Future Generation Computer Systems

TraTSA: A Transprecision Framework for Efficient Time Series Analysis

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full Length Article
Keywords:	Time Series Analysis; Transprecision Computing; Floating-Point Unit; FPGA; Parallel Architectures
Corresponding Author:	Ivan Fernandez, PhD Student University of Malaga: Universidad de Malaga Malaga, Malaga SPAIN
First Author:	Ivan Fernandez, PhD Student
Order of Authors:	Ivan Fernandez, PhD Student
	Ricardo Quislant
	Sonia Gonzalez-Navarro
	Eladio Gutierrez
	Oscar Plata
Abstract:	Time series analysis (TSA) comprises methods for extracting information in domains as diverse as medicine, seismology, speech recognition and economics. Matrix Profile (MP) is the state-of-the-art TSA technique, which provides the most similar neighbour to each subsequence of the time series. However, this computation requires a huge amount of floating-point (FP) operations, which are a major contributor (approx 50%) to the energy consumption in modern computing platforms. In this sense, Transprecision Computing has recently emerged as a promising approach to improve energy efficiency and performance by using less bits in FP operations while providing accurate results. In this work, we present TraTSA, the first transprecision framework for efficient time series analysis based on MP. TraTSA allows the user to deploy a high-performance and energy-efficient computing solution with the exact precision required by the TSA application. To this end, we first propose implementations of TraTSA for both commodity CPU and FPGA platforms. Second, we propose an accuracy metric to compare the results with the double precision MP. Third, we study MP's accuracy when using a transprecision approach. Finally, our evaluation shows that, while obtaining results accurate enough, the FPGA transprecision MP i) is 22.75x faster than a 72-core server, and ii) the energy consumption is up to 3.3x lower than the double precision executions.
Suggested Reviewers:	Giuseppe Tagliavini giuseppe.tagliavini@unibo.it High level of transprecision computing knowledge
	Eamonn Keogh eamonn@cs.ucr.edu Main author of Matrix Profile algorithm

Cover Letter

Iván Fernández Vega
Dpt. of Computer Architecture
University of Malaga (Spain)
+34692785519 | ivanfv@uma.es

Professor Michela Taufer July 31, 2021
Editor-in-Chief, Future Generation Computer Systems

Dear Michela Taufer and other readers,

Please find enclosed our manuscript entitled: "TraTSA: A Transprecision Framework for Efficient Time Series Analysis". We are submitting this work to Future Generation Computer Systems. We confirm that this manuscript has not been published elsewhere and is not under consideration by any other journal. This manuscript is a significantly extended version of a conference paper that we include in our submission, along with a summary of changes.

Our work aims to provide the community with a tool to jumpstart the research niche of transprecision time series analysis. Time series analysis is a relevant task able to extract information in a wide variety of domains. However, it requires a huge amount of floating point operations, which are a major contributor to energy consumption in modern computing platforms. To overcome this issue, transprecision computing lowers the number of bits in data representation, emerging as a paradigm to 1) reduce energy consumption, 2) improve performance and 3) limit arithmetic operators complexity. Based on this approach, we design both CPU and FPGA implementations of the state-of-the-art time series analysis, evaluating the benefits and tradeoffs when varying the data precision. To the best of our knowledge, this is the first work that explores transprecision time series analysis and proposes an FPGA-based solution to accelerate it, which is 22.75x times faster than a 72-core server.

We thank you in advance for considering our work. Should you have any request concerning this manuscript, please do not hesitate to contact me.

Sincerely,

Ivan Fernandez Vega

Author Names and their Contribution

Ivan Fernandez: Conceptualization, Investigation, Methodology, Software, Writing - Original Draft

Ricardo Quislant: Conceptualization, Investigation, Software, Writing - Review & Editing, Data Curation

Sonia Gonzalez-Navarro: Validation, Writing - Review & Editing, Formal analysis

Eladio Gutierrez: Conceptualization, Writing - Review & Editing, Supervision

Oscar Plata: Conceptualization, Writing - Review & Editing, Supervision

TraTSA: A Transprecision Framework for Efficient Time Series Analysis

Ivan Fernandez, Ricardo Quislant, Sonia González-Navarro, Eladio Gutierrez and Oscar Plata

==========

Summary of Changes

Our preliminary work accepted in SBAC-PAD'20 has been completely upgraded with the following changes:

- We improve the motivation section adding an analysis of the number of operations involved in a self-join analysis of a time series.
- We develop a new FPGA implementation that significantly speeds up the computation of transprecision time series analysis. This FPGA implementation is based on an Alveo U50 board and provides hardware support for arbitrary precision analysis. It takes advantage of several optimization techniques to efficiently exploit the FPGA resources.
- We create a wrapper that eases the user experience of the proposed framework. Concretely, this wrapper receives time series and its configuration file, and then it invokes the corresponding backend accordingly. Finally, it provides the user the output in a proper format.
- We add six time series in our evaluation, which are about ten times larger than the ones evaluated in our SBAC-PAD submission.
- We include the recalculation factor as part of the FPGA implementation. This recalculation factor allows the user to establish a threshold between accuracy and performance, by resetting the accumulated errors after a given interval. This feature is specially useful in long time series analysis.
- We extend our evaluation analyzing the added longer time series, comprehensively evaluating the accuracy of the results depending on the data precision and the recalculation factor.
- We add comparisons of the execution time and energy savings of our proposal with respect to commodity platforms.

TraTSA: A Transprecision Framework for Efficient Time Series Analysis

Ivan Fernandez[a], Ricardo Quislant[a], Sonia Gonzalez-Navarro[a], Eladio Gutierrez[a], Oscar Plata[a]

[a]*Department of Computer Architecture, University of Malaga, Spain*

Abstract

Time series analysis (TSA) comprises methods for extracting information in domains as diverse as medicine, seismology, speech recognition and economics. *Matrix Profile* (MP) is the state-of-the-art TSA technique, which provides the most similar neighbour to each subsequence of the time series. However, this computation requires a huge amount of floating-point (FP) operations, which are a major contributor (≈ 50%) to the energy consumption in modern computing platforms. In this sense, *Transprecision Computing* has recently emerged as a promising approach to improve energy efficiency and performance by using less bits in FP operations while providing accurate results.

In this work, we present *TraTSA*, the first transprecision framework for efficient time series analysis based on MP. TraTSA allows the user to deploy a high-performance and energy-efficient computing solution with the exact precision required by the TSA application. To this end, we first propose implementations of TraTSA for both commodity CPU and FPGA platforms. Second, we propose an accuracy metric to compare the results with the double precision MP. Third, we study MP's accuracy when using a transprecision approach. Finally, our evaluation shows that, while obtaining results accurate enough, the FPGA transprecision MP i) is 22.75× faster than a 72-core server, and ii) the energy consumption is up to 3.3× lower than the double precision executions.

Keywords: Time Series Analysis, Transprecision Computing, Floating-Point Unit, FPGA, Parallel Architectures

1. Introduction

Time series are ordered sets of samples of real-valued variables that can contain millions of observations. Time series analysis seeks extracting information in a large variety of domains such as epidemiology [1], DNA analysis [2], economics [3], speech recognition [4], traffic prediction [5], energy conservation [6], seismology [7], medicine [8] and many more [9]. Particularly, *motif* [10] (similarity) and *discord* [11] (anomaly) discovery are two of the most frequently used primitives in time series data mining [12, 13, 14, 15, 16, 17, 18]. These primitives try to solve the all-pairs-similarity-search problem (also known as similarity join). That is, given a time series sliced in subsequences, retrieving the most similar and dissimilar subsequences.

The state-of-the-art method for motif and discord discovery is *Matrix Profile* [19]. This method solves the similarity join problem and allows time-manageable computation of very large time series datasets. The similarity join has many applications [20] such as community discovery, duplicate text detection, collaborative filtering for social networks, clustering, and query refinement for web search.

In this work, we focus on this technique, which provides full joins without the need to specify a similarity threshold, a very challenging task in this domain. When analyzing a time series, the matrix profile is another time series representing the minimum distance subsequence for each subsequence (motifs). On the other hand, maximum values of the profile highlight the most dissimilar subsequences (discords).

We evaluate the latest Euclidean-based implementations of matrix profile (SCRIMP [21] and SCAMP [22]) and find that a huge number of floating-point (FP) arithmetic operations are needed in order to analyze even short time series. In this sense, *transprecision computing* [23] has recently emerged as a promising approach to i) improve energy efficiency, ii) provide better performance, iii) reduce area footprint, and iv) reduce memory bandwidth by tolerating some loss of accuracy in computed results. This paradigm reduces the number of bits for the exponent and the mantissa in FP operations in a flexible way, depending on the requirements of the application. It is well known that FP operations are a major contributor (≈ 50%) [24] to the energy consumption in modern computing platforms. Thus, transprecision has the potential to provide efficiently designed with the required precision by the application.

Our goal in this work is to provide a set of tools to jumpstart the research niche of transprecision time series analysis, to achieve high-performance and energy-efficient computing for a wide range of applications. This way, new platforms can be designed that benefit from reduced-bit-count FP operations tailored to each application (e.g. using a transprecision Floating-Point-Unit (FPU), like FPNew [25]). This opens up the opportunity to detect important events on mobile and embedded devices, where energy is a critical concern. Those devices can be used, for example, to prevent ecological disasters or medical issues (e.g. for early earthquake detection [26] or to predict a heart attack [27]).

Email addresses: ivanfv@uma.es (Ivan Fernandez), quislant@uma.es (Ricardo Quislant), sgn@uma.es (Sonia Gonzalez-Navarro), eladio@uma.es (Eladio Gutierrez), oplata@uma.es (Oscar Plata)

To this end, we introduce *TraTSA*, the *first* transprecision framework for time series analysis. TraTSA, which is open-source and will be freely available to the community, provides fast and user-friendly transprecision matrix profile computing thanks to its CPU and FPGA implementations. We evaluate TraTSA with use cases of real datasets from different domains and lengths, analyzing the trade-offs between arithmetic precision and result accuracy using a proposed metric. Additionally, we present the energy savings of a real transprecision FPU.

The contributions of this work are the following:

- We develop TraTSA, the first framework for flexible transprecision matrix profile computation, which includes three transprecision implementations (`TranSCRIMP`, `TranSCAMP` and `TranSCAMPfpga`) based on FlexFloat [28] and cpfp-FPGA [29] libraries. We integrate those implementations using a Python wrapper that provides a user-friendly interface.
- We propose a new metric, called *Top-K Accuracy*, to measure the accuracy in the discovery of motifs and discords of a transformed time series (e.g. lower precision) with respect to the original time series.
- We provide a detailed accuracy analysis of `TranSCRIMP`, `TranSCAMP` and `TranSCAMPfpga` using the *Top-K Accuracy* metric while varying exponent and mantissa bit count combinations.
- We evaluate the potential benefits of using a transprecision FPU for matrix profile, finding that energy efficiency can be improved up to 3.3× compared with double precision with proper precision adjustments, while providing results accurate enough.

2. Background

In this section, we provide a background on time series analysis based on matrix profile, transprecision computing and FPGA acceleration.

2.1. Time Series Analysis. The Matrix Profile

A *time series* T is a sequence of consecutive n data points t_i, $1 \leq i \leq n$, collected over time. Time series data is typically stored using a floating-point representation. Let be $T_{i,m}$ a subsequence of T, where i is the index of its first data point, T_i, and m is the number of data points in the subset, with $1 \leq i$, and $m \leq n$. In the literature, $T_{i,m}$ is also called a *window* of length m. Fig. 1 shows an example of a sinusoidal time series with two subsequences highlighted, $T_{i,m}$ and $T_{j,m}$.

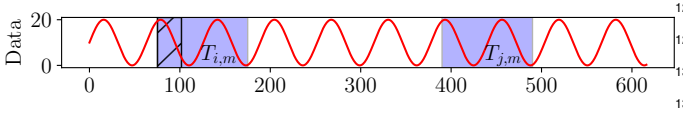

Figure 1: Example of two subsequences, $T_{i,m}$ and $T_{j,m}$, of a given time series. When computing the matrix profile P, subsequences starting in the exclusion zone of $T_{i,m}$ are ignored for their high similarity.

One way to measure the similarity between two time series subsequences is to use the *z-normalized Euclidean distance* [21], $d_{i,j}$, which is calculated as follows:

$$d_{i,j} = \sqrt{2m\left(1 - \frac{Q_{i,j} - m\mu_i\mu_j}{m\sigma_i\sigma_j}\right)} \quad (1)$$

where $Q_{i,j}$ is the dot product of subsequences $T_{i,m}$ and $T_{j,m}$. μ_x and σ_x are the mean and the standard deviation of the data points in $T_{x,m}$, respectively.

Now we can use this distance measure to find the most similar subsequences out of all subsequences of a time series T. There are three steps to this procedure. First, building a symmetric $(n–m+1)\times(n–m+1)$ matrix D, called *distance matrix*, with a window size m and a time series of length n. Each D cell, $d_{i,j}$, stores the distance between two subsequences, $T_{i,m}$ and $T_{j,m}$. Thus each row (or column) of D stores the distances between a subsequence of T and every subsequence of T. Second, finding the two subsequences whose distance is minimum (i.e. most similar sequences) in each row (or column) of D. This can be computed by building the *matrix profile*, P, which is a vector of size $n - m + 1$. Each cell P_i in P stores the minimum value found in the i^{th} row (or column) of D. Third, finding the indices of the most similar subsequences of the matrix profile. This requires building another vector I, called *matrix profile index*, of the same size as that of P, where $I_i = j$ if $d_{i,j} = P_i$. This way, P contains the minimum distances between subsequences of T while I is the vector of "pointers" to the location of these subsequences. Fig. 2 depicts an example of the distance matrix (D), the matrix profile (P), and the matrix profile index (I) for the time series in Fig. 1.

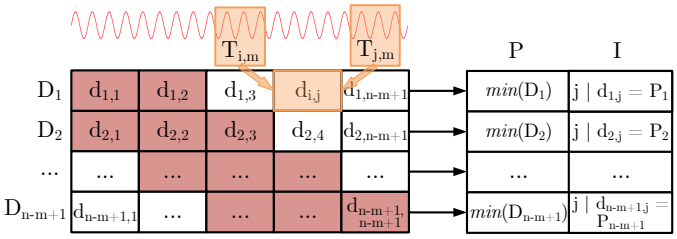

Figure 2: Computation of the matrix profile P and the profile index I from the distance matrix D. P_i is the minimum distance of each row (or column) D_i. I_i is the index of the subsequence providing the minimum.

Notice that the D matrix is symmetric and the neighboring subsequences of $T_{i,m}$ are highly similar to it (i.e. $d_{i,i+1} \approx 0$) due to the overlaps. Thus, we can exclude these subsequences from computing D to find similar subsequences other than the neighboring ones. This is done by defining an exclusion zone (red cells in Fig. 2) for each subsequence. In general, the exclusion zone for $T_{i,m}$ is $\frac{m}{4}$ [21].

From the complexity point of view, matrix profile algorithms present two main challenges: memory footprint and a huge number of floating point operations.

Memory footprint. The size of the distance matrix D can be huge for large time series, so it is not convenient to store it in

memory. For example, an earthquake sequence from a seismograph, consisting of 24000 elements [30], needs approximately 1.3GB of memory using double precision floating point representation. However, a 650000 electrocardiogram (ECG) time series from the MIT-BIH arrhythmia database [31] requires about 850GB of memory. Furthermore, we need to maintain the matrix profile P, the matrix profile index I and the time series T. With the aim to overcome this issue, matrix profile algorithms are designed to store only the matrix profile and the matrix profile index arrays, computing the minimum distances $d_{i,j}$ on the fly.

Number of floating-point operations. The number of operations required even for short time series is the dominant part of the algorithm and increases quicker than linear with the time series length. Table 1 depicts the number of operations performed for a time series of 180K elements. As an example for this particularly case, matrix profile needs to perform more than 64 billion multiplications. Thus, efficient floating-point units need to be designed to reduce the energy consumption and improve performance by increasing SIMD parallelism. Matrix profile implementations are usually based on double (64-bit) or single (32-bit) precision. The use of lower bit counts has not been well studied yet and is the main scope of this paper.

Table 1: Number of arithmetic floating-point operations performed in a time series of 180000 elements and a window size of 512 elements

Operation	Number	Operation	Number
Additions	16084915120	Subtractions	183664640
Multiplications	64340019200	FMAs	91832320
Inversions	16084915120	Comparisons	32170906916

2.2. Transprecision Computing

Transprecision Computing aims to boost energy efficiency and performance by exploiting numeric approximation in both hardware and software. It enables fine control over the precision of floating point arithmetic in space and time (where and when to use it). The key difference with approximate computing is that the first one guarantees the error while the latter one provides uncertainty. This approach leads to significant energy savings and performance improvements without sacrificing overall quality of results.

Transprecision computing can be applied to the entire algorithm by setting fixed exponent and mantissa widths for every floating point operation. However, it is possible to change the precision to different parts of the code to find a trade-off between the accuracy of the results and energy efficiency. Mixed precision of double and single floating point operations has been successfully used in the past [32, 33] with a significant gain in performance.

Besides the IEEE-754 standard, transprecision computing allows the use of arbitrary exponent and mantissa bit combinations. However, the design of arbitrary precision floating-point units (FPUs) can be challenging and presents difficulties when integrating in computing platforms. Because of this reason, already designed FPUs typically support a fixed number of exponent and mantissa combinations. As an example, Fig. 3 shows the types [24] that we use in this work for the energy evaluation.

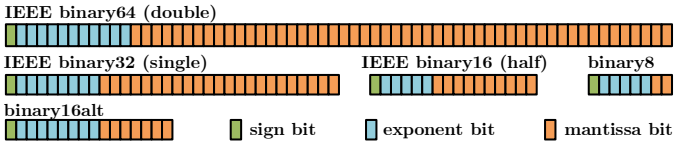

Figure 3: Overview of the floating-point types used for energy evaluation.

Table 2 summarizes the bit count for each floating point datatype, along with the approximate range and smallest number that can be represented. Notice that, while *binary16* and *binary8* have the same range, the first one provides narrower steps between numbers and more precision.

Table 2: Floating-point bit counts, ranges and smallest numbers

Type	Exp	Man	Range (≈)	Smallest (≈)
bin64	11	52	$\{-1.8 \cdot 10^{308}, 1.8 \cdot 10^{308}\}$	$2.2 \cdot 10^{-308}$
bin32	8	23	$\{-3.4 \cdot 10^{38}, 3.4 \cdot 10^{38}\}$	$1.2 \cdot 10^{-38}$
bin16	5	10	{-65504, 65504}	$6.1 \cdot 10^{-5}$
bin16alt	8	7	$\{-3.4 \cdot 10^{38}, 3.4 \cdot 10^{38}\}$	$1.2 \cdot 10^{-38}$
bin8	5	2	{-57344, 57344}	$6.1 \cdot 10^{-5}$

We find transprecision Floating-Point Units (FPU) already proposed in the literature, which aim to take advantage of transprecision computing in terms of energy, performance and area. The first silicon implementation of a 64-bit transprecision FPU can be found in [34], using 22nm technology node. This FPU supports the floating-point types depicted in Fig. 3, and the key idea behind it is to operate in scalars (64-bit) or in SIMD vectors of 2 elements (32-bit), 4 elements (16-bit) or 8 elements (8-bit). The authors evaluate the FPU using a RISC-V processor core (but can be used in other platforms, like FPGAs or ASICs) and obtain speedups up to 7.3× while reducing energy up to 7.94× with respect to *double* precision approaches. The source code of this FPU (known as FPnew) can be found in [25].

On the other hand, transprecision libraries aim to enable transprecision software emulation in commodity architectures. An example of CPU-based transprecision emulation is FlexFloat [28], which is written in C. The main advantage of this library is that it can be executed in almost any commodity platform, but it presents a main drawback: the significant overhead introduced by the software emulation. In this sense, we observe that the FlexFloat execution time increases the native IEEE double execution time by 200× for a given computing platform. One way to overcome this issue is the use of an FPGA architecture, including custom-precision units from cpfp-FPGA [29] library, which is intended for HLS synthesis. While not being a pure transprecision implementation from the architectural point of view (e.g. data is stored using 32 bits in memory and then converted to the desired width), it allows time-manageable evaluation of large time series (millions of elements). The key motivation to consider the FPGA acceleration is to avoid the emulation cost of the software-based one with the use of dedicated transprecision hardware.

2.3. FPGA Acceleration

Field Programmable Gate Arrays (FPGAs) [35] are programmable devices comprising logic blocks which connection and behavior can be defined by the user. This kind of architectures can be programmed to implement a wide variety of algorithms and even full-fledged systems, such as a complex RISC-V core [36].

The main advantage of using an FPGA to accelerate a given algorithm, compared with implementing a full system, is that more FPGA resources are available for the implementation of the required functionality. For instance, resources devoted to branch prediction or instruction decoding in the implementation of a general-purpose core can be used to place more floating-point units for the algorithm instead.

One example of an HPC FPGA board is the Xilinx Alveo U50. This PCIe board, which includes High-Bandwidth Memory [37], is a good candidate to accelerate memory-bound applications as time series analysis [38]. Thus, we use it in the evaluation of this work. Table 3 shows the specifications of the U50.

Table 3: Xilinx Alveo U50 Specifications

Parameter	**Alveo U50**
Look-up Tables (LUTs)	872K
Registers	1743K
DSP Slices	5952
Memory	8GB (HBM)
HBM Bandwidth	316GB/s (theoretical)
Internal SRAM	28MB
Int. SRAM Bandwidth	24TB/s
Max. Power	75W

3. TraTSA Framework

In this section we present an overview of the TraTSA framework and then describe its main components.

3.1. Overview of TraTSA

TraTSA is a Transprecision Framework for Time Series Analysis developed as a tool i) to perform design exploration of accelerators and ii) to tune current implementations. This way, computer architects can define the exact number of floating-point bits for exponent and mantissa, which potentially saves area and improves performance while reducing energy consumption. TraTSA is based on matrix profile, which is the state-of-the art method for flexible and exact time series analysis, particularly for motif and discord discovery. Concretely, we build TraTSA framework using i) the FlexFloat library to implement transprecision CPU versions of SCRIMP and SCAMP, ii) the cpfp-FPGA library to implement a transprecision FPGA implementation of SCAMP[1] and iii) Python to create a user-friendly wrapper.

[1]We do not consider implementing a transprecision FPGA version of SCRIMP since SCAMP provides better numeric stability and, as consequence, more robustness to reduced precision (see Section 5.2.1), being more amenable to transprecision approaches. However, due to the high-similar computation schemes of both algorithms, it is feasible to implement `TranSCRIMPfpga` with modest effort.

We present TraTSA's overview in Fig. 4. The *INPUT* and *OUTPUT* blocks in the figure represent TraTSA's wrapper. This wrapper is in charge of i) interpreting the configuration file (i.e. exponent and mantissa widths, selected CPU or FPGA backend, window size, among others) and obtaining the time series file provided by the user; ii) invoking the corresponding execution backend, and iii) collecting the results providing them to the user after proper formatting. TraTSA, being easily extensible to support additional algorithms, includes the following backends:

- `TranSCRIMP`. `TranSCRIMP` is a CPU C++ parallel transprecision implementation of SCRIMP algorithm. This implementation provides configurable precision arithmetic which is emulated via software using the FlexFloat library.
- `TranSCAMP`. `TranSCAMP` is a CPU C++ parallel transprecision implementation of SCAMP algorithm. This implementation provides configurable precision arithmetic which is emulated via software using the FlexFloat library.
- `TranSCAMPfpga`. `TranSCAMPfpga` is an FPGA HLS-based implementation of SCAMP algorithm. This implementation provides configurable precision arithmetic which is implemented via hardware using the cpfp-FPGA library. As we evaluate `TranSCAMPfpga` using a Xilinx Alveo U50 FPGA board, it is trivial to port it to other Alveo models.

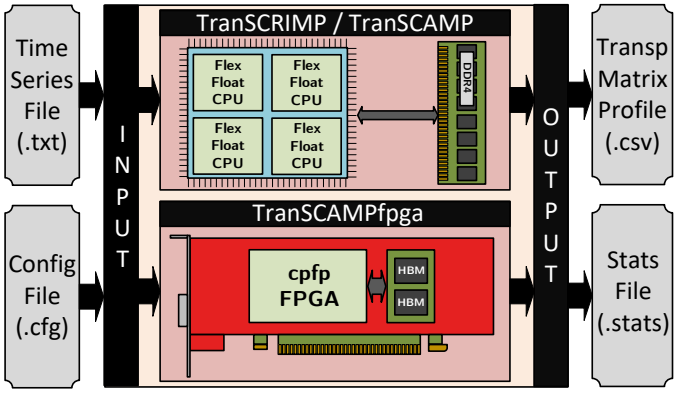

Figure 4: TraTSA overview and its components. The user provides a time series file (.txt) and a configuration file (.cfg) to the wrapper. Then, the wrapper invokes matrix profile either in the CPU or in the FPGA. Finally, the wrapper provides the user the transprecision matrix profile (.csv) and some statistics (.stats).

The key benefit of TraTSA is to provide a transprecision framework being i) portable enough to be executed in different execution environments according to the analysis requirements (i.e. length of the time series or if the user has access

to an FPGA or not), and ii) flexible enough to allow the possibility of exploring a wide range of exponent and mantissa combinations for any dataset. In this sense, both `TranSCRIMP` and `TranSCAMP` are designed to be used with time series of modest sizes (below 200K elements) and executed in commodity CPUs (desktops or high-end servers) due to the overheads of custom-precision types in those platforms. In contrast, `TranSCAMPfpga` is able to compute series of up to several million elements in manageable time thanks to the transprecision hardware support, at the cost of requiring an FPGA. Both CPU and FPGA backends can work simultaneously and join compute power.

3.2. Transprecision SCRIMP-CPU (`TranSCRIMP`)

`TranSCRIMP` is a CPU transprecision implementation of SCRIMP based on FlexFloat library. The key idea of SCRIMP [21] is to minimize the computation by exploiting the fact that the dot product can be updated incrementally for subsequences in the diagonal of the distance matrix, D (see Fig. 2). Consequently, the dot product can be expressed as follows:

$$Q_{i,j} = Q_{i-1,j-1} - T_{i-1}T_{j-1} + T_{i+m-1}T_{j+m-1} \tag{2}$$

The baseline implementation for our transprecision SCRIMP algorithm, `TranSCRIMP` in Fig. 5, is a vectorized-parallel version presented in [38]. It first precomputes the mean and standard deviation of each time series subsequence (line 1), and initializes the matrix profile array (line 3). Then, the distances between pairs of subsequences are calculated following the diagonals of the distance matrix (lines 4-26). The for loop is fully parallelized, with each thread computing a random subset of diagonals provided by their indices in the *diag* array in line 5.

For the first element of the diagonal, we need to compute the dot product of the first pair of subsequences (line 6) in parallel. The rest are updated following Eq. 2. For the proper vectorization of the dot product update, the algorithm separates the calculation of the diagonal in several steps: i) the products in Eq. 2 are calculated in parallel for *vectFact* elements of the diagonal (lines 14-15); ii) the previous dot product, q, is added to the element calculated in step i) (line 16); iii) the subsequent dot products are updated sequentially using the previous ones (lines 17-18) saving the last one in q for the next iteration of the diagonal (line 19); iv) distances are calculated in parallel (lines 20-21); and v) the profile is updated in parallel as well (lines 22-25).

The vectorization factor, *vectFact*, is given by the transprecision datatype width with respect to that of double precision (line 2). We highlight the lines of code which are to be executed using the transprecision approach. The algorithm is able to work in either one precision configuration or a mixed-precision one, thus the green (lower precision) and red (higher precision) marks.

Rather than using only double and single precision, we define a high and a low precision that can be set to any possible exponent and mantissa configuration, which provides further accuracy analysis opportunities. For the algorithms involved in this study, we store the time series codified with high precision and both the matrix profile and the precalculated statistics with low precision. For `TranSCRIMP`, Fig. 5 shows the lines of code that works with high precision highlighted in red, and those which work with low precision highlighted in green. The dot product calculations use high precision as they may require a larger numeric range. Distance calculation as well as calculations with means and standard deviations are performed with less precision.

3.3. Transprecision SCAMP-CPU (`TranSCAMP`)

`TranSCAMP` is a CPU transprecision implementation of SCAMP based on FlexFloat library. Whereas following a similar computation scheme to `TranSCRIMP`, `TranSCAMP` replaces the sliding dot product with a mean-centered-sum-of-products in order to reduce the floating-point rounding errors and the number of operations required [22]. The following equations can be precomputed in $O(n - m + 1)$ time, with $n - m + 1 = l$ being the length of P:

$$df_i = \frac{T_{i+m-1} - T_{i-1}}{2}, \quad 0 < i < l \tag{3}$$

$$dg_i = T_{i+m-1} - \mu_i + T_{i-1} - \mu_{i-1}, \quad 0 < i < l \tag{4}$$

$$ssq_i = \begin{cases} \sum_{k=0}^{m-1}(T_k - \mu_0)^2, & i = 0 \\ ssq_{i-1} + (T_{i+m-1} - \mu_i + & \\ + T_{i-1} - \mu_{i-1})(T_{i+m-1} - T_{i-1}) & 0 < i < l \end{cases} \tag{5}$$

$$\sigma_i = \sqrt{ssq_i}, \quad 0 \leq i < l \tag{6}$$

Eqs. 3 and 4 are terms used in the covariance update of Eq. 7, and the standard deviation (L2-norm of subsequence $T_{i,m} - \mu_i$) calculated in Eqs. 5 and 6 is used for the Pearson correlation coefficient depicted by Eq. 8. Notice the exclusion zone in the limits of Eq. 7 given by $\frac{m}{4}$.

$$\sigma_{i,j} = \begin{cases} \sum_{k=0}^{m-1}(T_k - \mu_0)(T_{k+j} - \mu_j), & i = 0, \frac{m}{4} < j < l \\ \sigma_{i-1,j-1} + df_i dg_j + df_j dg_i, & i > 0, \frac{m+4}{4} < j < l \end{cases} \tag{7}$$

$$P_{i,j} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j} \tag{8}$$

$$D_{i,j} = \sqrt{2m(1 - P_{i,j})} \tag{9}$$

The matrix profile can be derived incrementally for each diagonal of the distance matrix, Eq. 7, from the calculation of the covariance of two subsequences of the first row (first piece in Eq. 7). The Pearson correlation coefficient in Eq. 8 can be computed in fewer operations and it is more robust than the Euclidean Distance used by SCRIMP. Eq. 9 calculates the distance from the Pearson coefficient in $O(1)$. For `TranSCAMP`, we compute the covariance in Eq. 7 at high precision, which may have a large numeric range depending on the series. The correlation in Eq. 8, which varies between -1 and 1, is computed at low precision.

```
1:  μ, σ ← computeMeanDev(T, m);                          ▷ precalculation of statistics
2:  vectFact ← 8/sizeof(datatype);                        ▷ 8 bytes (double) by the size of the transprecision datatype
3:  P ← ∞;                                                ▷ matrix profile array initialization
4:  for idx ← tid * numDiag to (tid + 1) * numDiag − 1 do   ▷ parallel for
5:      i ← 0; j ← diag_idx;                              ▷ diag array contains the diagonals assigned to each thread
6:      q ← dotProduct(T_i,m, T_j,m);                     ▷ compute dot product for the first diagonal element (SIMD)
7:      d ← dist(m, q, μ_i, σ_i, μ_j, σ_j);               ▷ compute distance for the first element of diagonal
8:      if d < P_i then                                   ▷ update matrix profile and index arrays
9:          P_i ← d; I_i ← j;
10:     if d < P_j then                                   ▷ distance matrix is symmetric
11:         P_j ← d; I_j ← i;
12:     i ← i + 1;
13:     for j ← diag_idx to size(P) do
14:         for k ← 0 to vectFact − 1 do                  ▷ compute new dot product elements in parallel (SIMD)
15:             qs_k ← t_{i+m−1+k} t_{j+m−1+k} − t_{i−1+k} t_{j−1+k};
16:         qs_0 ← qs_0 + q                               ▷ update with the first dot product of the diagonal
17:         for k ← 1 to vectFact − 1 do                  ▷ update remaining dot products (sequentially)
18:             qs_k ← qs_k + qs_{k−1};
19:         q ← qs_{vectFact−1};                          ▷ store last dot product for next iteration
20:         for k ← 0 to vectFact − 1 do                  ▷ compute distances in parallel (SIMD)
21:             ds_k ← dist(m, qs_k, μ_{i+k}, σ_{i+k}, μ_{j+k}, σ_{j+k});
22:             if ds_k < P_{i+k} then                    ▷ update matrix profile and index arrays
23:                 P_{i+k} ← ds_k; I_{i+k} ← j + k;
24:             if ds_k < P_{j+k} then
25:                 P_{j+k} ← ds_k; I_{j+k} ← i + k;
26:         i ← i + vectFact;
```

Figure 5: Transprecision SCRIMP (`TranSCRIMP`) algorithm (transprecision operations highlighted).

3.4. Transprecision SCAMP-FPGA (`TranSCAMPfpga`)

`TranSCAMPfpga` is an FPGA transprecision implementation of SCAMP based on the cpfp-FPGA library. We include `TranSCAMPfpga` as part of TraTSA's backend to speed up the evaluation of transprecision time series analysis in those research environments where FPGAs are available. `TranSCAMPfpga` is tuned for a Xilinx Alveo U50 FPGA board, which includes High-Bandwidth-Memory (HBM) [39]. However, thanks to the C++-based HLS implementation, it is easy to port to higher-end Alveo boards or, with modest effort, to other FPGA platforms, as Intel Altera. We present an overview of `TranSCAMPfpga` in Fig. 6, including the kernels inside the two SLR regions of the FPGA and the HBM memory.

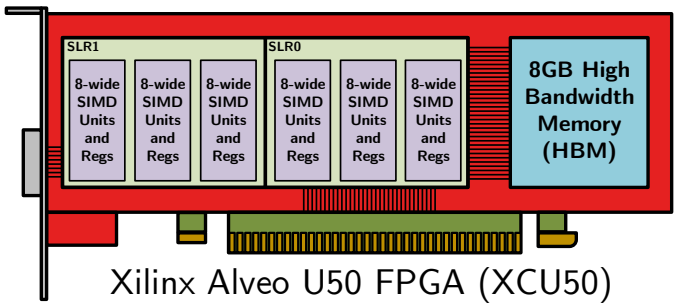

Figure 6: FPGA implementation overview. `TranSCAMPfpga` is composed of six kernels optimized for a Xilinx Alveo U50 FPGA that compute transprecision SCAMP algorithm using the data in the HBM.

We develop `TranSCAMPfpga` using Xilinx Vitis 2020.2 and a C++ HLS approach. Our implementation consists of i) a host side code, and ii) an FPGA side code (kernels), that compute matrix profile in a parallel-vectorized manner. To avoid the use of synchronization primitives, each kernel has exclusive access to its private matrix profile.

On the one hand, the host code, which is executed in the host CPU, is in charge of i) allocating the corresponding data structures in the FPGA and transferring the input vectors (i.e. time series and statistic data) from the host to the HBM cubes of the FPGA via PCI-Express, ii) defining the parallel scheduling partitioning, iii) invoking the FPGA, iv) transferring the output vectors (i.e. matrix profiles) from the HBM cubes of the FPGA to the host via PCI-Express, and v) performing the final matrix profile reduction.

On the second hand, the FPGA code is an HLS-based implementation that performs the transprecision computation of SCAMP. This FPGA implementation, intended for the evaluation of large time series (i.e. millions of elements), includes the possibility of defining a recalculation interval to reduce the ac-

cumulated errors across diagonals due to lower precision arithmetic. Concretely, instead of reusing the previous covariance to calculate the current one, it performs the complete centered-sum-of-products at certain interval. Regarding to performance, we tune this FPGA implementation to benefit as much possible from the available resources of the FPGA, exploiting the following techniques:

- **Parallelization across kernels**. Instead of having a huge kernel that computes the whole matrix profile of the given time series, we create several kernels (six in the case of the Alveo U50) which allows performing diagonal-level parallelization. Additionally, the use of six kernels efficiently exploits the HBM memory as accesses can be overlapped.

- **Vectorization inside kernels**. Instead of calculating one diagonal at a time, each kernel computes a set of diagonals (512) which allows exploiting loop unrolling techniques via arithmetic operator replication, and reusing data via *systolic arrays*. Figure 7 illustrates an example of the main advantage of vectorization using five diagonals. We notice that for the first elements of all diagonals, values of data structures need to be brought from the HBM memory. However, for the rest of the rows of all diagonals (as all diagonals increase i index at the same time), we only need to bring from memory one data value (the last one), regardless of the vectorization width. This can be achieved thanks to the systolic array approach, where previously fetched data are pushed forward to the next row while the new data value is read from memory at the same time. This operation is performed in parallel and takes only one step for all elements of the systolic array.

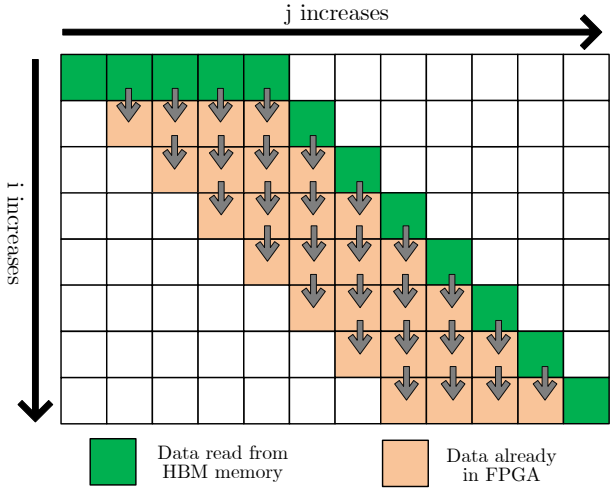

Figure 7: `TranSCAMPfpga` systolic array example.

- **Wide memory accesses**. We pack data that is requested from memory to improve the efficiency of the bus (i.e. bringing 512 bits in each request) while performing burst accesses at the same time.

We present `TranSCAMPfpga`'s execution flowchart in Fig. 8, which comprises an external loop that iterates over diagonals in batches of 512, and an inner loop to go over those diagonals following a SIMD approach. Notice that the *calculate_covariances* module is shared among two steps of the execution, which allows to save hardware resources as it is not used in all iterations of the *diag_calculation* loop.

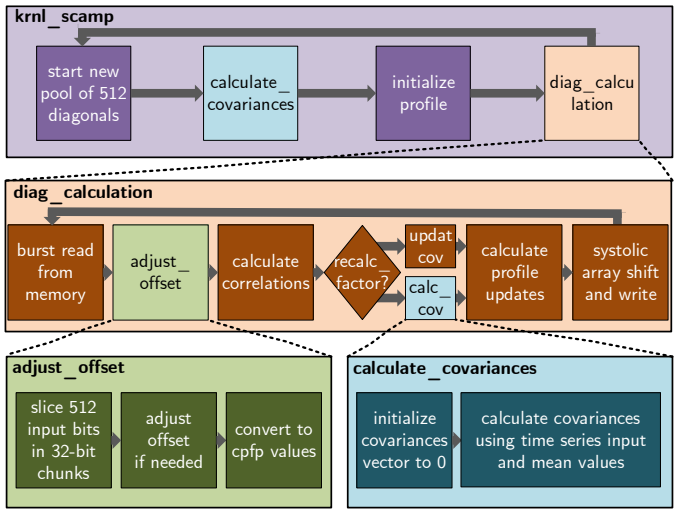

Figure 8: FPGA implementation diagram.

Finally, Table 4 shows the total resource utilization numbers by `TranSCAMPfpga` kernels in the Alveo U50 board. According to our observations using Vitis Profiler, the main bottleneck for the performance of this implementation is data movement. This means that increasing the number of arithmetic operations does not improve performance because i) global clock frequency is reduced to enable proper functionality, and ii) those additional arithmetic operations require increasing the number of memory ports, which leads to routing errors due to network congestion.

Table 4: `TranSCAMPfpga` kernel resource utilization

Param	Avail.	Used	Param	Avail.	Used
LUT	752672	308184 (41%)	URAM	640	0 (0%)
REG	1586939	494199 (31%)	BRAM	1164	714 (61%)
LUTm	389324	5802 (1.5%)	DSP	5936	828 (14%)

4. Top-K Accuracy Metric

Time series motifs [10] and discords [11] have been used for more than 15 years in the field of data mining for their capacity to find time series subsequences with special significance. In this section we define these special subsequences and propose a metric to measure the accuracy in the detection of motifs and discords from two time series.

The definitions of the motif and the Top-K motifs of a time series are presented in the remainder of this section.

Definition 1. *The motif M_1 of a time series T is the unordered pair of subsequences $\{T_{i,m}, T_{j,m}\}$ which is the most similar among all possible pairs:*

$$M_1 = \{T_{i,m}, T_{j,m}\} \Leftrightarrow dist(T_{i,m}, T_{j,m}) \leq dist(T_{u,m}, T_{v,m})$$
$$\forall i, j, u, v;\ i \neq j, u \neq v.$$

Definition 2. *The Top-K motifs $M_{1,K}$ of a time series T is the set of the first K motifs:*

$$M_{1,K} = \begin{cases} M_K \cup M_{1,K-1}, & K > 1 \\ M_1, & K = 1 \end{cases}$$

being M_K the motif (M_1) of the time series $T \setminus M_{1,K-1}, \forall K > 1$.

We can define the discord and the Top-K discords of a time series in a similar way:

Definition 3. *The discord D_1 of a time series T is the unordered pair of subsequences $\{T_{i,m}, T_{j,m}\}$ which is the most dissimilar among all possible pairs:*

$$D_1 = \{T_{i,m}, T_{j,m}\} \Leftrightarrow dist(T_{i,m}, T_{j,m}) \geq dist(T_{u,m}, T_{v,m})$$
$$\forall i, j, u, v;\ i \neq j, u \neq v.$$

Definition 4. *The Top-K discords $D_{1,K}$ of a time series T is the set of the first K discords:*

$$D_{1,K} = \begin{cases} D_K \cup D_{1,K-1}, & K > 1 \\ D_1, & K = 1 \end{cases}$$

being D_K the discord (D_1) of the series $T \setminus D_{1,K-1}, \forall K > 1$.

Being $MI_{1,K}$ and $DI_{1,K}$ the Top-K sets of unordered pair of indices $\{i, j\}$ of the unordered pair of subsequences in $M_{1,K}$ and $D_{1,K}$, respectively, we define the Top-K Motif/Discord Accuracy in the following way:

Definition 5. *The Top-K Motif (Discord) Accuracy AM (AD) of a time series T with respect to another time series TT is the number of coincidences among the unordered pairs in $MI^T_{1,K}$ and $MI^{TT}_{1,K}$ ($DI^T_{1,K}$ and $DI^{TT}_{1,K}$):*

$$AM^{T \to TT}_{1,K} = |MI^T_{1,K} \cap MI^{TT}_{1,K}|$$
$$AD^{T \to TT}_{1,K} = |DI^T_{1,K} \cap DI^{TT}_{1,K}|$$

Def. 5 is pointless when T and TT are time series from different applications. However, it can be useful when we want to know the degree of coincidence of the matrix profile of a time series and that of the same time series codified with less precision or affected by a transform operator. We can get to know if, after a transformation of the original time series, the matrix profile ends up unveiling the same motifs and discords, or a significant subset of them.

5. Experimental Evaluation

5.1. Methodology

We evaluate TraTSA's `TranSCRIMP`, `TranSCAMP` and `TranSCAMPfpga` backends in terms of accuracy, performance and energy savings. First, we compute the reference SCRIMP and SCAMP matrix profiles with double and single floating point precisions using the native C++ implementations. Second, we use the FlexFloat [28] library on an Intel Xeon Phi 7210 "Knights Landing" manycore processor [40] with 64 cores and 256 threads to evaluate `TranSCRIMP` and `TranSCAMP` on relatively short[2] time series (<200K elements). Third, we use the cpfp-FPGA library on a Xilinx Alveo U50 FPGA board to evaluate `TranSCAMPfpga` for larger time series (>200K elements). Fourth, we present performance comparisons between `TranSCAMP` and `TranSCAMPfpga` using different computing platforms. Finally, we use a transprecision FPU [34], proposed for RISC-V processors and suitable for FPGAs and ASICs, for our energy estimations, which are based on the operation breakdown statistics and the energy per operation numbers given in [34].

We use real-data time series and the Top-K Accuracy metric of Def. 5 to evaluate TraTSA. Table 5 summarizes the parameters of the time series we use for the experiments. *Song* corresponds to the song *London Bridge is Falling Down* [41] converted into Mel-frequency Cepstral Coefficients, which are commonly used in speech recognition [42]. *ECG_short* and *ECG* are extracted from an electrocardiogram signal from the European ST-T Database [43]. We select the 180000 and 1800000 first samples of the V4 electrode from ECG 0103, respectively. *Power_short* and *Power* are two time series extracted from fridge-freezer power consumption numbers collected over a whole year in a set of UK households [41]. *Seismology_short* and *Seismology* are two time series of seismic data collected by a seismograph in a geologically active region of the Long Valley Caldera, California [41]. We analyze the first 180000 and 1727990 samples of the original series, respectively. *Human Activity* comprises a time series with information of the optical flow of an actor performing activities, from picking up an object to talking on a mobile phone [41]. *Penguin Behavior* is a time series of penguin magnetometer telemetry [44]. *Speech* is an speech recorded time series where the first author of the paper reads a fragment (2 mins) of *El Quijote* book sampled at 16KHz. *IMU* is a time series extracted from the calibration of an Inertial-Measurement-Unit used in robotics [45]. Finally, *EPG* is insect electrical penetration graph data from [46].

[2]Notice that both SCRIMP and SCAMP perform profile calculations (P_i) based on a previous result along the diagonals (dot product and covariance, respectively). In such scenarios, it is possible that accumulated errors provide misleading results for large time series. However, as production implementations of SCRIMP and SCAMP are based on a tiled approach where the dot product or covariance is calculated from scratch for each 128K-512K elements of the diagonal, our evaluation and conclusions are also valid for time series of larger sizes. This is the main reason we include a recalculation factor in `TranSCAMPfpga`.

Table 5: Time series dataset parameterization

Time series	n	m	Max	Min	Scale
Song	20234	200	6.69	-56.48	1
ECG_short	180000	500	2.6	0.32	1
Power_short	180000	1325	14.0	0	0.1
Seismology_short	180000	50	6.96	-1.86	0.01
Human Activity	7997	120	2.51	-1.9	1
Penguin Behavior	109842	800	0.52	-0.21	1
Speech	1933944	16384	0.98	-1	1
ECG	1800000	512	3.39	-1.635	1
Power	1754985	1536	14.0	0	0.1
Seismology	1727990	64	23.29	-23.34	0.1
IMU	1756230	256	1.64588	-2.87	1
EPG	2000000	16384	72.26	-61.30	10

5.2. Results

5.2.1. Short Time Series Accuracy

We use the Top-K Accuracy metric of Def. 5 to evaluate our proposals. The value of K will depend on the number of significant events of a given time series, which will be eventually determined by a domain expert.

One way of setting K is defining a profile threshold for both motifs and discords. Figs. 9 and 10 present the Top-100 motif/discord accuracy with respect to double for a wide range of configurations of exponent and mantissa in `TranSCRIMP`, respectively. On the other hand, Figs. 11 and 12 present the same metrics in `TranSCAMP`, respectively. In most cases, single precision (8,23) provides 100% accuracy with respect to double precision. As can be noticed, most of the plots follow a square-like shape where the accuracy decreases dramatically after a given combination of exponent and mantissa. This may occur whether one or both of the following scenarios appear: i) the range of the exponent has been exceeded; ii) the precision provided by the mantissa is not enough for the calculations.

Comparing both algorithms, we can observe that SCAMP is more robust and presents a better numeric stability than SCRIMP for all the datasets. This fact can be noticed for *Penguin Behavior*, where a slight decrease in the length of the mantissa makes SCRIMP fail in detecting events, while SCAMP provides more margin in this reduction. We conclude that SCAMP is a better candidate for transprecision computing since it provides similar accuracy with lower bit count requirements for the exponent and mantissa.

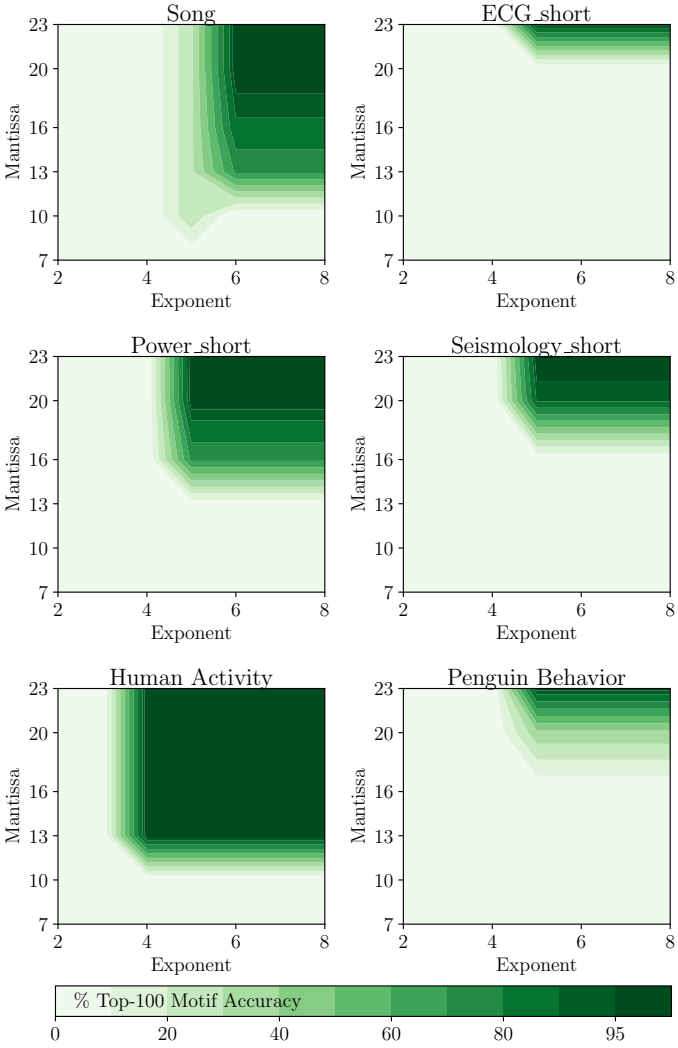

Figure 9: `TranSCRIMP` Top-100 motif accuracy with respect to double.

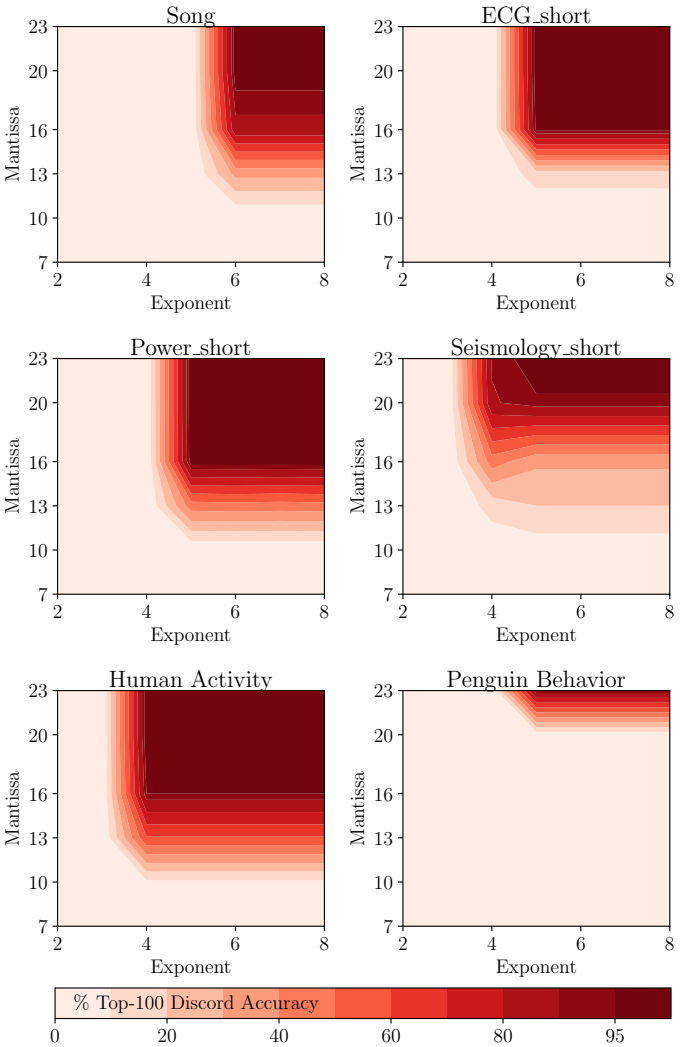

Figure 10: `TranSCRIMP` Top-100 discord accuracy with respect to double.

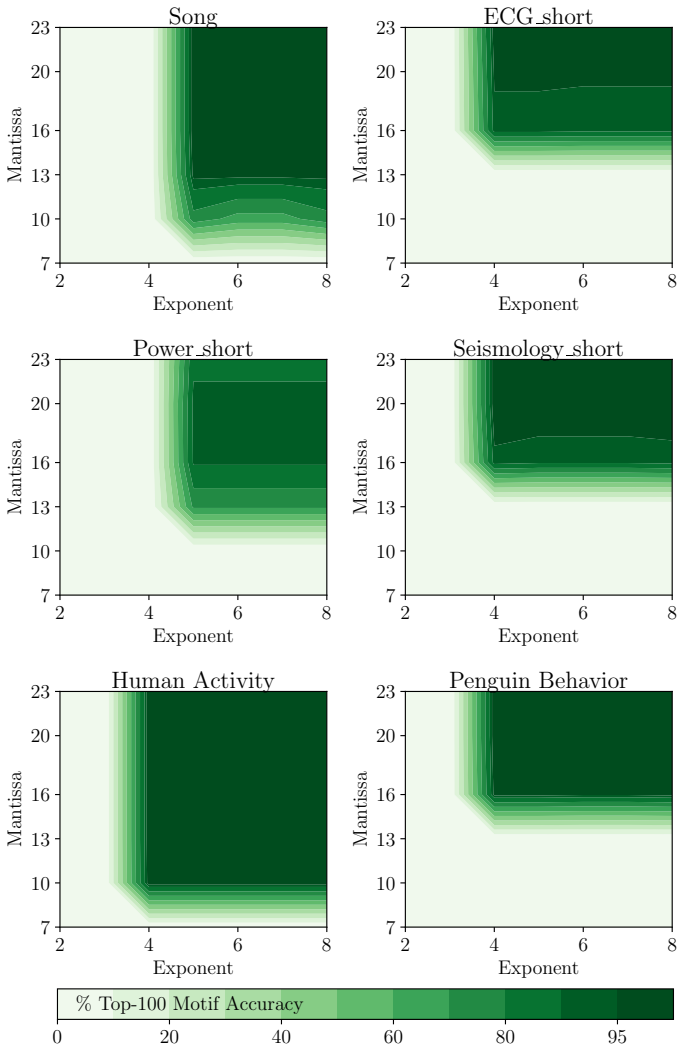

Figure 11: `TranSCAMP` Top-100 motif accuracy with respect to double.

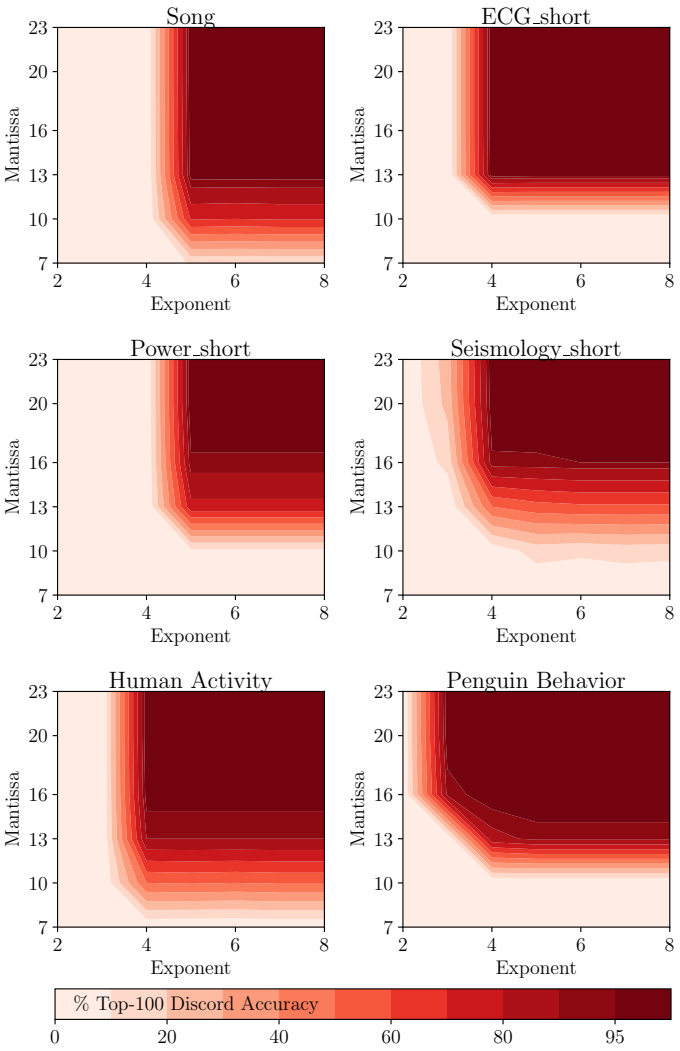

Figure 12: `TranSCAMP` Top-100 discord accuracy with respect to double.

We also evaluate how the profile shape changes when reducing the exponent/mantissa bitcount using `TranSCAMP`. Fig. 13 presents the transprecision profile for the *Song* dataset and several exponent/mantissa combinations (i.e. (7,13), (7,10), (5,13), (5,7) in the figures) with respect to a double-precision profile. We observe that for combinations of (7,13) and (7,10) the profile shape is well-preserved (i.e. both curves overlap). We need to reduce the FP bit count to a lower value (e.g. (5,7)) to find misleading results. We notice that, while in this (5,7) scenario the shape of the profile is somehow conserved with an offset, the profile index is not providing the exact matches with respect to the reference (double) solution. This fact is explained because there are potentially many values very close to each other along the profile, which can lead to high chances of profile index interchanging even if profile values slightly differ from the double-precision reference.

Fig. 14 shows the transprecision profile with respect to double for the *ECG_short* dataset. We observe that in this case the value of the mantissa plays a crucial role, since reducing it to a value lower than 13 bits leads to misleading results (i.e. the transprecision curve is not shown at all since it contains NaN values). This fact is explained because the *ECG_short* dataset is a very regular and monotonic time series, where profile values are smaller and mantissa becomes more relevant than exponent.

Fig. 15 shows the transprecision profile with respect to double for the *Power_short* dataset. We notice that the profile shape is well conserved for the (7,13), (7,10) and (5,13) exponent and mantissa combinations. However, the transprecision profile diverges from the oracle solution using lower bit counts, as shown in the (5,7) scenario. This dataset also benefits from larger mantissas since it follows a regular pattern, this explains why the (5,13) combination provides better results (i.e. the transprecision shape is better preserved) than the (7,10) one.

Fig. 16 shows the transprecision profile with respect to double for the *Seismology_short* dataset. We observe a similar behavior as in the *Song* dataset. While (7,13), (7,10) and (5,13) transprecision profiles present a well-conserved shape with respect to double, lower mantissa bit count (e.g. 7 bits) leads to misleading results and divergence.

Fig. 17 shows the transprecision profile with respect to dou-

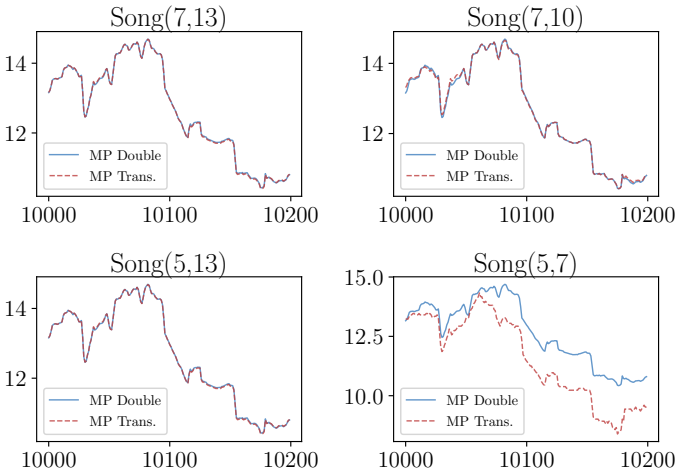

Figure 13: `TranSCAMP` *Song* profile with respect to double.

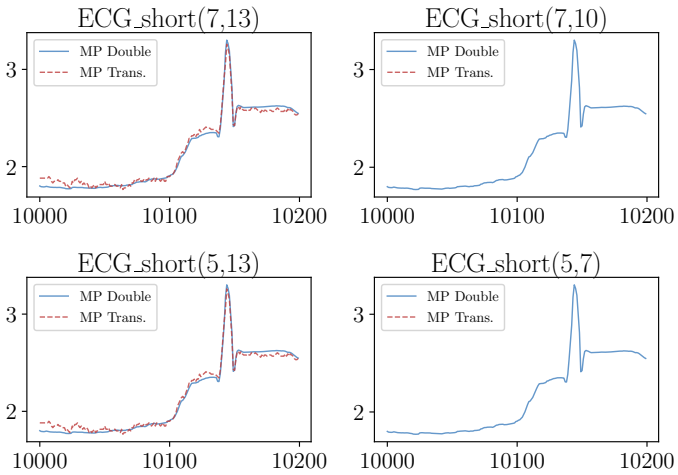

Figure 14: `TranSCAMP` *ECG_short* profile with respect to double.

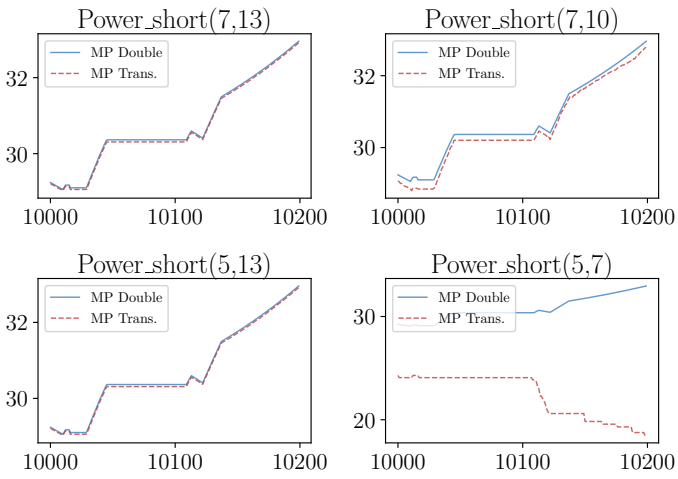

Figure 15: `TranSCAMP` *Power_short* profile with respect to double.

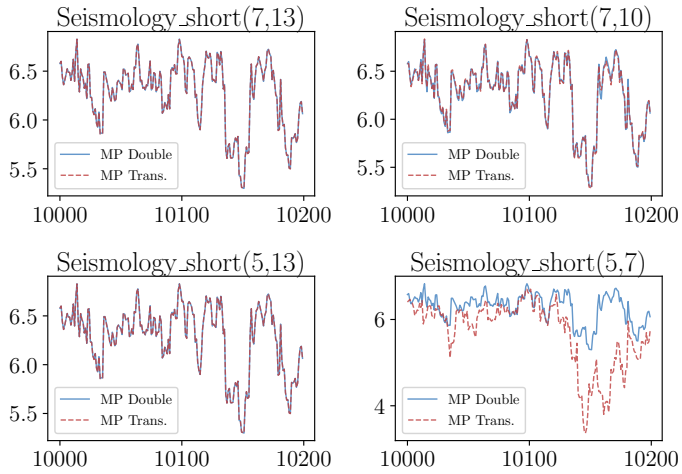

Figure 16: `TranSCAMP` *Seismology_short* profile with respect to double.

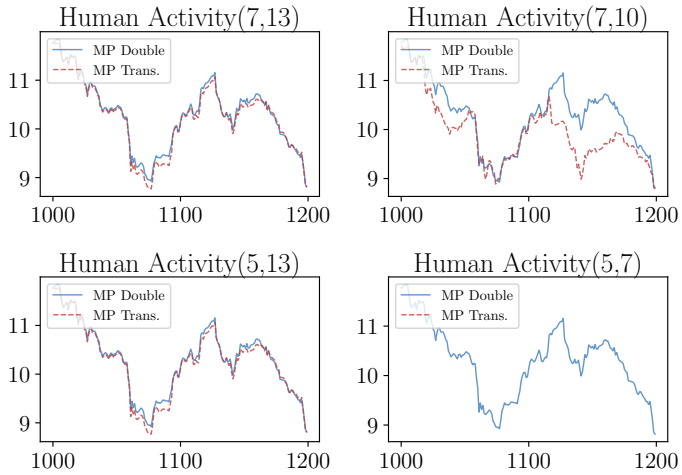

Figure 17: `TranSCAMP` *Human Activity* profile with respect to double.

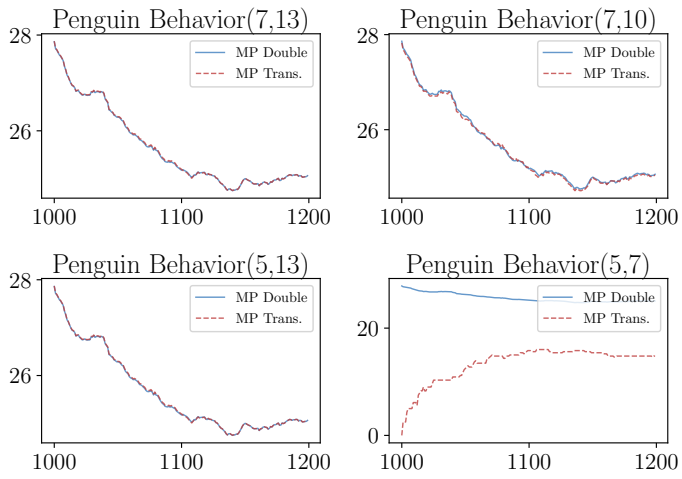

Figure 18: `TranSCAMP` *Penguin Behavior* profile with respect to double.

ble for the *Human Activity* dataset. We notice that this dataset presents a higher sensitivity to mantissa bit count, since even with 13 bits the profile does not match perfectly. However, as evaluated with the Top-K metric, the overall accuracy of the results is not affected in a higher degree than the other datasets, since *peaks* in the profile are distinguishable between them.

Fig. 18 shows the transprecision profile with respect to double for the *Penguin Behavior* dataset. This dataset presents a similar behavior as the *Power_short* one, as they have similar numeric ranges (i.e. exponent/mantissa of (7,13), (7,10) and

(5,13) provide well-conserved profile shape while (5,7) fails).

We also evaluate a mixed-precision approach for TranSCRIMP and TranSCAMP to further tune the bit counts required in different parts of those algorithms. Table 6 shows the results using mixed-precision for TranSCRIMP and TranSCAMP. We obtain better results with TranSCAMP for most datasets, as expected. It is worth noting the case of *Penguin Behavior*, where TranSCRIMP is not able to detect any of the events while TranSCAMP finds near 100% of them with the (8/23, 5/10) configuration.

Table 6: Mixed precision Top-100 accuracy results

			TranSCRIMP		TranSCAMP	
T	**High E/M**	**Low E/M**	**Accuracy Mot/Disc**	**Accu. ±10 Mot/Disc**	**Accuracy Mot/Disc**	**Accu. ±10 Mot/Disc**
Song	8/23	8/7	14/9	16/31	54/86	100/97
	”	5/10	38/0	99/0	95/99	100/100
	”	5/2	0/0	0/0	1/0	1/0
ECG	8/23	8/7	0/1	0/1	0/51	0/56
	”	5/10	10/57	10/60	25/99	30/100
	”	5/2	0/0	0/0	0/0	0/0
Power	8/23	8/7	47/31	67/95	39/25	78/99
	”	5/10	68/92	96/100	81/65	100/100
	”	5/2	0/0	0/0	0/0	0/0
Seis.	8/23	8/7	3/17	55/21	0/3	0/6
	”	5/10	7/68	86/70	12/40	15/45
	”	5/2	0/0	0/0	0/0	0/0
Hum.	8/23	8/7	72/24	80/63	91/84	99/92
	”	5/10	100/85	100/97	100/98	100/99
	”	5/2	0/3	0/4	0/0	0/1
Peng.	8/23	8/7	0/0	0/0	15/89	85/98
	”	5/10	0/0	0/0	81/99	100/99
	”	5/2	0/0	0/0	0/0	0/0

Overall, our evaluation shows that the mixed-precision configurations provide better accuracy results than using only one reduced precision configuration throughout the code. The experiments suggest that a mixed configuration of (8/23, 5/10) can be used for the majority of the applications analyzed in this work. In this sense, we can observe that SCAMP results for *Song* present 70% accuracy for a 5/10 configuration (see Figs. 11 and 12), whereas the accuracy increases to 95% using (8/23, 5/10) mixed precision (see Table 6). The rest of the series present a similar behavior with SCAMP: 0% accuracy for *ECG* with the 5/10 configuration, while up to 99% discord accuracy with the (8/23, 5/10)-mixed-precision configuration; 0% motif and 7% discord accuracy for *Power* with the 5/10 configuration, whereas 81% motif and 65% discord accuracy with mixed precision; and so on. SCRIMP presents a similar pattern as well.

We observe that the (8/23, 5/2)-mixed-precision configuration does not yield meaningful results for any of the analyzed datasets, since the low bit count for the mantissa (only two bits) does not provide enough resolution for the correlation calculations. It can be noticed that detecting discords (anomalies) is more accurate than detecting motifs (similarities). This can be due to the fact that similarities are low values of the profile which require more precision than discords, that typically are higher ones. This fact takes more relevance in very monotonic time series, where most of the subsequences are similar to each other (e.g. the beats in an electrocardiogram – *ECG*).

The number of detected events and its significance must be eventually determined by a domain expert. Thus, we can think that presenting a time series subsequence to an expert, in its context, and moved slightly to the left/right might end up with the expert coming to the same conclusion as if did not move the subsequence. For that reason we introduce the concept of Accuracy ±10 (see Table 6), which is the accuracy calculated for the motif/discord indices ranging in a ±10 interval. Using this metric the accuracy peaks 100% for most of the datasets.

Mixed precision configurations are aimed at balancing the trade-off between accuracy and performance (time/energy) of the matrix profile algorithms. We can reach high peaks of accuracy with the (8/23, 5/10) configuration while reducing time and energy consumption as described in Section 5.2.4.

5.2.2. Large Time Series Accuracy

We evaluate the accuracy of the larger time series using TranSCAMPfpga. To this end, we run the larger time series (between 1M and 2M elements) using the Alveo U50 FPGA, and evaluate the results using the Top-1000 accuracy metric (i.e. increasing K proportionally to time series length). We calculate time series statistics using double precision in the host side, which increases accuracy while it does not significantly affect the total execution time (statistics only take <1 sec using 1 thread in a Xeon Gold for 2M data points). This fact is particularly important in large time series, since statistics are calculated in an accumulative manner and single precision is not enough to avoid errors.

From the algorithmic point of view, TranSCAMPfpga differs from TranSCAMP in that i) TranSCAMPfpga includes the possibility to recalculate the covariance value from scratch after certain number of elements of a diagonal, and ii) TranSCAMP allows us to define mixed-precision configurations. The first feature resets the accumulated error due to calculation reuse across the diagonal, which increases accuracy in most datasets and allows to potentially evaluate time series of arbitrary length. We leave the mixed-precision evaluation of TranSCAMPfpga for future work.

Fig. 19 and 21 show the accuracy results obtained for motifs and discords, respectively. We observe that TranSCAMPfpga follows a similar behavior than the results obtained for its CPU counterpart implementation (TranSCAMP in Figs. 11 and 12). We make several observations here. First, analyzing the larger time series, we observe that single precision still provides almost 100% accuracy for all datasets when helped by covariance recalculation. Moreover, there is still margin to reduce the exponent and mantissa bit count below single precision depending on the application. Second, we find that the top part of the cubes are smaller in those series where the window size is larger (e.g. *Speech* or *EPG*). This fact is explained because larger window sizes have to deal with larger numbers, which may be out of the range of representation during calculations. And third, those datasets that present a randomness component (e.g. IMU) benefit from lower exponent and mantissa combinations, since the average distance is higher and there is no need for high precision to distinguish them.

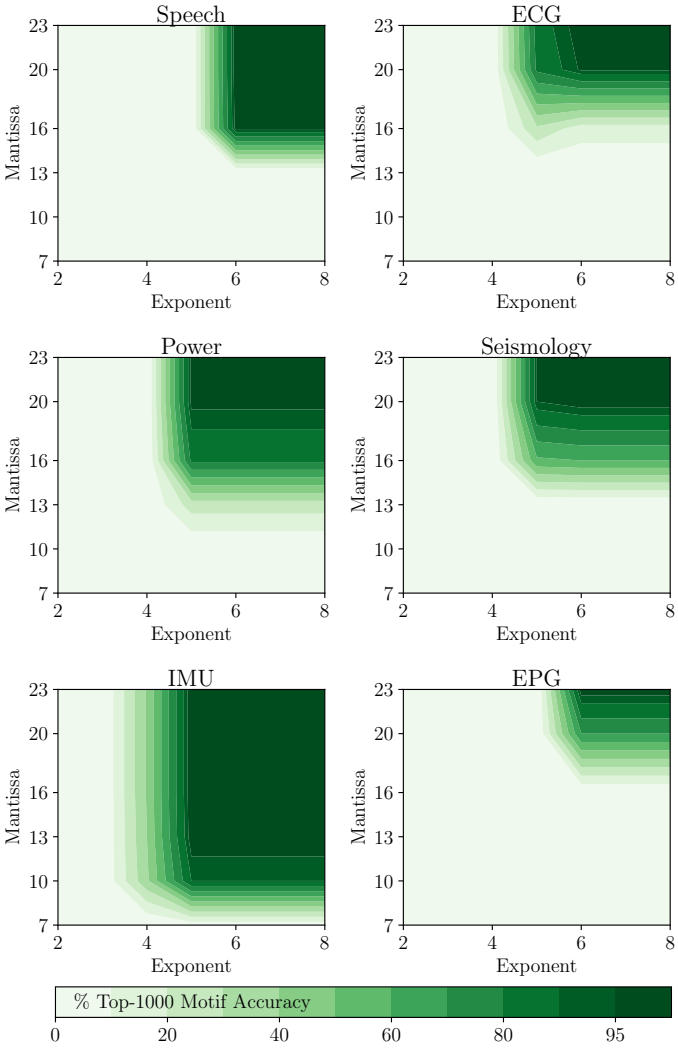

Figure 19: `TranSCAMPfpga` Top-1000 motif accuracy results with respect to double precision using a recalculation factor of 64K elements.

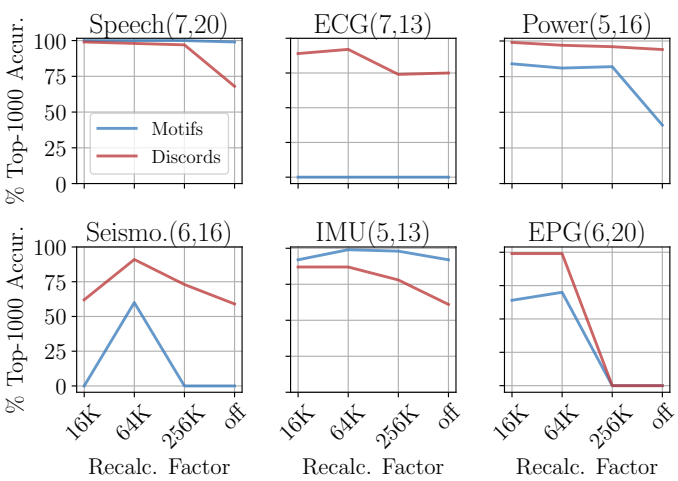

Figure 20: `TranSCAMPfpga` Top-1000 accuracy results with respect to double precision when varying the recalculation factor.

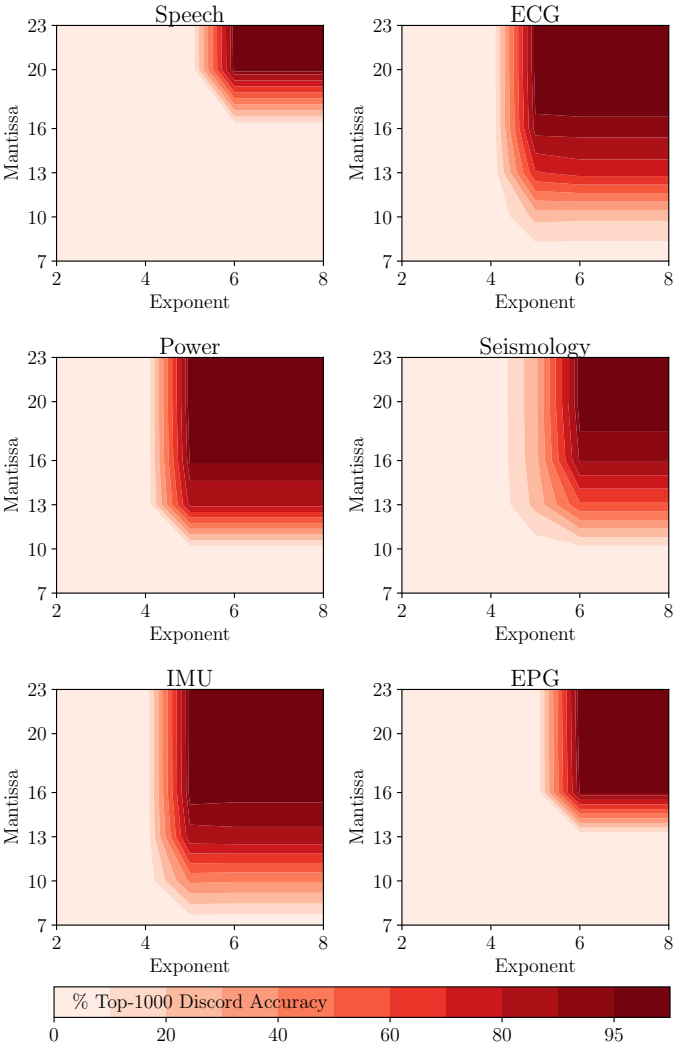

Figure 21: `TranSCAMPfpga` Top-1000 discord accuracy results with respect to double precision using a recalculation factor of 64K elements.

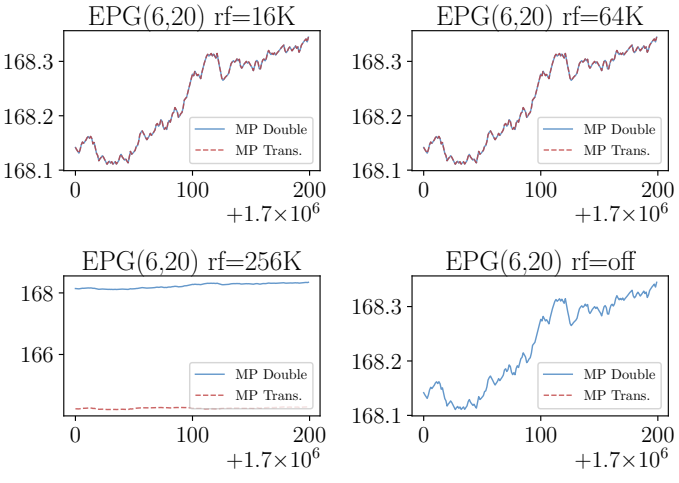

Figure 22: EPG matrix profile when varying the recalculation factor using `TranSCAMPfpga` for a given mantissa and exponent combination.

We also evaluate the effect of changing the recalculation on the accuracy. Fig. 20 shows the % of Top-1000 accuracy for the analyzed datasets with different exponent and mantissa combinations. We analyze *recalculation factors* of every 64K

and 256K elements of the diagonal against turning this feature off. We observe that 64K elements is the sweet-spot since it provides even better results than lower ones (e.g. 16K) without having great impact on performance. Lower recalculation factors imply significant impact on performance as dot product has to be calculated more frequently. In contrast, turning this feature off for series of ≈ 2M elements leads to misleading results.

An example of this fact can be noticed in Fig.22, where we present the matrix profile of the *EPG* dataset with respect to the double-precision one. Notice that while recalculation factors of 16K and 64K provide accurate results (i.e. the curves overlap), a 256K one makes the profile to always be close to 0 while turning this feature off leads to a NaN scenario.

5.2.3. TraTSA Performance

We evaluate the performance of TraTSA when running in different computing platforms. We focus our attention on comparing `TranSCAMP` versus `TranSCAMPfpga`, since `TranSCRIMP` and `TranSCAMP` present very similar execution times for a given platform.

Fig. 23 presents the execution times obtained when computing the series *Seismology_short* using a window of 512 elements and an exponent/mantissa combination of 7/10. As can be noticed, `TranSCAMPfpga` in an Alveo U50 outperforms `TranSCAMP` in commodity servers by 22.75× when using a 72-core Xeon Gold and by 52.65× when using a 64-core Intel Xeon Phi KNL. Compared to desktop computers, `TranSCAMPfpga` outperforms an Intel i7-8700 by 126× and by 313× an Intel i5-4570.

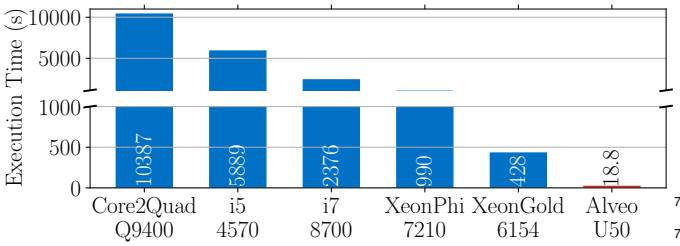

Figure 23: Execution time for different platforms when computing *Seismology_short*, using a window size of 512 elements (exp. 7 man. 10).

5.2.4. Energy Savings

First, we show an energy consumption comparison in Table 7 for the platforms in Fig. 23. As expected, the FPGA solution not only provides benefits in terms of performance but also reduced energy consumption. According to our measurements, the Alveo U50 consumes 24W when running `TranSCAMPfpga`. This is way less than the CPU evaluated platforms, where, for example, the Xeon Gold server reports a TDP of 2x200W while the Xeon Phi one reaches 215W.

Table 7: Energy consumption comparison

	Core2	i5	i7	XeonPhi	XeonGold	Alveo
Platform	Quad9400	4570	8700	7210	6154	U50
Power	95W	84W	65W	215W	2x200W	24W

Second, we present in Fig. 24 the normalized energy consumption of a) `TranSCRIMP` with respect to SCRIMP double, b) `TranSCAMP` with respect to SCAMP double and c) `TranSCAMP` with respect to `TranSCRIMP` with different precisions, respectively. As the energy is calculated with respect to the number of FPU operations performed by the algorithms, the proportion holds independently of the time series. We can observe a 60% energy reduction when using single precision instead of double for both algorithms. Furthermore, we can not only expect a reduction in time for this configuration due to an improved use of the memory hierarchy, but also because of SIMD capabilities, allowing two single precision elements computed at a time. We observe that single precision provides the same accuracy than double in the majority of cases (see Figs. 9, 10, 11, 12, 19 and 21).

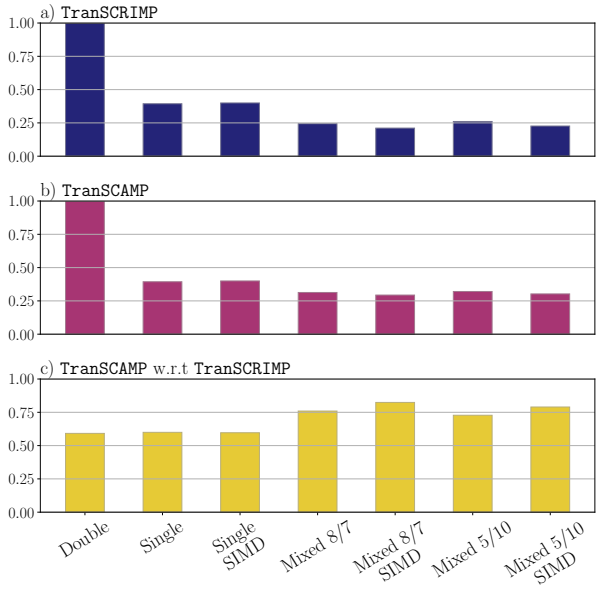

Figure 24: Normalized FPU energy using FPNew.

It is possible to reduce the energy consumption even further using mixed precision. Our `TransSCRIMP` and `TranSCAMP` mixed precision configurations can yield up to 50% and 25% energy reduction over the single precision approach respectively. The savings for `TranSCRIMP` are more pronounced since there are more operations computed in low precision. The dot product is computed in high precision but the distance in Eq. 1 is calculated in low precision. However, the distance in `TranSCAMP` is given by the Pearson correlation coefficient in Eq. 8 which entails fewer operations. We confirm this fact when comparing `TranSCAMP` with respect to `TranSCRIMP` in plot c) of Fig. 24, where `TranSCAMP` reduces the energy consumption between 18% and 40% for the same given precision. There are also a significant number of comparisons computed in low precision in both algorithms, although the energy cost of this operation is not as high as that of multiplications or sums, so the savings are not so high either. On the other hand, the SIMD support of the FPU yields roughly the same energy numbers but opens the opportunity to improve the performance even more (up to 4 operations at a time).

6. Related Work

Time series analysis. Multiple techniques for time series motif and discord discovery can be found in the literature: [47, 10, 48, 49, 50, 51, 52], including probabilistic solutions [10], spatio-temporal models [12], indexing [53], symbolic representation [54] or Euclidean distances [19]. A survey of time series motif discovery can be found in [16]. Regarding to matrix profile [19], we find multiple efforts to increase time series analysis efficiency and performance based on commodity CPU architectures, high-performance GPUs, heterogeneous systems and even Processing-Near-Memory [55] in [38, 56, 57, 22, 58].

Transprecision computing. We can find the project of this paradigm in [23], a CPU transprecision library for emulation in [28], FPGA transprecision libraries in [29, 59] and also hardware implementations [34, 60, 24, 61, 62].

Time series analysis using transprecision computing. The authors of matrix profile provide a precision evaluation in [22]. They explain how constant regions are a source of numerical instability due to zero standard deviation. Additionally, they provide a comparison between STOMP [57] and SCAMP, reporting maximum absolute error for each execution. However, this work i) does not compare SCRIMP vs SCAMP, although they compare against GPU-STOMPopt which is based on the same diagonal approach as SCRIMP but without the anytime property; ii) only considers standard IEEE floating point representations, and iii) does not discuss whether motifs and discords are conserved in a measurable manner.

To the best of our knowledge, there are no previous transprecision time series analysis works, and particularly for matrix profile. In this paper, we extend the work in [63].

7. Conclusions and Future Work

This work studies how time series analysis benefits from a transprecision approach, and introduces TraTSA, a framework that allows defining the exact needed precision according to the requirements of the specific application. The proposed `TranSCRIMP`, `TranSCAMP` and `TranSCAMPfpga` implementations will help the community to design energy-efficient time series analysis solutions based on transprecision RISC-V processors, FPGAs or ASICs while minimizing area and power requirements. Our FPGA-based solution is 22.75× faster than a 72-core Xeon server thanks to the hardware transprecision support and the use of optimization techniques. Additionaly, we study how matrix profile algorithms can benefit from an already presented transprecision FPU. In this sense, our analysis reveals that, for a variety of applications, the energy consumption of the matrix profile algorithms is reduced up to 3.3× compared with double precision while obtaining accurate results.

Future work comprises the evaluation of the transprecision analysis of time series in complete implementations of RISC-V processors and ASIC devices, analyzing how both performance and energy are affected by this approach in a complete system environment (e.g. evaluating the benefits in terms of data movement using the reduced datatypes).

References

[1] K. Bhaskaran, A. Gasparrini, S. Hajat, L. Smeeth, B. Armstrong, Time series regression studies in environmental epidemiology, International Journal of Epidemiology 42 (2013) 1187–1195.

[2] Z.-G. Yu, V. Anh, Time series model based on global structure of complete genome, Chaos, Solitons & Fractals 12 (2001) 1827–1834.

[3] M. Nerlove, D. M. Grether, J. L. Carvalho, Analysis of Economic Time Series: a Synthesis, Academic Press, 2014.

[4] N. Desai, K. Dhameliya, V. Desai, Feature extraction and classification techniques for speech recognition: a review, International Journal of Emerging Technology and Advanced Engineering 3 (2013) 367–371.

[5] L. Li, X. Su, Y. Zhang, Y. Lin, Z. Li, Trend modeling for traffic time series analysis: An integrated study, IEEE Transactions on Intelligent Transportation Systems 16 (2015) 3430–3439.

[6] S. Firth, T. Kane, REFIT: smart homes and energy demand reduction, `www.refitsmarthomes.org/index.php/data`, 2021.

[7] C. E. Yoon, O. O'Reilly, K. J. Bergen, G. C. Beroza, Earthquake detection through computationally efficient similarity search, Science Advances 1 (2015).

[8] C.-K. Peng, S. Havlin, H. E. Stanley, A. L. Goldberger, Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, Chaos: An Interdisciplinary Journal of Nonlinear Science 5 (1995) 82.

[9] R. H. Shumway, D. S. Stoffer, Time series analysis and its applications, 4 ed., Springer-Verlag, 2017.

[10] B. Chiu, E. Keogh, S. Lonardi, Probabilistic discovery of time series motifs, in: 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2003, pp. 493–498.

[11] E. Keogh, J. Lin, S.-H. Lee, H. Van Herle, Finding the most unusual time series subsequence: algorithms and applications, Knowledge and Information Systems 11 (2007) 1–27.

[12] A. McGovern, D. H. Rosendahl, R. A. Brown, K. K. Droegemeier, Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction, Data Mining and Knowledge Discovery 22 (2011) 232–258.

[13] C. Cassisi, M. Aliotta, A. Cannata, P. Montalto, D. Patanè, A. Pulvirenti, L. Spampinato, Motif discovery on seismic amplitude time series: the case study of Mt Etna 2011 eruptive activity, Pure and Applied Geophysics 170 (2013) 529–545.

[14] B. Szigeti, A. Deogade, B. Webb, Searching for motifs in the behaviour of larval Drosophila Melanogaster and Caenorhabditis Elegans reveals continuity between behavioural states, Journal of The Royal Society Interface 12 (2015).

[15] P. Garrard, V. Nemes, D. Nikolic, A. Barney, Motif discovery in speech: application to monitoring Alzheimer's disease, Current Alzheimer Research 14 (2017) 951–959.

[16] S. Torkamani, V. Lohweg, Survey on time series motif discovery, Data Mining and Knowledge Discovery 7 (2017) e1199.

[17] E. Cartwright, M. Crane, H. J. Ruskin, Financial time series: motif discovery and analysis using VALMOD, in: International Conference on Computational Science (ICCS), 2019, pp. 771–778.

[18] S. Kanarachos, S.-R. G. Christopoulos, A. Chroneos, M. E. Fitzpatrick, Detecting anomalies in time series data via a deep learning algorithm combining wavelets, neural networks and Hilbert transforms, Expert Systems with Applications 85 (2017) 292–304.

[19] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, E. Keogh, Matrix Profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets, in: IEEE 16th International Conference on Data Mining (ICDM), 2016, pp. 1317–1322.

[20] R. J. Bayardo, Y. Ma, R. Srikant, Scaling up all pairs similarity search, in: 16th International Conference on World Wide Web (WWW), 2007, pp. 131–140.

[21] Y. Zhu, C.-C. M. Yeh, Z. Zimmerman, K. Kamgar, E. Keogh, Matrix Profile XI: SCRIMP++: time series motif discovery at interactive speeds, in: 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2018.

[22] Z. Zimmerman, K. Kamgar, N. S. Senobari, B. Crites, G. Funning, P. Brisk, E. Keogh, Matrix Profile XIV: scaling time series motif discovery with GPUs to break a quintillion pairwise comparisons a day and

beyond, in: ACM Symposium on Cloud Computing (SoCC), 2019, pp. 74–86.

[23] A. C. I. Malossi, M. Schaffner, A. Molnos, L. Gammaitoni, G. Tagliavini, A. Emerson, A. Tomás, D. S. Nikolopoulos, E. Flamand, N. Wehn, The transprecision computing paradigm: concept, design, and applications, in: Design, Automation & Test in Europe Conference & Exhibition (DATE), 2018.

[24] G. Tagliavini, S. Mach, D. Rossi, A. Marongiu, L. Benini, A transprecision floating-point platform for ultra-low power computing, in: Design, Automation & Test in Europe Conference & Exhibition (DATE), 2018.

[25] FPnew Source Code, `https://github.com/pulp-platform/fpnew`, 2021.

[26] R. M. Allen, Q. Kong, R. Martin-Short, The MyShake Platform: a global vision for earthquake early warning, Pure and Applied Geophysics 177 (2020) 1699–1712.

[27] K. H. C. Li, F. A. White, T. Tipoe, T. Liu, M. C. Wong, A. Jesuthasan, A. Baranchuk, G. Tse, B. P. Yan, The current state of mobile phone apps for monitoring heart rate, heart rate variability, and atrial fibrillation: narrative review, JMIR Mhealth Uhealth 7 (2019) e11606.

[28] G. Tagliavini, A. Marongiu, L. Benini, FlexFloat: a software library for transprecision computing, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 39 (2020) 145–156.

[29] R. DiCecco, L. Sun, P. Chow, FPGA-based training of convolutional neural networks with a reduced precision floating-point library, in: International Conference on Field Programmable Technology (ICFPT), 2017.

[30] E. Vogel, G. Saravia, D. Pastén, V. Muñoz, Time-series analysis of earthquake sequences by means of information recognizer, Tectonophysics 712–713 (2017) 723–728.

[31] G. B. Moody, R. G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Engineering in Medicine and Biology Magazine 20 (2001) 45–50.

[32] A. Buttari, J. Dongarra, J. Kurzak, P. Luszczek, S. Tomov, Using mixed precision for sparse matrix computations to enhance the performance while achieving 64-bit accuracy, ACM Transactions on Mathematical Software 34 (2008).

[33] X. S. Li, J. W. Demmel, D. H. Bailey, G. Henry, Y. Hida, J. Iskandar, W. Kahan, S. Y. Kang, A. Kapur, M. C. Martin, B. J. Thompson, T. Tung, D. J. Yoo, Design, implementation and testing of extended and mixed precision BLAS, ACM Transactions on Mathematical Software 28 (2002) 152–205.

[34] S. Mach, F. Schuiki, F. Zaruba, L. Benini, A 0.80 pJ/flop, 1.24 Tflop/sW 8-to-64 bit transprecision floating-point unit for a 64 bit RISC-V processor in 22nm FD-SOI, in: IFIP/IEEE 27th International Conference on Very Large Scale Integration (VLSI-SOC), 2019.

[35] J. Rose, A. El Gamal, A. Sangiovanni-Vincentelli, Architecture of field-programmable gate arrays, Proceedings of the IEEE 81 (1993) 1013–1029.

[36] A. Kurth, P. Vogel, A. Capotondi, A. Marongiu, L. Benini, HERO: Heterogeneous embedded research platform for exploring RISC-V manycore accelerators on FPGA, arXiv 1712.06497 [cs.AR] (2017).

[37] D. U. Lee, K. W. Kim, K. W. Kim, H. Kim, J. Y. Kim, Y. J. Park, J. H. Kim, D. S. Kim, H. B. Park, J. W. Shin, J. H. Cho, K. H. Kwon, M. J. Kim, J. Lee, K. W. Park, B. Chung, S. Hong, 25.2 A 1.2v 8Gb 8-channel 128GB/s high-bandwidth memory (HBM) stacked DRAM with effective microbump I/O test methods using 29nm process and TSV, in: IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014, pp. 432–433.

[38] I. Fernandez, A. Villegas, E. Gutierrez, O. Plata, Accelerating time series motif discovery in the Intel Xeon Phi KNL processor, The Journal of Supercomputing 75 (2019) 7053–7075.

[39] H. Jun, J. Cho, K. Lee, H.-Y. Son, K. Kim, H. Jin, K. Kim, HBM (High Bandwidth Memory) DRAM technology and architecture, in: IEEE International Memory Workshop (IMW), 2017.

[40] J. Jeffers, J. Reinders, A. Sodani, Intel Xeon Phi Processor High Performance Programming: Knights Landing Edition, Morgan Kaufmann, 2016.

[41] C. M. Yeh, H. V. Herle, E. Keogh, Matrix Profile III: the Matrix Profile allows visualization of salient subsequences in massive time series, in: IEEE 16th International Conference on Data Mining (ICDM), 2016.

[42] A. V. Oppenheim, R. W. Schafer, From frequency to quefrency: a history of the cepstrum, IEEE Signal Processing Magazine 21 (2004) 95–106.

[43] A. Taddei, G. Distante, M. Emdin, P. Pisani, G. B. Moody, C. Zeelenberg, C. Marchesi, The european ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography, European Heart Journal 13 (1992) 1164–1172.

[44] Penguin Data, `https://www.cs.ucr.edu/~eamonn/MatrixProfile.html`, 2021.

[45] D. Zuñiga-Noël, A. Jaenal, R. Gomez-Ojeda, J. Gonzalez-Jimenez, The UMA-VI dataset: visual–inertial odometry in low-textured and dynamic illumination environments, The International Journal of Robotics Research 39 (2020) 1052–1060.

[46] S. Alaee, K. Kamgar, E. Keogh, Matrix Profile XXII: exact discovery of time series motifs under DTW, in: IEEE 16th International Conference on Data Mining (ICDM), 2020.

[47] P. Patel, E. Keogh, J. Lin, S. Lonardi, Mining motifs in massive time series databases, in: IEEE International Conference on Data Mining (ICDM), 2002.

[48] Y. Tanaka, K. Iwamoto, K. Uehara, Discovery of time-series motif from multi-dimensional data based on MDL principle, Machine Learning 58 (2005) 269–300.

[49] D. Yankov, E. Keogh, J. Medina, B. Chiu, V. Zordan, Detecting time series motifs under uniform scaling, in: 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Miningi (KDD), 2007, pp. 844–853.

[50] H. Tang, S. S. Liao, Discovering original motifs with different lengths from time series, Knowledge-Based Systems 21 (2008) 666–671.

[51] P. Nunthanid, V. Niennattrakul, C. A. Ratanamahatana, Discovery of variable length time series motif, in: 8th Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand - Conference, 2011.

[52] A. Balasubramanian, J. Wang, B. Prabhakaran, Discovering multidimensional motifs in physiological signals for personalized healthcare, Journal of Selected Topics in Signal Processing 10 (2016) 832–841.

[53] B. K. Yi, C. Faloutsos, Fast time sequence indexing for arbitrary $\mathcal{L}_p$ norms, in: 26th International Conference on Very Large Data Bases (VLDB), 2000, pp. 385–394.

[54] J. Lin, E. Keogh, L. Wei, S. Lonardi, Experiencing SAX: a novel symbolic representation of time series, Data Mining and Knowledge Discovery 15 (2007) 107–144.

[55] O. Mutlu, S. Ghose, J. Gómez-Luna, R. Ausavarungnirun, A modern primer on processing in memory, arXiv 2012.03112 [cs.AR] (2020).

[56] J. C. Romero, A. Vilches, A. Rodríguez, A. Navarro, R. Asenjo, ScrimpCo: scalable matrix profile on commodity heterogeneous processors, The Journal of Supercomputing 76 (2020) 9189–9210.

[57] Y. Zhu, Z. Zimmerman, N. S. Senobari, C.-C. M. Yeh, G. Funning, A. Mueen, P. Brisk, E. Keogh, Matrix Profile II: exploiting a novel algorithm and GPUs to break the one hundred million barrier for time series motifs and joins, in: IEEE 16th International Conference on Data Mining (ICDM), 2016.

[58] I. Fernandez, R. Quislant, E. Gutiérrez, O. Plata, C. Giannoula, M. Alser, J. Gómez-Luna, O. Mutlu, NATSA: A near-data processing accelerator for time series analysis, in: 38th IEEE International Conference on Computer Design (ICCD), 2020.

[59] D. B. Thomas, Templatised soft floating-point for high-level synthesis, in: IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), 2019.

[60] G. Stazi, F. Silvestri, A. Mastrandrea, M. Olivieri, F. Menichelli, Synthesis time reconfigurable floating point unit for transprecision computing, in: International Conference on Applications in Electronics Pervading Industry, Environment and Society (ApplePies), 2018.

[61] A. Borghesi, G. Tagliavini, M. Lombardi, L. Benini, M. Milano, Combining learning and optimization for transprecision computing, in: 7th ACM International Conference on Computing Frontiers (CF), 2020.

[62] C. Sudarshan, J. Lappas, C. Weis, D. M. Mathew, M. Jung, N. Wehn, A lean, low power, low latency DRAM memory controller for transprecision computing, in: 20th International Conference on Embedded Computer Systems (SAMOS), 2019.

[63] I. Fernandez, R. Quislant, E. Gutierrez, O. Plata, Energy-efficient time series analysis using transprecision computing, in: IEEE 32nd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), 2020.

Ivan Fernandez

Ivan Fernandez received his B.S. degree in Computer Engineering and his M.S. degree in Mechatronics Engineering from University of Malaga in 2017 and 2018, respectively. He is currently working toward the PhD degree at the University of Malaga. His current research interests include processing in memory, near-data processing, stacked memory architectures, high-performance computing, transprecision computing and time series analysis.

Ricardo Quislant

Ricardo Quislant received the M.Sc. degree in computer engineering from the University of Granada, and the Ph.D. from the University of Malaga, Spain, in 2006 and 2012, respectively. Currently, he is working as working as an assistant professor in the Department of Computer Architecture at the University of Malaga. His main research interests include computer memory system and high-performance computing, with special regard to transactional memory.

Sonia Gonzalez-Navarro

Sonia Gonzalez-Navarro received the B.S. and M.S. degrees in Mathematics in 2000 and the Ph.D. degree in Computer Science in 2006, both from the University of Malaga, Spain. She is currently an assistant professor in the Computer Architecture Department at the University of Malaga. Her research interests include computer arithmetic, floating point number computation, high-performance architectures for data-intensive applications and near-data processing.

Eladio Gutierrez

Eladio Gutierrez received the M.Sc. and Ph.D. degrees in Telecommunication Engineering both from the University of Malaga, Spain, in 1995 and 2001 respectively. Since 2003 he has been an associate professor in the Department of Computer Architecture of the University of Malaga. His research interests include parallel architectures and algorithms, graphics processing units and automatic parallelization.

Oscar Plata

Oscar Plata earned a M.S. in Physics in 1985 and a Ph.D. in Physics in 1989, both from the University of Santiago de Compostela, Spain. He started as Assistant Professor in the same University where he became Associated Professor in 1990. He moved to the University of Malaga in 1995, where he became Full Professor in the Computer Architecture Dept. in 2002. His research interests are related to high performance computing and parallel architectures.

Ivan Fernandez Photo

Click here to access/download;Author Photo;photo_ivan_fernandez.jpeg

Ricardo Quislant Photo

Click here to access/download;Author Photo;photo_ricardo_quislant.jpeg

Sonia Gonzalez Photo

Click here to access/download;Author Photo;photo_sonia_gonzalez.png

Eladio Gutierrez Photo

Click here to access/download;Author Photo;photo_eladio_gutierrez.jpeg

Oscar Plata Photo

Click here to access/download;Author Photo;photo_oscar_plata.png

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Conference paper