



UvA-DARE (Digital Academic Repository)

Using Evolutionary Algorithms to Find Cache-Friendly Generalized Morton Layouts for Arrays

Swatman, S.N.; Varbanescu, A.-L.; Pimentel, A.D.; Salzburger, A.; Krasznahorkay, A.

DOI

[10.48550/arXiv.2309.07002](https://doi.org/10.48550/arXiv.2309.07002)

[10.1145/3629526.3645034](https://doi.org/10.1145/3629526.3645034)

Publication date

2024

Document Version

Final published version

Published in

ICPE '24

License

CC BY

[Link to publication](#)

Citation for published version (APA):

Swatman, S. N., Varbanescu, A.-L., Pimentel, A. D., Salzburger, A., & Krasznahorkay, A. (2024). Using Evolutionary Algorithms to Find Cache-Friendly Generalized Morton Layouts for Arrays. In *ICPE '24: Proceedings of the 15th ACM/SPEC International Conference on Performance Engineering : May 7-11, 2024, London, United Kingdom* (pp. 83–94). Association for Computing Machinery. <https://doi.org/10.48550/arXiv.2309.07002>, <https://doi.org/10.1145/3629526.3645034>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)



Using Evolutionary Algorithms to Find Cache-Friendly Generalized Morton Layouts for Arrays

Stephen Nicholas Swatman*
University of Amsterdam
Amsterdam, The Netherlands
s.n.swatman@uva.nl

Ana-Lucia Varbanescu
University of Twente
Enschede, The Netherlands
a.l.varbanescu@utwente.nl

Andy D. Pimentel
University of Amsterdam
Amsterdam, The Netherlands
a.d.pimentel@uva.nl

Andreas Salzburger
CERN
Geneva, Switzerland
andreas.salzburger@cern.ch

Attila Krasznahorkay
CERN
Geneva, Switzerland
attila.krasznahorkay@cern.ch

ABSTRACT

The layout of multi-dimensional data can have a significant impact on the efficacy of hardware caches and, by extension, the performance of applications. Common multi-dimensional layouts include the canonical row-major and column-major layouts as well as the Morton curve layout. In this paper, we describe how the Morton layout can be generalized to a very large family of multi-dimensional data layouts with widely varying performance characteristics. We posit that this design space can be efficiently explored using a combinatorial evolutionary methodology based on genetic algorithms. To this end, we propose a chromosomal representation for such layouts as well as a methodology for estimating the fitness of array layouts using cache simulation. We show that our fitness function correlates to kernel running time in real hardware, and that our evolutionary strategy allows us to find candidates with favorable simulated cache properties in four out of the eight real-world applications under consideration in a small number of generations. Finally, we demonstrate that the array layouts found using our evolutionary method perform well not only in simulated environments but that they can effect significant performance gains—up to a factor ten in extreme cases—in real hardware.

CCS CONCEPTS

• **Software and its engineering** → **Software performance**; • **Mathematics of computing** → **Combinatorial optimization**; • **Information systems** → **Data layout**.

KEYWORDS

Morton curve, Z-order curve, space-filling curves, array layout, multi-dimensional data, evolutionary algorithms, caching, locality

*Also with CERN.



This work is licensed under a Creative Commons Attribution International 4.0 License.

ICPE '24, May 7–11, 2024, London, United Kingdom
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0444-4/24/05
<https://doi.org/10.1145/3629526.3645034>

ACM Reference Format:

Stephen Nicholas Swatman, Ana-Lucia Varbanescu, Andy D. Pimentel, Andreas Salzburger, and Attila Krasznahorkay. 2024. Using Evolutionary Algorithms to Find Cache-Friendly Generalized Morton Layouts for Arrays. In *Proceedings of the 15th ACM/SPEC International Conference on Performance Engineering (ICPE '24)*, May 7–11, 2024, London, United Kingdom. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3629526.3645034>

1 INTRODUCTION

Structured multi-dimensional data are ubiquitous in high-performance computing [9]: three-dimensional fluid simulations, dense linear algebra operations, and stencil kernels are just a few examples of applications which rely fundamentally on multi-dimensional arrays. In spite of the importance of such applications, however, most modern computer systems have one-dimensional memories: from the perspective of the programmer, memory is nothing more than a very large one-dimensional array of bytes. This discrepancy between application requirements and hardware design requires programmers to carefully consider *array layouts*: injective functions which translate multi-dimensional indices into one-dimensional memory addresses.

Although array layouts do not impact the functional properties of programs, choosing a suitable layout can significantly impact application performance in modern processors with complex cache hierarchies [48]. Exploiting these caches is of critical importance to achieving high performance in all but purely compute-bound applications, but doing so requires locality of access—both temporal and spatial—in memory. Kernels often exhibit locality in multiple dimensions, and a well-chosen array layout maximizes the degree to which this application-level locality is translated to the address-level locality that caches are designed to exploit; as a result, that layout increases the efficacy of hardware caching and—by extension—the performance of an application.

Data in two-dimensions is commonly laid out in *row-major* order (shown in Figure 2a for an 8×8 array) or *column-major* order (Figure 2t) which provide good locality of access in a single dimension, but poor locality in all others. Thankfully, the design space for data orderings—in two dimensions or more—extends far beyond these canonical layouts: the *Morton* layout (Figure 2f), for example, is a layout based on a space-filling curve which provides balanced locality between multiple dimensions [46, 62]. Our work explores a family of data layouts which generalize the Morton order, and allow

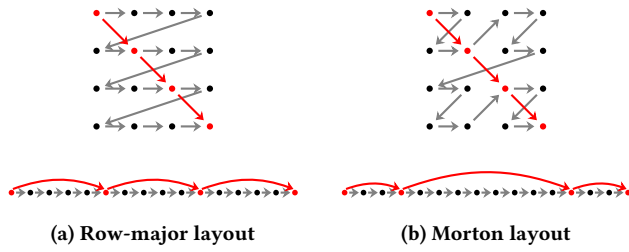


Figure 1: Two-dimensional arrays laid out in memory along the gray arrows. An application accesses the array diagonally along the red arrows. Application locality is shown above, memory locality is shown below.

us to carefully tune the cache behavior in multiple dimensions to match a given application.

The design space of the aforementioned family of data layouts is dauntingly large; indeed, the number of possible layout for arrays at scales applicable to real-world problems is so large that it renders exhaustive search infeasible. In order to find suitable array layouts in tractable amounts of time, we propose to employ genetic algorithms—heuristics known to be able to efficiently find high-quality solutions in large search spaces [35]. To this end, we design a chromosomal representation of Morton-like array layouts, as well as a fitness function that uses cache simulation to estimate the performance of individual array layouts. Finally, we evaluate our evolutionary strategy and the array layouts it discovers.

In short, our paper makes the following contributions:

- We characterize the design space given by a generalization of the Morton array layout, and we show that the size of this design space renders exhaustive search infeasible (Section 3);
- We propose an evolutionary methodology based on genetic algorithms for exploring the aforementioned design space based on the simulated cache-friendliness of layouts (Section 4);
- We design and execute a series of experiments to assess the accuracy of our fitness function, the efficacy of our evolutionary process, and the performance of the discovered array layouts, showing that our method can improve performance up to a factor ten (Section 5).

2 BACKGROUND AND RELATED WORK

In this section, we provide a brief overview of the basic concepts and notations which are essential to the remainder of this paper, and highlight relevant related work.

2.1 Indexing Functions and Canonical Layouts

Dense n -dimensional arrays can be imagined as structured grids in which each element is assigned to exactly one point in \mathbb{N}^n . In most modern processors, multi-dimensional arrays are a software-level abstraction over the one-dimensional memory of the machine; in order to actually access multi-dimensional data, we need to define a function that converts indices in n dimensions to memory

addresses¹. We refer to the class of such functions as *indexing functions*, and they are isomorphic to *array layouts*. In short, an n -dimensional indexing functions is an injective (often bijective) function of the following type, where N_i represents the size of the array in the i th dimension, \times is the generalised Cartesian product, and $\llbracket a, b \rrbracket$ is the integer interval from a to b :

$$f : \times_{i=0}^{n-1} \llbracket 0, N_i - 1 \rrbracket \rightarrow \left[\left[0, \left(\prod_{i=0}^{n-1} N_i \right) - 1 \right] \right] \quad (1)$$

In a multi-dimensional grid, we denote the elements along a given axis—that is to say, the sequence of elements for which all indices except one are fixed—as *fibers* [41]. In a two-dimensional case, fibers along the x -axis are known as *rows*, and fibers along the y -axis as *columns*. In order to facilitate the description of arrays in three or more dimensions, we use the term *mode- m fibers* to describe fibers along the m th dimension, such that mode-0 fibers are synonymous with rows, mode-1 fibers refer to columns, and so forth.

The most common group of multi-dimensional indexing functions are the *canonical* layouts, sometimes known as the *lexicographic* layouts or, in the two-dimensional case, the *row- and column-major* layouts. In a canonical layout, one-dimensional array indices are calculated according to Equation 2, in which x_0, \dots, x_{n-1} are components of the n -dimensional index, and N_0, \dots, N_{n-1} represent the size of the array in each dimension:

$$f(x_0, \dots, x_{n-1}; N_0, \dots, N_{n-1}) = \sum_{i=0}^{n-1} \left(\prod_{j=0}^{i-1} N_j \right) x_i \quad (2)$$

An important corollary of Equation 2 is that the mode-0 fibers are contiguous in memory i.e., Equation 3 holds:

$$f(x_0 + 1, x_1, \dots, x_{n-1}) = f(x_0, x_1, \dots, x_{n-1}) + 1 \quad (3)$$

It is worth noting that the calculation of addresses in column-major layout—in which the mode-1 fibers are contiguous—is also given by Equation 2, with the order of the indices and sizes swapped. The canonical array layouts achieve perfect spatial locality in one dimension: if a kernel accesses memory along mode- m fibers, then a canonical layout where the m th dimension is major will provide the optimal translation between locality in the multi-dimensional space to locality in memory. Many real world applications, however, exhibit locality in multiple dimensions; a kernel might, for example, iterate diagonally over an array; an example of this—and the resulting locality in memory—is given in Figure 1a.

The performance of canonical storage layouts has been studied extensively. Park et al. discuss methods for compensating for the weaknesses of canonical layouts using tiling and recursive layouts [48]. Similarly, Kowarschik and Weiß propose a variety of strategies that mitigate cache misses in canonical storage layouts for numerical applications [42]. Weinberg et al. propose a metric for the locality of array layouts [66]. Jang et al. analyze the performance of access patterns in multi-dimensional data in graphics processing

¹In reality, address calculations must also consider array offsets (the address of the first element) and scales (the size of each element). We skip over these complications as they are handled transparently by address generation units in modern hardware, and they affect all array layouts in the same manner.

units (GPUs) [40]. Che et al. propose a method for automatically optimizing storage layouts [16].

2.2 Morton Layouts

The Morton order is a notable example of a non-canonical array layout that provides balanced locality in multiple dimensions. It is conceptually simple to understand, efficient to implement in commodity hardware (as we will show in Section 3.3), and it has been shown to positively affect the efficacy of hardware caches: Al-Kharusi and Walker show the efficacy of the Morton layout in molecular dynamics applications [5], Perdacher et al. describe its benefits in matrix decomposition [51], and Thiyagalingam et al. provide an in-depth performance analysis of this array layout in a range of kernels [62]. Chatterjee et al. show the applicability of Morton layouts—as well as other non-canonical layouts—in matrix multiplication [15], and this work is expanded upon in [14]. Applications of the Morton order in more than two dimensions have been studied by Pawłowski et al. [50]. Mellor-Crummey et al. show the applicability of array layouts based on space-filling curves—like the Morton layout—for irregular applications [44]. The practical applicability of the Morton layout is further evidenced by the OPIE compiler, which employs Morton array layouts natively [24].

The performance benefits of the Morton layout stem from its spatial structure: an example—which justifies why this layout is sometimes known as the *Z-order layout*—is given in Figure 1b; note the difference in locality in the address space compared to the canonical layout (Figure 1a). The Morton order layout has also been applied to data movement in parallel systems by Walker and Skjellum [65], and Deford and Kalyanaraman have applied the layout to workload distribution in parallel processes [20]. Bader explores a variety of applications of space-filling curves in scientific programs [10]. Armbrust et al. explore the application of Morton curves for the storage of databases, reducing the total amount of data read from persistent storage [8]; although the aforementioned paper considers a much higher level of abstraction than the methods in this paper—which operate at the level of hardware caches—we believe that the methods presented in this paper may generalize to a broader range of applications, including databases.

In the Morton order, multi-dimensional indices can be converted to one-dimensional addresses in a variety of ways. The Moser–de Bruijn sequence [36] is commonly used as it allows efficient conversions in two dimensions, but this method requires us to store the Moser–de Bruijn sequence in memory, and accessing this sequence causes additional overhead. Therefore, we prefer a different method based on the interleaving of the (unsigned) binary representation of multi-dimensional indices. As an example, the two-dimensional index (3, 5) can be bijectively mapped into one-dimensional memory by finding the binary expansions of the indices i.e., (011₂, 101₂), and interleaving the bits yielding 100111₂ = 39₁₀. This is equivalent to first dilating and shifting the binary expansions of the numbers, and then taking their bitwise disjunction (OR): the first index is dilated yielding 000101₂ while the second index is dilated and shifted left yielding 100010₂. Taking the bitwise disjunction of these numbers yields the same address as using the interleaving strategy. The computation of Morton indices through bit manipulation can be extended to an arbitrary number of dimensions; the

three-dimensional index (3, 5, 4) expands to (011₂, 101₂, 100₂), and the resulting memory address is 110001011₂ = 395₁₀. Note that the relative significance of bits in each of the input indices is preserved in the output address. Gottschling et al. present the idea that the Morton layout can be generalized by allowing arbitrary bit-interleaving orders [27, 28], which is foundational to our work. This idea is further expanded on by Walker [64].

2.3 Genetic Algorithms

Genetic algorithms are a class of heuristics introduced by Holland which are designed to solve optimization and search problems by emulating the process of evolution as it happens in the natural world [34]. In genetic algorithms, *generations of individuals* i.e., sets of candidate solutions, iteratively explore a design space through genetic operators. In particular, *crossover* operators model the combination of the genetic material of two or more individuals [49], and *mutation* operators model random changes to the gene pool [57]. In genetic algorithms, individuals are removed from the population based on their *fitness* i.e., the quality of the solution they represent to the problem posed [55]. Genetic algorithms have seen successful application in an extremely broad range of fields, ranging from drug discovery [61] to music composition [25]. Genetic algorithms have also proven useful for design space exploration in computer systems; Pimentel shows that they can be applied in the design of embedded systems [52]. Sapra and Pimentel show that a broader class of evolutionary approaches can be used in the design of neural networks [54]. The optimization problem we consider in this paper is combinatorial in nature, and the application of genetic algorithms to such problems has also been extensively studied and proven across a variety of domains [7, 26, 31, 47].

3 GENERALIZED MORTON LAYOUTS

The Morton layout functions by interleaving the bits of the input indices in a fixed pattern: bits are drawn from each of the inputs in a round-robin manner. In this section, we generalize this idea, allowing bits to be interleaved in arbitrary order. This gives rise to more specialized layouts with different structure and, as a result, different extra-functional properties [27, 28, 64]. Figure 2 shows all 20 layouts that are given by different bit interleaving orders for an 8 × 8 array. As with the standard Morton layout, the generalized Morton layout can be applied to any number of dimensions. As an example, the following three-dimensional layout selects two bits from the second index, one bit from the third index, then two bits from the first index, etc.:

$$f(011_2, 101_2, 100_2) = \begin{array}{r} 000011000_2 \\ \vee 000100001_2 \\ \hline 10000000_2 \\ \hline 100111001_2 \end{array} = 313_{10} \quad (4)$$

Our goal is to find Morton-like layouts i.e., bit-interleaving patterns, that improve application performance through an increase in cache efficacy. In this section, we will show that the design space for such layouts is very large, motivating the use of genetic algorithms. This necessitates a chromosomal representation of layouts, which we also present in this section. In addition, we describe how the canonical layouts can be described using the same representation,

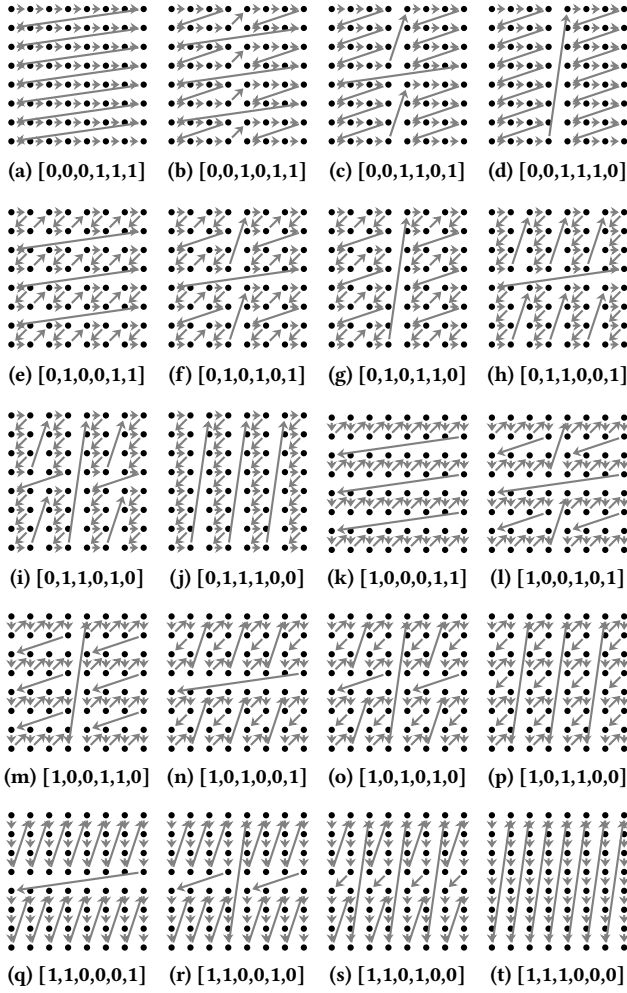


Figure 2: All 20 layouts for 8×8 arrays generated by the family of indexing schemes described in Section 3. Note that Figure 2a corresponds to a row-major layout, while Figure 2t corresponds to a column-major layout.

and we delve into practical considerations such as the computational cost of computing indices and support for same-instruction multiple-data (SIMD) processing.

3.1 Enumerating Layouts

We can characterize Morton-like layouts by the bit scattering pattern applied to each of the inputs (e.g., for Equation 4, the first index is scattered to the fourth, fifth, and eighth bits). However, such a characterization is unsound in the sense that it allows us to describe invalid layouts: if two bits from any of the input indices are mapped onto the same bit in the output, the bitwise disjunction becomes an information-destroying operation and the layout becomes non-injective—that is, it would cause multiple multi-dimensional indices to map onto the same location in memory, making the layout unusable.

We can instead characterize layouts in a manner that is both complete and sound by enumerating the *source* of each bit in the output index. In the remainder of this work we shall denote array layouts using sequences of the form $[i_0, \dots, i_{n-1}]$, indicating the source indices in order of increasing bit significance: the least significant bit in the output index is drawn from the i_0 th input index, the second-least significant bit is drawn from the i_1 th input, and the most significant bit is drawn from the i_{n-1} th input. Note that each input bit must be used once and only once: whenever a bit is to be drawn from a given input index, we implicitly use the least significant bit for that input which has not yet been consumed. For the layout shown in Equation 4, the two least significant bits are drawn from the second input, the third-least significant bit is drawn from the third input, and so forth: the resulting array layout is denoted using the sequence $[1, 1, 2, 0, 0, 1, 2, 0, 2]$.

The aforementioned characterization of multi-dimensional layouts gives rise to families of layouts. The family of layouts over n inputs, where each input has b_0, \dots, b_{n-1} bits, is isomorphic to the set of permutations of the multiset $S = \{0 : b_0, \dots, n-1 : b_{n-1}\}$. We denote this set of permutations as $\mathfrak{S}(S)$. For convenience, we obviate the intermediate multiset such that $\mathfrak{S}'(b_0, \dots, b_{n-1}) = \mathfrak{S}(\{0 : b_0, \dots, n-1 : b_{n-1}\})$. We can then determine the total number of possible layouts as the number of multiset permutations of \mathfrak{S}' [13, p. 42]:

$$|\mathfrak{S}'(b_0, \dots, b_{n-1})| = \binom{\sum_{i=0}^{n-1} b_i}{b_0, \dots, b_{n-1}} = \frac{(\sum_{i=0}^{n-1} b_i)!}{\prod_{i=0}^{n-1} (b_i!)} \quad (5)$$

3.2 Including Canonical Layouts

It is worth noting that canonical layouts over arrays for which the size in each dimension is a power of two are, in fact, members of the family of Morton-like layouts. In order to sketch an informal argument for this, we recall that the indexing function for an n -dimensional canonical layout given array sizes N_0, \dots, N_{n-1} is defined as in Equation 2. If we assume that all sizes are powers of two, then the product of these sizes is guaranteed to be itself a power of two. Because multiplication by powers of two can be interpreted as a left-ward shift, the canonical layouts shift each input index x_0, \dots, x_n to a specific location in the binary expansion of the output index. Furthermore, because we assume $\forall i : x_i < N_i$, each bit in the output is determined by exactly one of the input indices; this allows us to interpret the summation as a series of bit-wise disjunctions, exactly like the definition of our Morton-like layouts. In general, a mode-0-major canonical layout of a $2^{b_0} \times \dots \times 2^{b_{n-1}}$ array can be characterized—in the the scheme defined in Section 3.1—by contiguous subsequences of bits, each drawn from the same index i.e., a sequence of the following form:

$$\underbrace{[0, \dots, 0]}_{b_0 \text{ times}}, \underbrace{[1, \dots, 1]}_{b_1 \text{ times}}, \dots, \underbrace{[n-1, \dots, n-1]}_{b_{n-1} \text{ times}} \quad (6)$$

Canonical layouts with different major axes can be constructed by changing the order of the contiguous subsequences. The fact that the canonical layouts are members of the Morton-like family of array layouts allows us to evaluate the performance of these layouts in the exact same framework as the rest of the Morton-like layouts, and we will exploit this in Section 5.

3.3 Hardware-Accelerated Indexing

It is tempting to extend the aforementioned ideas to even more exotic indexing functions, like the Hilbert array layout [6, 32, 67]. The computational cost of many such functions renders them impractical, however: if the cost of computing memory addresses is too large, any performance gained by improving the cache-friendliness of a program will be negated. The Morton-like layouts we consider in this work allow efficient index calculations on modern commodity hardware, which we demonstrate in this section.

Under canonical array layouts, indices are calculated either iteratively through repeated addition and multiplication, or in parallel through parallel multiplication followed by reduction through addition. In n -dimensional cases both approaches require $n - 1$ additions and $n - 1$ multiplications, operations which can be efficiently performed on virtually all processors. Specifically, the Intel Haswell and AMD Zen 3 microarchitectures—on which we focus in this work—can perform 64-bit register addition (`ADD r64 r64`) with a latency 1 cycle and a reciprocal throughput of 0.25 cycles, while they can execute multiplication (`IMUL r64 r64`) with a latency of 3 cycles and a reciprocal throughput of 1 cycle [1].

Our bit-interleaving array layouts rely, in n -dimensional cases, on $n - 1$ bitwise disjunctions and n bit-scatter operations. Such disjunctions (`OR r64 r64`) can be performed with a latency of 1 cycle and a reciprocal throughput of 0.25 cycles—the same as the `ADD` instruction—on both of the aforementioned microarchitectures. We perform the bit-scattering operation using the *parallel bit deposition* (`PDEP r64 r64 r64`) instruction, which is included in the BMI2 extension to the x86-64 instruction set [4]. The Intel Haswell and AMD Zen 3 microarchitectures both perform bit deposition with a latency of 3 cycles and a reciprocal throughput of 1 cycle, identical to the `IMUL` instruction. It follows that Morton-like indexing requires—in theory—only a single additional instruction over canonical index calculation.

The hardware extension required to perform bit deposition is widely supported: BMI2 has been included in Intel processors starting with the Haswell microarchitecture (2013) [29], and in AMD processors starting with the Excavator microarchitecture (2015), albeit in a limited fashion; AMD processors gained full hardware support for these instructions starting with the Zen 3 microarchitecture (2020) [22]².

In order to further evaluate the competitiveness of Morton-like layouts compared to canonical layouts, we analyze implementations of both indexing schemes over a range of dimensionalities as compiled by GCC 12.3 and CLANG 15.0 using OSACA 0.5.2 [43]. All code was compiled using the `-O2` optimization flag. The results of this analysis are shown in Figure 3. Over the range of dimensionalities considered, the canonical layouts are consistently faster i.e., require fewer cycles to compute, than the Morton-like layouts. However, the difference in performance—approximately one cycle—is relatively small and overshadowed by the number of cycles saved due to a reduction in cache misses. Furthermore, we focus primarily on memory-bound applications, in which a small increase in index calculation time is unlikely to affect performance. We conclude,

²Pre-Zen 3 processors supported parts of the BMI2 instruction set—the `PEXT` and `PDEP` instructions in particular—through emulation in microcode rather than in hardware, making them very slow.

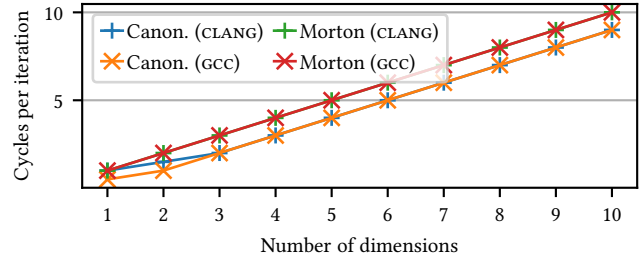


Figure 3: Throughput of a kernel calculating array indices using canonical layouts as well as Morton-like layouts on the Intel Haswell microarchitecture as given by OSACA.

therefore, that Morton-like layouts are competitive with canonical layouts strictly in terms of address computation costs.

3.4 Support for SIMD

An important consideration in the design of array layouts is the ability to vectorize kernels through single-instruction multiple-data (SIMD) operations. Canonical layouts guarantee the contiguity of fibers in the array, which facilitates the (automated) vectorization (e.g., the application of SIMD) of many operations, and this benefit is lost when applying the array layouts discussed in this paper. However, we posit that there remains ample opportunity to accelerate computation on Morton-like arrays using SIMD, and we argue this by distinguishing two classes of computation patterns.

The first class consists of *unstructured* patterns in which data is operated on element-wise without spatial context i.e., without consideration of nearby elements; a prominent example of such an operation is matrix *addition*. In such applications, SIMD can be trivially applied to the underlying one-dimensional memory, regardless of the layout of the data: since elements can be added point-wise in any order, doing so in the order in which the data is laid out in memory is both feasible and enables SIMD.

The second class of problems consists of *structured* patterns in which operations must be performed in a specific order. A prime example of such an operation is matrix *multiplication* where the inner product of fibers must be computed. In such cases, it is imperative that fibers can be accessed in contiguous blocks. The size of these blocks depends on the vectorization technology used as well as the size of the data type: in the x86 instruction set, SSE vectorisation requires two consecutive double-precision numbers or four consecutive single-precision numbers [38]; the much wider ARM SVE instruction set extension [58] may require up to thirty-two consecutive double-precision numbers or sixty-four single-precision numbers.

In order to facilitate vectorization for structured patterns of computation, we can impose certain constraints on the array layouts we consider. Indeed, if the n least-significant bits of an interleaving pattern are all drawn from the m th input index, then the layout guarantees that the mode- m fibers in the array are contiguous in blocks of 2^n elements. This requirement can be incorporated into the selection of array layouts; for example, we can enable efficient AVX2 vectorisation (with a vector width of 256 bits) using single-precision (32-bit) floating point numbers by ensuring that the three

least significant bits in an array layout are drawn from the same source. In other words, we can easily constrain our search space to include only array layouts with properties that favor vectorization, and we believe that doing so will enable SIMD-accelerated computation arrays laid out in Morton-like orders.

4 EXPLORATION THROUGH EVOLUTION

The canonical set of indexing bijections for laying out multi-dimensional memory is small: for two-dimensional data, there are two possible layouts, and the performance of these layouts can be evaluated using exhaustive benchmarks [23, 59, 62]. Exhaustively exploring the family of indexing function outlined in Section 3, however, is impractical owing to the sheer number of permissible permutations. Importantly, the number of canonical layouts increases only with the number of *dimensions*, while the number of Morton-like layouts increases with both the number of dimensions and the *size* of the array in each of those dimensions. By Equation 5, a small 4×4 array (indexed by two bits in each dimension) can be laid out in $(2+2)!/2!2! = 6$ ways. A larger array of size 4096×4096 (twelve bits in each dimension) can be laid out in $(12+12)!/12!12! = 2\,704\,156$ ways. A three-dimensional array of size $256 \times 256 \times 256$ has the same number of elements as the aforementioned 4096×4096 array, but permits $(8+8+8)!/8!8!8! = 9\,465\,511\,770$ permutations. As these examples indicate, the number of possible permutations quickly scales beyond what can be feasibly explored through exhaustive search; in order to tackle the explosive growth in the design space for Morton-like layouts, we propose the use of genetic algorithms (Section 2.3).

4.1 Genetic Algorithm Configuration

In this work, we employ a relatively simple (λ, μ) -ES genetic algorithm [34, 56]. The chromosomal representations of array layouts is identical to the characterization given in Section 3.1, and this gives rise to a combinatorial optimization problem. We facilitate the recombination of array layouts into novel layouts using the ordered crossover (OX) operator [18], and we employ inversion-based mutation [21]. Our approach differs from classical genetic algorithms in only one significant way: our initial population is not chosen randomly from the solution space. Instead, the initial populations for our evolutionary experiments always consist of two individuals, depicting two canonical layouts for a given array size, as described in Section 3.2. We choose to do this to ensure that our initial populations are unbiased and deterministic, allowing us to more easily assess the efficacy of our genetic strategy.

4.2 Fitness Function Design

There are two general strategies for evaluating the performance i.e., fitness, of a given array layout under a given cache hierarchy and access pattern: measurement and simulation. In order to assess fitness through *measurement*, we execute a program on actual hardware and measuring the running time of the process. Although such a fitness function is conceptually simple, it suffers from two primary flaws: (1) measurements are noisy and may suffer from run-to-run variance, which may hinder the performance of genetic algorithms [45]—in particular, our genetic algorithm is vulnerable to noise stemming from cache pollution effects; and (2) measurements

require access to the target hardware, which may be inconvenient or even impossible—for example, in hardware-software co-design scenarios, where the hardware under consideration does not (yet) exist. For these reasons, we choose not to base our fitness function on measurements.

Instead, we employ *simulation* for which we need a simulator that can accurately compare the cache performance for different access-patterns on the same cache hierarchy. For this, we selected PYCACHESIM, a component of the KERNCRAFT toolkit [30]. We use PYCACHESIM by simulating an access pattern such as matrix multiplication and registering the relevant trace of load and store operations. After all accesses have been recorded, we force a write-back of the caches and collect the number of hits and misses in each cache level. We combine the number of hits in every cache level as well as in main memory with the latency of retrieving data from each of these levels to compute the total number of cycles spent retrieving data from the cache hierarchy. Given an array layout I , an access pattern A and a simulated cache hierarchy H , we calculate the total number of cycles using the following equation, in which $L_{i\text{hit}}$, $L_{i\text{miss}}$, and $L_{i\text{lat}}$ represent the number of hits, the number of misses, and the latency of the i th cache level, and M represents main memory:

$$C(I; A, H) = M_{\text{hit}}(I; A, H)M_{\text{lat}}(H) + \sum_i L_{i\text{hit}}(I; A, H)L_{i\text{lat}}(H) \quad (7)$$

From this, we compute an approximation of the number of accesses performed per cycle, giving rise to a higher-is-better fitness function defined as follows:

$$F(I; A, H) = \frac{L1_{\text{hit}}(I; A, H) + L1_{\text{miss}}(I; A, H)}{L1_{\text{lat}}(H) \cdot C(I; A, H)} \quad (8)$$

Intuitively, the numerator in Equation 8 counts the total number of memory accesses, as all accesses either hit or miss in L1. The denominator, then, estimates the total number of cycles spent retrieving data from the various cache levels. The denominator is multiplied by a normalizing factor equal to the latency of the L1 cache; it follows from Equation 7 that the achievable performance is softly bound by the reciprocal of the L1 access latency. Indeed, this performance is achieved if and only if all accesses hit the L1 cache. Normalizing the fitness function using the L1 cache latency improves our ability to compare results between different cache hierarchies.

5 EVALUATION

We evaluate the efficacy of the methods hitherto discussed by demonstrating that (1) our fitness function is well-chosen i.e., that it correlates with performance measurements in real hardware; that (2) our evolutionary process is capable of finding novel array layouts with favorable cache properties; and that (3) the layouts which are found by our evolutionary process actually lead to relevant performance gains in real hardware. Our validation is based on eight distinct access patterns and two processors with distinct cache hierarchies.

5.1 Experimental Setup

We consider a set of eight access patterns loosely based on the selection of algorithms used by Thiyyagalingam et al. [62]. The access

```

1 template <concepts::array<2> M>
2 void mm_ijk(const M & A, const M & B, M & C) {
3     const auto m = C.get_size();
4     for (std::size_t i = 0; i < m; ++i) {
5         for (std::size_t j = 0; j < m; ++j) {
6             typename M::value_type acc = 0.;
7             for (std::size_t k = 0; k < m; ++k)
8                 acc += A.load(i, k) * B.load(k, j);
9             C.store(acc, i, j);
10        }
11    }
12 }

```

Listing 1: Example of how an access pattern (MMIJK) is described in C++. Metaprogramming allows the same source to be used for both simulation and execution on real hardware.

<pre> 1 caches: 2 L1: 3 sets: 64 4 ways: 8 5 line: 64 6 replacement: LRU 7 write_back: true 8 store_to: L2 9 load_from: L2 10 latency: 4 11 L2: 12 sets: 512 13 ways: 8 14 line: 64 15 replacement: LRU 16 write_back: true 17 store_to: L3 18 load_from: L3 19 victim_to: L3 20 latency: 12 21 L3: 22 sets: 25600 23 ways: 16 24 line: 64 25 replacement: LRU 26 write_back: true 27 latency: 36 28 memory: 29 first: L1 30 last: L3 31 latency: 200 </pre>	<pre> 1 caches: 2 L1: 3 sets: 64 4 ways: 8 5 line: 64 6 replacement: LRU 7 write_back: true 8 store_to: L2 9 load_from: L2 10 latency: 7 11 L2: 12 sets: 1024 13 ways: 8 14 line: 64 15 replacement: LRU 16 write_back: true 17 store_to: L3 18 load_from: L3 19 victim_to: L3 20 latency: 12 21 L3: 22 sets: 32768 23 ways: 16 24 line: 64 25 replacement: LRU 26 write_back: true 27 latency: 46 28 memory: 29 first: L1 30 last: L3 31 latency: 200 </pre>
---	--

(a) Intel Xeon E5-2660 v3

(b) AMD EPYC 7413

Listing 2: Two examples of cache specifications for different CPU models. Note that these configurations are approximations of the true cache hierarchies.

patterns were picked to represent common real-world applications (dense linear algebra and fluid dynamics), to represent both two-dimensional and three-dimensional applications, and to differ in critical properties such as memory size and number of loads and stores. A description of the access patterns we consider in this paper is given in Table 1.

All our access patterns are described using C++ code—see the example in Listing 1—which ensures high performance as opposed to the Python code used for our evolutionary processes; the interaction between the C++ and Python components of our project is managed using PYBIND11 [39]. We use template meta-programming to generalize our access patterns in such a way that a single definition can be used for both simulation and benchmarking without loss of performance due to run-time polymorphism; this eliminates any possible discrepancies between the code used for simulation and the code used for measurement.

We conduct our experiments on two different CPUs: the Intel Xeon E5-2660 v3 [37] based on the Haswell microarchitecture [29], and the AMD EPYC 7413 [2] based on the Zen 3 microarchitecture [22]. When we perform experiments on non-simulated Haswell processors we use the the DAS-6 cluster [11], whereas we use a machine located at CERN for experiments on Zen 3 processors. When we perform experiments based on simulation, we use the the DAS-6 cluster [11] and configure our cache simulator according to the cache configurations shown in Listing 2a for the Haswell processor, and Listing 2b for the Zen 3 processor. Note that the cache configurations are based on the accessibility of caches from a single core. This is especially relevant for the L3 cache on the Zen 3 chip, which is shared across groups of cores rather than the entire CPU: in the case of the AMD EPYC 7413, the CPU comes equipped with 128 MiB of L3 cache, but only 32 MiB is accessible from any single core [22]. We simplify the cache replacement policies of the actual hardware by assuming LRU caches (i.e., caches with a least-recently-used eviction policy); in reality, the Haswell caches employ eviction policies consistent with tree-PLRU (tree-based pseudo-LRU) for the L1 and L2 caches [1, 63], while the L3 cache is consistent with a set-dueling-controlled adaptive insertion policy [1, 53]. Cache sizes were gathered from specification documents [3, 29], while cache latencies were obtained optimistically from sources on the fastest load-to-use latencies [3, 17]. The Zen 3 L1 cache has a fastest load-to-use latency of four cycles for integers and seven cycles for floating point values [3]—we use the latter in our simulations. Finally, we assume a constant 200 cycle access latency for main memory in both systems.

5.2 Fitness Function Validation

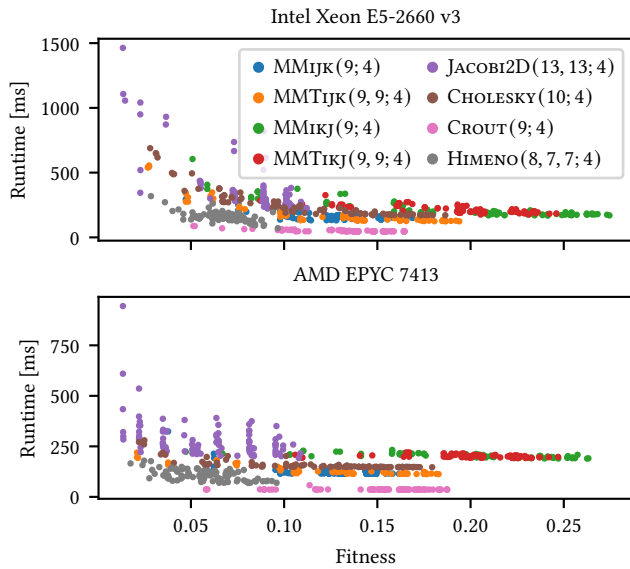
The fitness function we use in our evolutionary process (Section 4.2) is based on simulation results because simulation yields significant benefits over empirical measurements, primarily in terms of determinism and in the ability to simulate future hardware. However, this strategy is not without risk: the simulation we perform is based on a non-cycle-accurate simulator, uses simplified cache hierarchies, and ignores computation entirely. Consequently, we must evaluate the usefulness of our fitness function by establishing its correlation with execution time in real hardware.

Ideally, the running time of a kernel using a given array layout would correlate inversely linearly with our fitness function, therefore ensuring two important properties. Firstly and most importantly, it guarantees that running time decreases monotonically with the value of the fitness function, such that an array layout with a higher fitness value is guaranteed to run more quickly; this allows us to establish a ranking of layouts and enables us to reliably select the best-performing array layout. Secondly, linear correlation guarantees proportionality between fitness and running time, which facilitates the weighted selection of individuals.

To evaluate the degree to which the aforementioned criteria are met, we randomly select one hundred array layouts for each of the eight access patterns given in Table 1. We then evaluate the simulated fitness and measure the running time in real hardware of each pair of array layout and access pattern. The fitness functions of the pairs are calculated in parallel, as they are designed to be deterministic and impervious to cache pollution or resource contention.

Table 1: Overview of the access patterns used for evaluation, including the use of memory and the number of loads and stores.

Access pattern	Description	Mem. size	Loads	Stores
MMIJK($m; s$)	Multiplication of two $2^m \times 2^m$ matrices, both of s -byte real numbers.	$3 \cdot s \cdot 2^{2m}$ B	$2 \cdot 2^{3m}$	2^{2m}
MMTIJK($m, n; s$)	Multiplication of a $2^m \times 2^n$ matrix by a transposed $2^m \times 2^n$ matrix.	$s \cdot (2 \cdot 2^{m+n} + 2^{2n})$ B	$2 \cdot 2^{2m+n}$	2^{2m}
MMIKJ($m; s$)	Same as MMIJK($m; s$) with the order of the inner loops switched.	$3 \cdot s \cdot 2^{2m}$ B	$3 \cdot 2^{3m}$	2^{3m}
MMTIKJ($m, n; s$)	Same as MMTIJK($m, n; s$) with the order of the inner loops switched.	$s \cdot (2 \cdot 2^{m+n} + 2^{2n})$ B	$3 \cdot 2^{2m+n}$	2^{2m+n}
JACOBI2D($m, n; s$)	Four-point stencil kernel over a $2^m \times 2^n$ array of s -byte real numbers.	$2 \cdot s \cdot 2^{m+n}$ B	$\sim 4 \cdot 2^{m+n}$	2^{m+n}
CHOLESKY($m; s$)	Cholesky–Banachiewicz decomposition of a $2^m \times 2^m$ matrix.	$2 \cdot s \cdot 2^{2m}$ B	$2 \cdot 2^{2m}$	$\sim \frac{1}{2} \cdot 2^{2m}$
CROUT($m; s$)	Crout decomposition of a $2^m \times 2^m$ matrix of s -byte real numbers.	$2 \cdot s \cdot 2^{2m}$ B	$\frac{7}{2} \cdot 2^{2m}$	2^{2m}
HIMENO($m, n, p; s$)	Nineteen-point Himeno stencil [33] over $2^m \times 2^n \times 2^p$ arrays.	$12 \cdot s \cdot 2^{m+n+p}$ B	$24 \cdot 2^{m+n+p}$	2^{m+n+p}

**Figure 4: Scatter plot of the fitness and measured running time on an Intel Xeon E5-2660 v3 CPU and AMD EPYC 7413 for randomly chosen array layouts.**

The empirical benchmarks are performed sequentially, ensuring that the benchmark is the sole user of the processor caches. All measurements are repeated ten times, and we report the mean and standard deviation of the running time.

The results of this experiment are shown in Figure 4. The coefficient of variation of the measurements never exceeded a value of $c_v = 0.0801$. Accordingly, we have opted to omit error bars from the figure. Upon visual inspection, it is clear that the correlation between our fitness function and running time is not linear, although the two do appear correlated. We confirm our suspicions of correlation by computing Pearson’s coefficient of correlation (ρ_p) and Spearman’s coefficient of rank correlation (ρ_s); the resulting statistics are given in Table 2. We observe that our fitness function and running time correlate moderately to strongly with running time for the Intel Xeon E5-2660 v3 processor, although the correlation is weaker for the AMD EPYC 7413 processor. Although it is clear that there is space for the fitness function to be improved, we believe that it correlates sufficiently with running time to enable its use in genetic algorithms.

Table 2: Pearson’s coefficient of correlation (ρ_p) and Spearman’s coefficient of rank correlation (ρ_s) between our simulation-based fitness function and true running time.

Access pattern	Intel E5-2660 v3		AMD EPYC 7413	
	ρ_p	ρ_s	ρ_p	ρ_s
MMIJK(9; 4)	-0.672	-0.480	-0.648	-0.489
MMTIJK(9, 9; 4)	-0.810	-0.896	-0.863	-0.823
MMIKJ(9; 4)	-0.845	-0.815	-0.800	-0.838
MMTIKJ(9, 9; 4)	-0.777	-0.744	-0.291	-0.405
JACOBI2D(13, 13; 4)	-0.760	-0.769	-0.390	-0.428
CHOLESKY(10; 4)	-0.827	-0.953	-0.725	-0.892
CROUT(9; 4)	-0.846	-0.663	-0.213	-0.704
HIMENO(8, 7, 7; 4)	-0.607	-0.475	-0.561	-0.496

5.3 Genetic Algorithm Performance

To evaluate our evolutionary process (Section 4) as a whole, we intend to verify that it can, indeed, find Morton-like array layouts that have a higher simulated fitness than the canonical layouts. To this end, we perform the evolutionary process for each combination of our two simulated processors and eight access patterns, giving rise to a total of sixteen experiments. For all of these experiments, we configure our genetic algorithm to use $\mu = 20$, $\lambda = 20$, and a mutation rate of 25%. We simulate a total of 20 generations in each case.

Figure 5 shows a violin plot of the fitness distribution of all individuals considered during the evolutionary process. Figure 6 shows the evolution of population fitness over the course of our experiments. Note that each of these experiments represents a single evolutionary process. We notice that for the MMTIJK, MMIKJ, JACOBI2D, and HIMENO access patterns, our method does not manage to discover any layouts with higher fitness than the initial population of canonical layouts. In the experiment on the MMIJK access pattern, we discover layouts with a fitness 149.8% higher than the canonical layouts on the Intel Xeon E5-2660 v3 processor, and we improve on the fitness of canonical layouts by 187.5% for the AMD EPYC 7413. We also find layouts with improved fitness for the MMTIKJ (109.6% and 141.1% for the Intel and AMD processors, respectively), CHOLESKY (26.4% and 36.8%), and CROUT (545.9% and 541.1%) access patterns. It is notable that we are able to find layouts with high fitness in few generations.

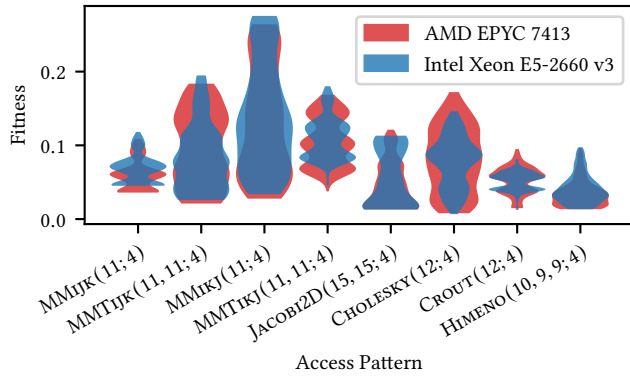


Figure 5: Distribution of the fitness values for all individuals found across evolution experiments for eight access patterns and two processors.

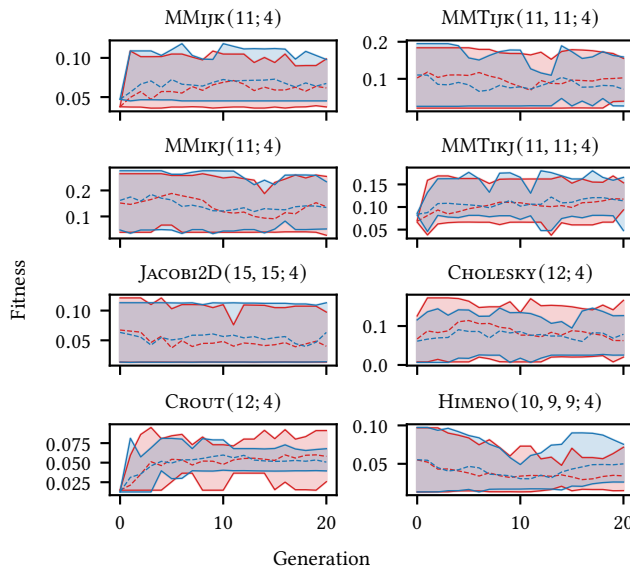


Figure 6: Range of fitness values across eight experiments for the Intel Xeon E5-2660 v3 (blue) and AMD EPYC 7413 (red). Mean fitness values are given by the dashed lines.

5.4 Real-World Performance

In order to evaluate whether the layouts identified by our evolutionary algorithms as superior to canonical layouts are indeed better, we evaluate them on real hardware. We collect the fittest individual from each of the successful evolution experiments—i.e., experiments in which our method improved upon canonical layouts, as indicated by the top boundary in Figure 6 exceeding the maximum fitness in the first generation—and evaluate the performance of those layouts compared to the canonical layouts on real hardware. Given that our genetic algorithm discovered superior layouts for four access patterns—MMiJK, MMTiKJ, CHOLESKY, and CROUT—and that we evaluate a discovered layout and two canonical layouts for each access pattern, this gives rise to twenty-four experiments.

Table 3: Comparison of running time between the best-performing canonical layout and the best-performing layout found by our evolutionary process for four access patterns.

Access pattern	Best can.	Best evo.	Speedup
Intel Xeon E5-2660 v3			
MMiJK(11;4)	17.84 s	10.94 s	63.1%
MMTiKJ(11,11;4)	18.13 s	13.96 s	29.9%
CHOLESKY(12;4)	11.84 s	11.43 s	3.6%
CROUT(12;4)	158.54 s	43.72 s	262.6%
AMD EPYC 7413			
MMiJK(11;4)	37.71 s	9.58 s	293.8%
MMTiKJ(11,11;4)	32.35 s	15.21 s	112.6%
CHOLESKY(12;4)	9.72 s	9.55 s	1.0%
CROUT(12;4)	232.84 s	21.03 s	1007.0%

We repeat each experiment ten times to compensate for run-to-run variance.

The results of our experiments are shown in Table 3; they show that some access patterns—the CHOLESKY pattern in particular—benefit very little from our method, with speed-ups ranging from small on the Haswell processor to insignificant on the Zen 3 processor. The matrix multiplication access patterns benefit more, and performance for these access patterns is improved significantly. The CROUT access pattern stands out as achieving very large speedup—up to a factor ten—from our method. It is worth noting that, in most cases, the Zen 3 processor benefits more from our evolutionary methodology than the Haswell processor; we do not currently have a satisfactory explanation for this behavior.

It is important to note that we do not claim to have discovered a novel way of performing matrix multiplication or matrix decomposition that outperforms existing implementations. Indeed, our experiments are based on relatively naive implementations of these algorithms; high-performance implementations of matrix multiplication commonly rely on tiling to significantly improve the cache behavior of the application [48], and the performance of tiled matrix multiplication surpasses what we achieve in this paper. The purpose of the methodology described in this paper, rather, is to provide an alternative way of improving the cache behavior of an application in a manner which is fully agnostic of the application: unlike tiling and other application-specific optimizations, our methodology of altering the array layouts can be applied to any multi-dimensional problem without the need for application-specific knowledge. In addition, our approach requires few code changes, making it easy to implement.

6 LIMITATIONS AND THREATS TO VALIDITY

Throughout this work, we evaluate cache efficacy through a simplified lens which may reduce the applicability of our methods in more complex, real-world applications. Indeed, we consider accesses to memory in isolation, decoupled from computation and cache-polluting effects. We assume single-threaded execution without scheduling, which means that our caches will not be polluted by processes sharing (parts of) the cache hierarchy, nor will the application have its cached data evicted due to context switching. We

also assume scalar, in-order execution of memory accesses. Finally, we take an optimistic view of cache latencies, using the fastest load-to-use latencies provided by hardware manufacturers; in real-world scenarios, cache latencies may be both more pessimistic and less stable than we assume. The results shown in Section 5.4 indicate, however, that our fitness function is sufficiently accurate to be effective in real hardware.

In addition, the family of array layouts described in this work requires array sizes to be powers of two in each dimension. In applications where this is not the case, arrays must be over-allocated. For n -dimensional applications, using the layouts described in this paper requires over-allocation by a factor of $\mathcal{O}(2^n)$. Furthermore, applications using such layouts must consider the use of SIMD vectorization: it remains an open question which operations on arrays laid out in non-standard ways can be (automatically) vectorized. We have argued for the feasibility of SIMD in Morton-like arrays in Section 3.4.

Finally, our work considers only multiset permutations, in which the rank significance of bits in the input indices is preserved. This decision is based on current commodity hardware, which is capable of efficiently permuting bits only under this condition. There exists an even larger family of layouts in which rank bit significance is not preserved³; such layouts could be of practical use in theoretical future processors with more advanced bit manipulation instructions, or in current FPGA and ASIC devices which permit the implementation of custom bit manipulation operations. Although we have not tested our approach on this further generalization, we are confident that an evolutionary approach like the one presented in this paper could be beneficial in exploring this (even larger) design space.

7 REPRODUCIBILITY AND REUSABILITY

The evolutionary algorithms, scripts for the processing and visualisation of data, and other software used in this paper are permanently archived on Zenodo [60], and have been made available at doi:10.5281/zenodo.10567243. The aforementioned artifact also contains all data that was gathered and processed during the work presented in this paper. For more information about the use of the artifact accompanying this paper, see the included README file.

8 CONCLUSIONS AND FUTURE WORK

In this paper, we have discussed a generalization of the Morton layout for multi-dimensional data and we have shown that there exist families of array layouts with strongly varying cache behavior which, in turn, impact the performance of applications. We have shown how these layouts can be systematically described, and that the number of possible layouts quickly exceed the limits of what can be feasibly explored using exhaustive search. We have proposed a method based on evolutionary algorithms for the exploration of the design space of such layouts. We have evaluated the fitness of different array layouts using cache simulation and we have presented results indicating that our fitness function correlates with real world performance. Furthermore, we have shown that the

methodology described in this paper can be used to improve the performance of applications on real hardware by up to ten times.

In the future, we intend to investigate the use of multi-objective optimization using NSGA-II [19] in order to find array layouts that provide favorable cache behavior across multiple applications. We also intend to explore more advanced genetic algorithms which are known to perform well in combinatorial problems, such as RKGA [12] and BRKGA [26]. It is our belief that exploring more evolutionary strategies will give us more insight into the convergence properties of various methods, and allow us to select the most efficient one. Although our fitness function correlates with real-world performance, the correlation is not perfect; we believe that the efficacy of our method could be improved through the development of more advanced fitness function, perhaps through the use of machine learning methods. In particular, we believe that the field of metric learning may enable us to develop more accurate fitness functions, and we aim to explore this avenue of research in the future. Finally, we aim to expand our research to a broader range of access patterns and hardware, including graphics processing units (GPUs).

ACKNOWLEDGMENTS

The work presented in this paper was done in the context of the CERN Doctoral Student Programme. Many of the experimental results shown in this paper were gathered on the Advanced School for Computing and Imaging (ASCI) DAS-6 compute cluster [11].

REFERENCES

- [1] Andreas Abel and Jan Reineke. 2019. uops.info: Characterizing Latency, Throughput, and Port Usage of Instructions on Intel Microarchitectures. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems* (Providence, Rhode Island, USA) (ASPLOS '19). Association for Computing Machinery, New York, New York, USA, 673–686. <https://doi.org/10.1145/3297858.3304062>
- [2] Advanced Micro Devices, Inc. 2014. *AMD EPYC 7413*. Advanced Micro Devices, Inc. <https://www.amd.com/en/products/cpu/amd-epyc-7413>
- [3] Advanced Micro Devices, Inc. 2020. Software optimization guide for the AMD family 19h processors.
- [4] Amogh Akshintala, Bhushan Jain, Chia-Che Tsai, Michael Ferdman, and Donald E. Porter. 2019. x86-64 Instruction Usage among C/C++ Applications. In *Proceedings of the 12th ACM International Conference on Systems and Storage* (Haifa, Israel) (SYSTOR '19). Association for Computing Machinery, New York, New York, USA, 68–79. <https://doi.org/10.1145/3319647.3325833>
- [5] Ibrahim Al-Kharusi and David W. Walker. 2019. Locality properties of 3D data orderings with application to parallel molecular dynamics simulations. *The International Journal of High Performance Computing Applications* 33, 5 (2019), 998–1018. <https://doi.org/10.1177/1094342019846282>
- [6] João Nuno Ferreira Alves, Luís Manuel Silveira Russo, and Alexandre Francisco. 2022. Cache-Oblivious Hilbert Curve-Based Blocking Scheme for Matrix Transposition. *ACM Trans. Math. Softw.* 48, 4, Article 37 (dec 2022), 28 pages. <https://doi.org/10.1145/3555353>
- [7] Edward J. Anderson and Michael C. Ferris. 1994. Genetic Algorithms for Combinatorial Optimization: The Assemble Line Balancing Problem. *ORSA Journal on Computing* 6, 2 (1994), 161–173. <https://doi.org/10.1287/ijoc.6.2.161>
- [8] Michael Armbrust, Tathagata Das, Liwen Sun, Burak Yavuz, Shixiong Zhu, Mukul Murthy, Joseph Torres, Herman van Hovell, Adrian Ionescu, Alicja Luszczak, Michał undefiniedwitakowski, Michał Szafranski, Xiao Li, Takuya Ueshin, Mostafa Mokhtar, Peter Boncz, Ali Ghodsi, Sameer Paranjpye, Pieter Senster, Reynold Xin, and Matei Zaharia. 2020. Delta lake: high-performance ACID table storage over cloud object stores. *Proc. VLDB Endow.* 13, 12 (aug 2020), 3411–3424. <https://doi.org/10.14778/3415478.3415560>
- [9] Krste Asanovic, Rastislav Bodik, James Demmel, Tony Keaveny, Kurt Keutzer, John Kubiatowicz, Nelson Morgan, David Patterson, Koushik Sen, John Wawrzynek, David Wessel, and Katherine Yelick. 2009. A View of the Parallel Computing Landscape. *Commun. ACM* 52, 10 (oct 2009), 56–67. <https://doi.org/10.1145/1562764.1562783>

³That is to say, the layout $[0_0, 0_1, 1_0, 1_1]$ (which draws its least significant bit from the least significant bit of the first index) is distinct from the layout $[0_1, 0_0, 1_0, 1_1]$ (which instead draws its least significant bit from the *second-least* significant bit of the first index).

- [10] Michael Bader. 2012. *Space-filling curves: an introduction with applications in scientific computing*. Vol. 9. Springer Science & Business Media, Berlin, Heidelberg, Germany.
- [11] Henri Bal, Dick Epema, Cees de Laat, Rob van Nieuwpoort, John Romein, Frank Seinstra, Cees Snoek, and Harry Wijshoff. 2016. A Medium-Scale Distributed System for Computer Science Research: Infrastructure for the Long Term. *Computer* 49, 05 (5 2016), 54–63. <https://doi.org/10.1109/MC.2016.127>
- [12] James Carl Bean. 1994. Genetic Algorithms and Random Keys for Sequencing and Optimization. *ORSA Journal on Computing* 6, 2 (1994), 154–160. <https://doi.org/10.1287/ijoc.6.2.154>
- [13] Richard A. Brualdi. 1977. *Introductory combinatorics* (5 ed.). Pearson Education, London, United Kingdom.
- [14] Siddhartha Chatterjee, Vibhor V. Jain, Alvin R. Lebeck, Shyam Mundhra, and Mithuna Thottethodi. 1999. Nonlinear Array Layouts for Hierarchical Memory Systems. In *Proceedings of the 13th International Conference on Supercomputing* (Rhodes, Greece) (ICS '99). Association for Computing Machinery, New York, New York, USA, 444–453. <https://doi.org/10.1145/305138.305231>
- [15] Siddhartha Chatterjee, Alvin R. Lebeck, Praveen K. Patnala, and Mithuna Thottethodi. 1999. Recursive Array Layouts and Fast Parallel Matrix Multiplication. In *Proceedings of the Eleventh Annual ACM Symposium on Parallel Algorithms and Architectures* (Saint Malo, France) (SPA '99). Association for Computing Machinery, New York, New York, USA, 222–231. <https://doi.org/10.1145/305619.305645>
- [16] Shuai Che, Jeremy W. Sheaffer, and Kevin Skadron. 2011. Dymaxion: Optimizing Memory Access Patterns for Heterogeneous Systems. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis* (Seattle, Washington) (SC '11). Association for Computing Machinery, New York, New York, USA, Article 13, 11 pages. <https://doi.org/10.1145/2063384.2063401>
- [17] Vali Codreanu, Joerg Hertzler, Cristian Morales, Jorge Rodriguez, Ole Widar Saastad, and Martin Stachon. 2017. Best practice guide Haswell / Broadwell.
- [18] Lawrence Davis. 1985. Applying Adaptive Algorithms to Epistatic Domains. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 1* (Los Angeles, California, USA) (IJCAI'85). Morgan Kaufmann Publishers Inc., San Francisco, California, USA, 162–164.
- [19] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and Thirunavukarasu-Murariyan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197. <https://doi.org/10.1109/4235.996017>
- [20] Daryl Deford and Ananth Kalyanaraman. 2013. Empirical Analysis of Space-Filling Curves for Scientific Computing Applications. In *2013 42nd International Conference on Parallel Processing* (Lyon, France). IEEE, New York, New York, USA, 170–179. <https://doi.org/10.1109/ICPP.2013.26>
- [21] Agoston E. Eiben and Jim E. Smith. 2015. *Representation, Mutation, and Recombination*. Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, 49–78. https://doi.org/10.1007/978-3-662-44874-8_4
- [22] Mark Evers, Leslie Barnes, and Mike Clark. 2022. The AMD Next-Generation “Zen 3” Core. *IEEE Micro* 42, 3 (2022), 7–12. <https://doi.org/10.1109/MM.2022.3152788>
- [23] Matteo Frigo, Charles E. Leiserson, Harald Prokop, and Sridhar Ramachandran. 1999. Cache-oblivious algorithms. In *40th Annual Symposium on Foundations of Computer Science* (New York, New York, USA) (FOCS '99). IEEE, New York, New York, USA, 285–297. <https://doi.org/10.1109/SFCS.1999.814600>
- [24] Steven T. Gabriel and David S. Wise. 2004. The Opie Compiler from Row-Major Source to Morton-Ordered Matrices. In *Proceedings of the 3rd Workshop on Memory Performance Issues: In Conjunction with the 31st International Symposium on Computer Architecture* (Munich, Germany) (WMPPI '04). Association for Computing Machinery, New York, New York, USA, 136–144. <https://doi.org/10.1145/1054943.1054962>
- [25] Andrew Gartland-Jones and Peter Copley. 2003. The Suitability of Genetic Algorithms for Musical Composition. *Contemporary Music Review* 22, 3 (2003), 43–55. <https://doi.org/10.1080/0749446032000150870>
- [26] José Fernando Gonçalves and Mauricio G. C. Resende. 2011. Biased random-key genetic algorithms for combinatorial optimization. *Journal of Heuristics* 17, 5 (01 Oct 2011), 487–525. <https://doi.org/10.1007/s10732-010-9143-1>
- [27] Peter Gottschling, David S. Wise, and Michael D. Adams. 2007. Representation-Transparent Matrix Algorithms with Scalable Performance. In *Proceedings of the 21st Annual International Conference on Supercomputing* (Seattle, Washington, USA) (ICS '07). Association for Computing Machinery, New York, New York, USA, 116–125. <https://doi.org/10.1145/1274971.1274989>
- [28] Peter Gottschling, David S. Wise, and Adwait Joshi. 2009. Generic support of algorithmic and structural recursion for scientific computing. *International Journal of Parallel, Emergent and Distributed Systems* 24, 6 (2009), 479–503. <https://doi.org/10.1080/17445760902758560>
- [29] Per Hammarlund, Alberto J. Martinez, Atiq A. Bajwa, David L. Hill, Erik Hallnor, Hong Jiang, Martin Dixon, Michael Derr, Mikal Hunsaker, Rajesh Kumar, Randy B. Osborne, Ravi Rajwar, Ronak Singhal, Reynold D'Sa, Robert Chappell, Shiv Kaushik, Srinivas Chennupati, Stephan Jourdan, Steve Gunther, Tom Piazza, and Ted Burton. 2014. Haswell: The Fourth-Generation Intel Core Processor. *IEEE Micro* 34, 2 (2014), 6–20. <https://doi.org/10.1109/MM.2014.10>
- [30] Julian Hammer, Jan Eitzinger, Georg Hager, and Gerhard Wellein. 2017. Kerncraft: A Tool for Analytic Performance Modeling of Loop Kernels. In *Tools for High Performance Computing 2016* (Stuttgart, Germany). Christoph Niethammer, José Gracia, Tobias Hilbrich, Andreas Knüpfer, Michael M. Resch, and Wolfgang E. Nagel (Eds.). Springer International Publishing, Cham, Switzerland, 1–22. https://doi.org/10.1007/978-3-319-56702-0_1
- [31] Brian Hegerty, Chih-Cheng Hung, and Kristen Kasprak. 2009. A comparative study on differential evolution and genetic algorithms for some combinatorial problems. In *Proceedings of the 8th Mexican International Conference on Artificial Intelligence* (Guanajuato, Mexico), Vol. 9. Springer Verlag, Berlin, Heidelberg, Germany, 13.
- [32] David Hilbert. 1891. Ueber die stetige Abbildung einer Linie auf ein Flächenstück. *Math. Ann.* 38, 3 (1891), 459–460. <https://doi.org/10.1007/BF01199431>
- [33] Ryutaro Himeno. 2001. The Riken Himeno CFD Benchmark. <https://i.riken.jp/en/supercom/documents/himenobmt/>
- [34] John Henry Holland. 1992. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT Press, Cambridge, Massachusetts, USA.
- [35] John Henry Holland. 1992. Genetic algorithms. *Scientific American* 267, 1 (1992), 66–73.
- [36] OEIS Foundation Inc. 2023. Moser–de Bruijn sequence, entry A000695 in The On-Line Encyclopedia of Integer Sequences. <https://oeis.org/A000695>
- [37] Intel Corporation. 2014. *Intel Xeon Processor E5-2660 v3*. Intel Corporation. <https://ark.intel.com/content/www/us/en/ark/products/81706/intel-xeon-processor-e52660-v3-25m-cache-2-60-ghz.html>
- [38] Intel Corporation. 2023. Intel 64 and IA-32 Architectures Software Developer’s Manual.
- [39] Wenzel Jakob, Jason Rhinelander, and Dean Moldovan. 2017. pybind11 – Seamless operability between C++11 and Python. <https://github.com/pybind/pybind11>
- [40] Byunghyun Jang, Dana Schaa, Perhaad Mistry, and David Kaeli. 2011. Exploiting Memory Access Patterns to Improve Memory Performance in Data-Parallel Architectures. *IEEE Transactions on Parallel and Distributed Systems* 22, 1 (2011), 105–118. <https://doi.org/10.1109/TPDS.2010.107>
- [41] Tamara G. Kolda and Brett W. Bader. 2009. Tensor Decompositions and Applications. *SIAM Rev.* 51, 3 (2009), 455–500. <https://doi.org/10.1137/07070111X>
- [42] Markus Kowarschik and Christian Weiß. 2003. *An Overview of Cache Optimization Techniques and Cache-Aware Numerical Algorithms*. Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, 213–232. https://doi.org/10.1007/3-540-36574-5_10
- [43] Jan Laukemann, Julian Hammer, Johannes Hofmann, Georg Hager, and Gerhard Wellein. 2018. Automated Instruction Stream Throughput Prediction for Intel and AMD Microarchitectures. In *2018 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)* (Dallas, Texas, USA). IEEE, New York, New York, USA, 121–131. <https://doi.org/10.1109/PMBS.2018.8641578>
- [44] John Mellor-Crummey, David Whalley, and Ken Kennedy. 2001. Improving Memory Hierarchy Performance for Irregular Applications Using Data and Computation Reorderings. *International Journal of Parallel Programming* 29, 3 (01 Jun 2001), 217–247. <https://doi.org/10.1023/A:1011119519789>
- [45] Brad L. Miller and David E. Goldberg. 1995. Genetic algorithms, tournament selection, and the effects of noise. *Complex systems* 9, 3 (1995), 193–212.
- [46] Guy Macdonald Morton. 1966. *A Computer Oriented Geodetic Data Base and a New Technique in File Sequencing*. Technical Report. International Business Machines Company.
- [47] Heinz Mühlenbein. 1991. Parallel genetic algorithms, population genetics and combinatorial optimization. In *Parallelism, Learning, Evolution* (Wildbad Kreuth, Germany). Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, 398–406.
- [48] Neungsoo Park, Bo Hong, and Viktor K. Prasanna. 2003. Tiling, block data layout, and memory hierarchy performance. *IEEE Transactions on Parallel and Distributed Systems* 14, 7 (2003), 640–654. <https://doi.org/10.1109/TPDS.2003.1214317>
- [49] G. Pavaai and T. V. Geetha. 2016. A Survey on Crossover Operators. *ACM Comput. Surv.* 49, 4, Article 72 (dec 2016), 43 pages. <https://doi.org/10.1145/3009966>
- [50] Filip Pawłowski, Bora Uçar, and Albert-Jan Nicholas Yzelman. 2019. A multi-dimensional Morton-ordered block storage for mode-oblivious tensor computations. *Journal of Computational Science* 33 (2019), 34–44. <https://doi.org/10.1016/j.jocs.2019.02.007>
- [51] Martin Perdacher, Claudia Plant, and Christian Böhm. 2020. Improved Data Locality Using Morton-order Curve on the Example of LU Decomposition. In *2020 IEEE International Conference on Big Data (Big Data)* (Atlanta, Georgia, USA). IEEE, New York, New York, USA, 351–360. <https://doi.org/10.1109/BigData50022.2020.9378385>
- [52] Andy D. Pimentel. 2017. Exploring Exploration: A Tutorial Introduction to Embedded Systems Design Space Exploration. *IEEE Design & Test* 34, 1 (2017), 77–90. <https://doi.org/10.1109/MDAT.2016.2626445>
- [53] Moinuddin K. Qureshi, Aamer Jaleel, Yale N. Patt, Simon C. Steely, and Joel Emer. 2008. Set-Dueling-Controlled Adaptive Insertion for High-Performance Caching. *IEEE Micro* 28, 1 (2008), 91–98. <https://doi.org/10.1109/MM.2008.14>

- [54] Dolly Sapra and Andy D. Pimentel. 2020. Constrained Evolutionary Piecemeal Training to Design Convolutional Neural Networks. In *Trends in Artificial Intelligence Theory and Applications. Artificial Intelligence Practices*. Springer International Publishing, Cham, Switzerland, 709–721.
- [55] S.N. Sivanandam and S.N. Deepa. 2008. *Genetic Algorithms*. Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, 15–37. https://doi.org/10.1007/978-3-540-73190-0_2
- [56] Adam Slowik and Halina Kwasnicka. 2020. Evolutionary algorithms and their applications to engineering problems. *Neural Computing and Applications* 32, 16 (01 Aug 2020), 12363–12379. <https://doi.org/10.1007/s00521-020-04832-8>
- [57] M. Srinivas and Lalit M. Patnaik. 1994. Genetic algorithms: a survey. *Computer* 27, 6 (1994), 17–26. <https://doi.org/10.1109/2.294849>
- [58] Nigel Stephens, Stuart Biles, Matthias Boettcher, Jacob Eapen, Mbou Eyole, Giacomo Gabrielli, Matt Horsnell, Grigorios Magklis, Alejandro Martinez, Nathanael Premillieu, Alastair Reid, Alejandro Rico, and Paul Walker. 2017. The ARM Scalable Vector Extension. *IEEE Micro* 37, 2 (2017), 26–39. <https://doi.org/10.1109/MM.2017.35>
- [59] Stephen Nicholas Swatman, Ana-Lucia Varbanescu, Andy D. Pimentel, Andreas Salzburger, and Attila Krasznahorkay. 2023. Systematically Exploring High-Performance Representations of Vector Fields Through Compile-Time Composition. In *Proceedings of the 2023 ACM/SPEC International Conference on Performance Engineering (Coimbra, Portugal) (ICPE '23)*. Association for Computing Machinery, New York, New York, USA, 55–66. <https://doi.org/10.1145/3578244.3583723>
- [60] Stephen Nicholas Swatman, Ana-Lucia Varbanescu, Andy D. Pimentel, Andreas Salzburger, and Attila Krasznahorkay. 2024. *Artifact for "Using Evolutionary Algorithms to Find Cache-Friendly Generalized Morton Layouts for Arrays"*. <https://doi.org/10.5281/zenodo.10567243>
- [61] Lothar Terfloth and Johann Gasteiger. 2001. Neural networks and genetic algorithms in drug design. *Drug Discovery Today* 6 (2001), 102–108. [https://doi.org/10.1016/S1359-6446\(01\)00173-8](https://doi.org/10.1016/S1359-6446(01)00173-8)
- [62] Jeyarajan Thiyaalingam, Olav Beckmann, and Paul H. J. Kelly. 2006. Is Morton layout competitive for large two-dimensional arrays yet? *Concurrency and Computation: Practice and Experience* 18, 11 (2006), 1509–1539. <https://doi.org/10.1002/cpe.1018>
- [63] Pepe Vila, Pierre Ganty, Marco Guarnieri, and Boris Köpf. 2020. CacheQuery: Learning Replacement Policies from Hardware Caches. In *Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation (London, United Kingdom) (PLDI 2020)*. Association for Computing Machinery, New York, New York, USA, 519–532. <https://doi.org/10.1145/3385412.3386008>
- [64] David W. Walker. 2018. Morton ordering of 2D arrays for efficient access to hierarchical memory. *The International Journal of High Performance Computing Applications* 32, 1 (2018), 189–203. <https://doi.org/10.1177/1094342017725568>
- [65] David W. Walker and Anthony Skjellum. 2023. The Impact of Space-Filling Curves on Data Movement in Parallel Systems. <https://doi.org/10.48550/arXiv.2307.07828> [cs.DC]
- [66] Jonathan Weinberg, Michael O. McCracken, Erich Strohmaier, and Allan Snively. 2005. Quantifying Locality In The Memory Access Patterns of HPC Applications. In *SC '05: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing* (Seattle, Washington, USA). IEEE, New York, New York, USA, 50–50. <https://doi.org/10.1109/SC.2005.59>
- [67] Albert-Jan Nicholas Yzelman and Rob Hendrik Bisseling. 2012. A Cache-Oblivious Sparse Matrix–Vector Multiplication Scheme Based on the Hilbert Curve. In *Progress in Industrial Mathematics at ECMI 2010*. Springer Berlin Heidelberg, Berlin, Heidelberg, Germany, 627–633.