479892236921.

Haque, Sardar Anisul and Shahadat Hossain. "A Note on the Performance of Sparse Matrix-vector Multiplication with Column Reordering". *2009 International Conference on Computing, Engineering and Information*. 2009, pp. 23–26. DOI: 10.1109/ICC.2009.40.

Heras, D.B. et al. "Modeling and improving locality for the sparse-matrix–vector product on cache memories". *Future Generation Computer Systems* 18.1 (2001), pp. 55–67. ISSN: 0167-739X. DOI: 10.1016/S0167-739X(00)00075-3.

Karypis, George and Vipin Kumar. "A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs". *SIAM Journal on Scientific Computing* 20.1 (1998), pp. 359–392. DOI: 10.1137/S1064827595287997.

Liu, Wai-Hung and Andrew H Sherman. "Comparative analysis of the Cuthill-McKee and the reverse Cuthill-McKee ordering algorithms for sparse matrices". *SIAM Journal on Numerical Analysis* 13.2 (1976), pp. 198–213.

Merrill, Duane and Michael Garland. "Merge-Based Sparse Matrix-Vector Multiplication (SpMV) Using the CSR Storage Format". *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*. Association for Computing Machinery, 2016. DOI: 10.1145/2851141.2851190.

Oliker, Leonid et al. "Effects of Ordering Strategies and Programming Paradigms on Sparse Matrix Computations". *SIAM Review* 44.3 (2002), pp. 373–393. DOI: 10.1137/S00361445003820.

Pinar, Ali and Michael T. Heath. "Improving Performance of Sparse Matrix-Vector Multiplication". *Proceedings of the 1999 ACM/IEEE Conference on Supercomputing*. Portland, Oregon, USA: Association for Computing Machinery, 1999. DOI: 10.1145/331532.331562.

Temam, O. and W. Jalby. "Characterizing the behavior of sparse algorithms on caches". *Supercomputing '92: Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*. 1992, pp. 578–587. DOI: 10.1109/SUPERC.1992.236646.

Yzelman, A. N. and Rob H. Bisseling. "Cache-Oblivious Sparse Matrix–Vector Multiplication by Using Sparse Matrix Partitioning Methods". *SIAM Journal on Scientific Computing* 31.4 (2009), pp. 3128–3154. DOI: 10.1137/080733243.

Zhao, Haoran et al. "Exploring Better Speculation and Data Locality in Sparse Matrix-Vector Multiplication on Intel Xeon". *2020 IEEE 38th International Conference on Computer Design (ICCD)*. 2020, pp. 601–609. DOI: 10.1109/ICCD50377.2020.00105.

6 HISTORY OF CHANGES

Version	Author(s)	Date	Comment
0.5	Johannes Langguth	28.04.2023	Initial draft
0.5.1	Didem Unat	02.05.2023	Final version for submission

Table 6 *Document History of Changes*