H2020-FETHPC-2014: GA 671633

NLAFET Working Note 18

An Auto-Tuning Framework for a NUMA-Aware Hessenberg Reduction Algorithm

Mahmoud Eljammaly, Lars Karlsson, and Bo Kågström

October, 2017

Document information

This preprint report is also published as Report UMINF 17.19 at the Department of Computing Science, Umeå University, Sweden.

This version: October 27, 2017.

Acknowledgements

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under the grant agreement number 671633.

An Auto-Tuning Framework for a NUMA-Aware Hessenberg Reduction Algorithm*

Mahmoud Eljammaly
Department of
Computing Science
mjammaly@cs.umu.se

Lars Karlsson
Department of
Computing Science
larsk@cs.umu.se

Bo Kågström
Department of Computing
Science and HPC2N
bokg@cs.umu.se

Abstract

The performance of a recently developed Hessenberg reduction algorithm greatly depends on the values chosen for its tunable parameters. The search space is huge combined with other complications makes the problem hard to solve effectively with generic methods and tools. We describe a modular auto-tuning framework in which the underlying optimization algorithm is easy to substitute. The framework exposes sub-problems of standard auto-tuning type for which existing generic methods can be reused. The outputs of concurrently executing sub-tuners are assembled by the framework into a solution to the original problem.

Keywords: Auto-tuning, Tuning framework, Binning, Search space decomposition, Multistage search, Hessenberg reduction, NUMA-aware.

1 Introduction

The motivation behind this work starts from the distributed parallel multishift QR algorithm [9], which is the key step in solving large dense unsymmetric eigenvalue problems. On the critical path of the distributed QR algorithm lies a costly process known as Aggressive Early Deflation (AED) [2, 3]. The purpose of AED is two-fold: to detect and deflate converged eigenvalues and to generate shifts for subsequent QR iterations. Aggressive early

*NLAFET Working Note 18. Report UMINF 17.19, Dept. Computing Science, Umeå University, SE 901 87 Umeå, Sweden.

deflation is composed of three major parts: Schur decomposition, eigenvalue reordering, and Hessenberg reduction. The AED process is currently a bottleneck in the distributed QR algorithm and we therefore aim to accelerate it in the hopes of thereby improving the performance and scalability of the QR algorithm. We recently developed a new NUMA-aware Hessenberg reduction algorithm [7] based on the Parallel Cache Assignment (PCA) technique [5, 4, 10]. The performance of the new algorithm depends greatly on the values chosen for its tunable parameters. Auto-tuning is required due to both the large number of parameters (four per iteration; see Section 2.1 ahead) and the interactions between different parameters.

In this paper, we propose a modular auto-tuning framework that helps with the tuning process. In particular, the framework tries to search the huge search space efficiently by partitioning the parameters into subsets that are tuned independently, grouping similar sub-problems into the same bin and tune them as one, and searching in multiple stages (first coarsely and then finely). The framework by itself is not a complete solution. At the heart of the framework is a generic module for optimizing a sub-problem of standard type. The framework provides a clean interface to generic optimization methods and extends them into an auto-tuner for the complex and non-standard original problem. Besides the main benefit of reducing the complex optimization problem into something more manageable, this architecture has the added benefit of making it easy to experiment with different search algorithms.

The framework works as pre- and post-processing layers around the Hessenberg algorithm. The interactions between the framework and the algorithm are as follows. The user provides to the framework an input matrix $A \in \mathbb{R}^{n \times n}$ and the number, p, of available cores. Based on n and p, the framework chooses specific values for all the algorithmic parameters of the Hessenberg algorithm. The framework then executes the Hessenberg algorithm on A with the specified parameters. The output matrices H and Q are returned to the user. Simultaneously, the Hessenberg algorithm feeds back internal time measurements to the framework for use in the tuning process.

The rest of the paper is organized as follows. Section 2 describes the details of the new implementation of blocked Hessenberg reduction. The algorithmic parameters and their interaction are identified and discussed in Section 2.1. Section 3 describes techniques used within the framework to efficiently search the huge search space. Section 4 describes the architecture of the auto-tuning framework. Section 5 shows experimental results. Finally, Section 6 sums up the paper and outlines future work.

2 NUMA-Aware Hessenberg Reduction

Hessenberg reduction is a similarity transformation that maps a matrix $A \in \mathbb{R}^{n \times n}$ to an upper Hessenberg matrix $H = Q^T AQ$, where Q is an orthogonal matrix. The current state-of-the-art algorithm [12] performs the reduction in a blocked manner. The matrix is reduced iteratively one block of columns (called a *panel*) at a time from left to right. Each panel is reduced column-by-column using Householder reflectors. The reflectors are also applied to the rest of the matrix to update it. Most of the work associated with the updates are delayed. More precisely, one iteration consists of two phases: a *(panel) reduction phase*, in which the panel is reduced, and an *(delayed) update phase*, in which the delayed updates are fully applied. Let $I - VTV^T$ denote the compact WY representation of all reflectors from one panel reduction. One iteration logically applies the similarity transformation

$$A \leftarrow (I - VTV^T)^T A(I - VTV^T) = (I - VTV^T)^T (A - YV^T), \quad (1)$$

where $Y = AVT$; see [12] for details. Our recently developed NUMA-aware parallel variant of [12] is summarized in Algorithm 1.

Figure 1 shows the shapes of A, V, and Y after the first k columns of A have been reduced. Here b refers to the width (number of columns) of the next panel.

In the reduction phase, the panel A_{22} is reduced. To reduce a column in A_{22}, the column is first updated using (1), lines 6 to 7, and then reduced by a Householder reflection, line 8. The reflection is augmented into the compact WY representation. This process affects Y_2, T, and V, lines 9 to 15.

The operations in the reduction phase are mainly matrix–vector multiplications, which therefore makes the whole phase memory-bound. The most expensive operation is a large matrix–vector multiplication involving $A_{2,2:3}$ during the computation of $\mathbf{y}$, lines 9 to 10. To perform this multiplication efficiently, our NUMA-aware algorithm [7] uses the PCA technique [5, 4, 10]. This makes for efficient utilization of the aggregate cache capacity and more localized access to main memory. Applying PCA means to (logically or physically) distribute the data over the threads/cores and then distribute the work according to the owner-computes rule. Concretely, before the start of the reduction phase, $A_{2,2:3}$ is partitioned into uniform row blocks and each block is assigned to one thread.

The NUMA-aware algorithm provides two alternative parallelization strategies for the reduction phase. In the *partial parallelization strategy*, multi-threading is used only for the most expensive multiplication (i.e., lines 9 to 10) while in the *full parallelization strategy* multi-threading is used for most of the operations.

In the update phase, Y_1 is efficiently computed directly from its definition $Y = AVT$ and A_{12}, A_{13}, and A_{23} are updated using (1), lines 16 to 19. All

operations in this phase are efficient matrix multiplications, which makes it compute-bound. All computations in the update phase are parallelized.

Algorithm 1: Parallel blocked Hessenberg reduction using PCA.

```
   // Outer loop over panels
1  foreach panel do
      // Select strategy
2     if s = full then p ← t_r else p ← 1
3     Partition A, V, and Y as in Figure 1 with panel width b
4     Assign and redistribute data to workers
      // Reduction Phase
5     foreach column a in panel A_{2,2} do
6        parfor i ← 1 : p do
7           Update column a^(i) of A_{2,2}
8        Construct a Householder reflection that reduces column a of
           A_{2,2}
9        parfor i ← 1 : t_r do
10          Compute column y^(i) of Y_2
11       parfor i ← 1 : p do
12          Compute column t^(i)
13       parfor i ← 1 : p do
14          Update column y^(i) of Y_2 using t
15       Augment Y, T, and V
      // Update Phase
16    parfor i ← 1 : t_u do
17       Update A_{2,3} from the right and left
18       Compute block Y_1 of Y
19       Update A_{1,2:3} from the right
```

2.1 Algorithmic Parameters

There are four families of tunable parameters in the NUMA-aware algorithm (see Table 1). There is one instance of each parameter *per iteration* of the algorithm, which means that there are $4N$ parameters to tune if there are N iterations. A complicating factor is that N in turn depends on the values chosen for the panel width parameters (b). Since our particular context (as a part of AED) implies that n might be relatively small compared to p, it may turn out to be sub-optimal to use all available cores, especially towards the end of the computation. The parameters t_{r} and t_{u} therefore specify the number of threads/cores ($\leq p$) to use in the reduction and update phases,

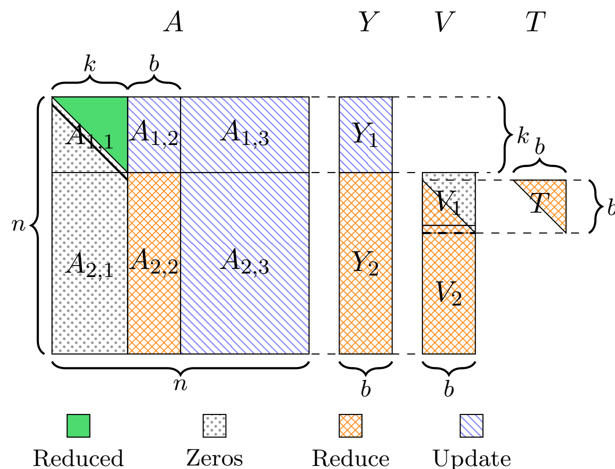

Figure 1: Partitioning of matrix A after reducing the first k columns, and Y and V will be used to reduce the panel A_{22}.

respectively.

Table 1: The four families of algorithmic parameters.

Parameter name	*Type*	*Domain*	*Affected phases*
Panel width (b)	Integer	$\{1, \ldots, n-k\}$	Both
Parallelization strategy (s)	Category	$\{\text{FULL}, \text{PARTIAL}\}$	Reduction
No. of reduction threads (t_r)	Integer	$\{1, \ldots, p\}$	Reduction
No. of update threads (t_u)	Integer	$\{1, \ldots, p\}$	Update

3 Techniques Used within the Framework

At the heart of the framework is a *search module* (see Section 4.1 ahead), which abstracts any standard auto-tuning method behind a generic interface. The main aim of the framework is to extend the very limited capability of the tuning algorithm within the search module into a complete auto-tuner for the original Hessenberg reduction problem. The framework achieves this by employing three specific techniques described in this section.

3.1 Decomposition into Independent Sub-Problems

Since the algorithm consists of an outer loop with non-overlapping iterations, it is reasonable to assume that parameters from different iterations are largely uncoupled. However, the four parameters within an iteration do strongly interact and must therefore be tuned together. This leads to the

thought of decomposing the problem of tuning all $4N$ parameters at once into tuning N independent sets of 4 parameters. Yet, since the number of iterations, N, depends on one of the parameter families (the panel width) this idea cannot be directly applied.

By analyzing Algorithm 1 and Figure 1 it becomes clear that the shape of A at the start of an iteration depends only on n and k. We (logically) associate a sub-problem with each (valid) pair (n, k). The sub-problem for (n, k) is defined as finding optimal parameter settings for the four parameters in the upcoming iteration. But optimal in what sense? Minimizing the time will not work since the panel width affects both the amount of work and the progress made. Instead the objective function (for the sub-problem) is to maximize the performance

$$P = \frac{F_{\mathrm{r}} + F_{\mathrm{u}}}{T_{\mathrm{r}} + T_{\mathrm{u}}},$$

where F_{r} and F_{u} are the flop counts for the reduction and update phases, respectively, and T_{r} and T_{u} are the corresponding wall clock times.

We collect the values of the parameters for one sub-problem into a 4-tuple referred to as a *parameter tuple*. We arrange all the N parameter tuples as columns (from left to right) of a table referred to as a *parameter table*. The objective for the auto-tuner represented by the framework is to find a parameter table that minimizes the total execution time.

3.1.1 Concurrent Solution of Several Sub-Problems.

The size of a parameter table depends on the number of iterations, which in turn depends on the chosen panel widths. For an input matrix of fixed size n, there are as many as $n-2$ possible sub-problems (n, k) for $k = 0, 1, \ldots, n-3$. Any particular parameter table therefore consists of parameter tuples extracted from some subset of the sub-problems.

One full execution of the Hessenberg algorithm makes use of N parameter tuples provided by the framework and in turn provides feedback (in the form of time measurements) that can be used by the framework to make progress on N sub-problems. In other words, the framework can concurrently solve several sub-problems. But note, however, that exactly *which subset* of the $n-2$ sub-problems are relevant for a given execution depends on the chosen panel widths. See Figure 2 for an illustration of the relationships between sub-problems, parameter tuples, and parameter tables. The framework logically keeps track of $n-2$ partially solved sub-problems and after each particular execution of the Hessenberg algorithm is able to make progress on some subset of them.

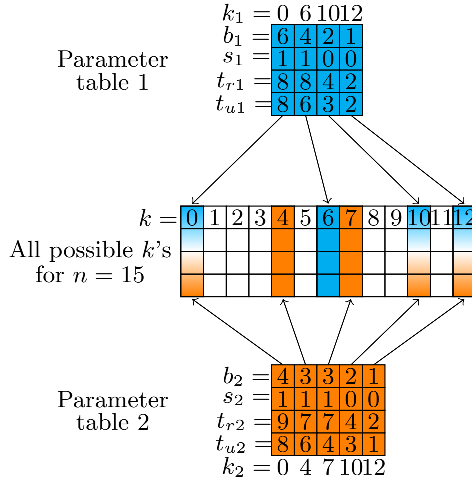

Figure 2: Two examples of parameter tables for $n = 15$.

3.2 Binning Similar Sub-Problems

Two distinct sub-problems (n, k) and (n', k') are similar if $n \approx n'$ and $k \approx k'$ simply because the shapes of all operands are similar. What this means is that we could (with some loss of accuracy) treat the two as one single sub-problem. This has several benefits. First, it reduces the total number of sub-problems that need to be solved. Second, it allows the effort invested into making progress on one sub-problem to benefit also other (similar) sub-problems.

Specifically, we group adjacent sub-problems into bins and tune each bin as if it represents a single sub-problem. The bins are rectangular of size $\Delta_n \times \Delta_k$ as illustrated by the example in Figure 3 for $\Delta_n = 2$ and $\Delta_k = 3$. In particular, the sub-problems $(10, 4)$ and $(9, 6)$ belong to the same bin $(4, 2)$.

3.3 Searching in Multiple Stages

Parameter tuples that yield good performance have a strong tendency (in this application) to cluster in one region of the search space. By performing the search in multiple stages, we can (potentially) more rapidly localize the search to this promising region. The idea is to start with a sparse but well distributed subset of the search space in the first stage of the search. Once (near-)convergence is reached, the search space is made denser and also restricted to a region around the converged point in subsequent stages.

For example, consider the two-stage search in Figure 4 which involves only t_{r} and t_{u} for simplicity. The goal is to optimize within the domain $\{1, \ldots, 10\}$. In the first stage, we choose the sparse but well distributed

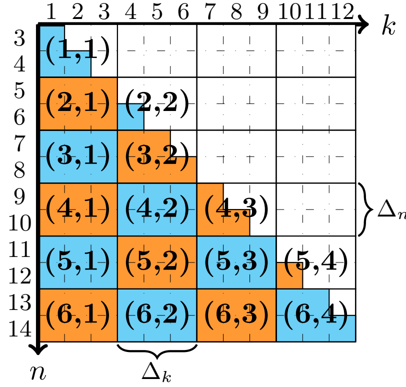

Figure 3: Binning of (12×12) space using bins of size (2×3).

sub-domain $\{1, 4, \ldots, 10\}$ (large green dots). Suppose the search in the first stage converges to the point $(t_\mathrm{r}, t_\mathrm{u}) = (4, 7)$ (red cross). Then we include more points and restrict the search in the second stage to the sub-domain $\{2, \ldots, 6\}$ for t_r and $\{5, \ldots, 9\}$ for t_u.

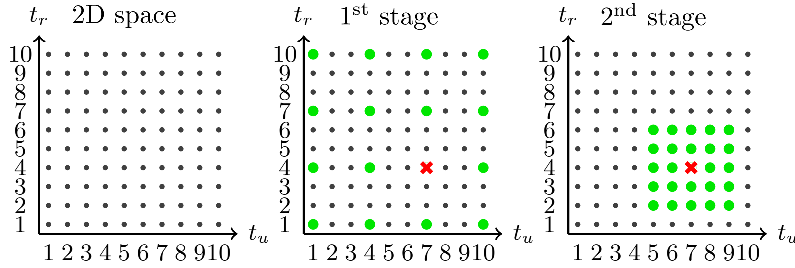

Figure 4: Two-stage search for 2D parameter space.

4 The Framework's Architecture

This section describes the software architecture of the framework. There are three modules: the Search Module, the Management Module, and the Database Module (see Figure 5).

4.1 The Search Module

The purpose of the Search Module is to encapsulate some standard autotuning method behind an abstract interface. The framework does not provide any implementation of this module by itself.

The Search Module has two primary functions: choose a parameter tuple for a given sub-problem and advance the search for a given sub-problem

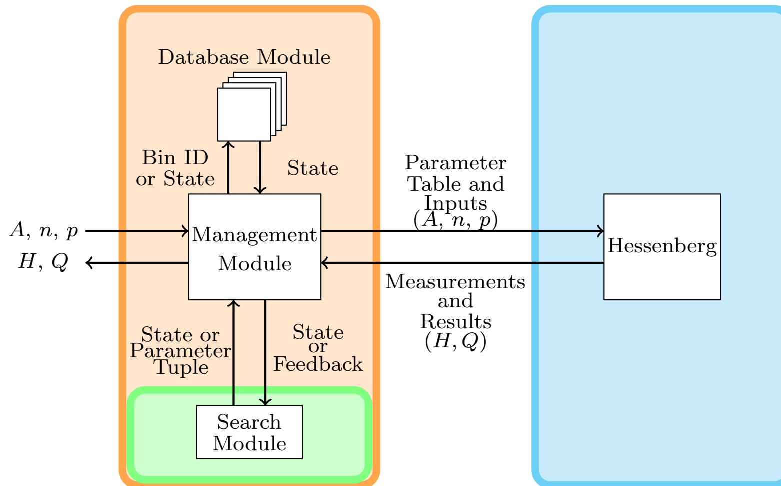

Figure 5: Modular diagram of the auto-tuning framework and the interface with the new Hessenberg reduction algorithm.

by one step in response to feedback. The module implementation itself is supposed to be state-less. A search state is instead encapsulated by the implementation into an opaque object[1] that is externally managed by the framework (see Sections 4.2 and 4.3 ahead). Since the specifics of what constitutes a "search state" depends entirely on the implementation of the Search Module, the framework views these objects as binary blobs[2] with no structure.

The Search Module exposes the following interface:

- CREATE–STATE: Creates a new state.
- SELECT–PARAMETERS: Chooses a parameter tuple for the next iteration.
- RECEIVE–FEEDBACK: Receives feedback from the previous execution.
- UPDATE–STATE: Performs one search step using previous feedback.
- CHECK–CONVERGENCE: Check if the search has converged.

4.2 The Management Module

The Management Module provides the glue that binds all the other modules together with the user input/output and the Hessenberg algorithm.

The core functions of the Management Module are as follows:

[1] An object whose content and structure are not concretely known.
[2] Collection of data stored in binary as a single entry.

- Construct the next parameter table to use.
- Run the Hessenberg algorithm with the chosen parameter table.
- Feed back measurements to the active sub-problems.

Construct parameter table. Starting from $k = 0$ and repeatedly calling the SELECT–PARAMETERS function of the Search Module (and updating $k \leftarrow k + b$ in between), a complete parameter table can be constructed column by column from left to right. The search state to use is either fetched from the Database Module or initialized using CREATE–STATE. Binning is applied before looking up a search state. The process of constructing a parameter table also implicitly selects the subset of *active sub-problems*, i.e., sub-problems which are going to be used in the next execution. So *before* calling SELECT–PARAMETERS, the function UPDATE–STATE is called on to make one step in the optimization algorithm (except initially when there is no feedback available). Furthermore, if the CHECK–CONVERGENCE function signals convergence, then the state is re-initialized with the search space used in the next stage of the multi-stage search.

Run the Hessenberg algorithm. The parameter table is passed alongside the other inputs to the Hessenberg algorithm. The computed matrices are output to the user.

Feed back measurements. The internal time measurements from each iteration are fed back to the active sub-problem search states using the RECEIVE–FEEDBACK function. The active states are kept in the Management Module until the measurements are fed back and afterwards they sent to the Database Module.

4.3 The Database Module

The Database Module stores the binary blobs representing the opaque search states. The search states are indexed by the bin coordinates (bin ID).

5 Experimental Results

The framework by itself cannot be meaningfully tested since it is dependent on an implementation of the Search Module. So in order to test the framework we implemented the Search Module using the *Nelder-Mead* algorithm [11]. This is neither the best nor the worst choice of algorithm. Ultimately the choice is not so important since the aim of this section is to show that the framework is able to make gradual improvements of the overall performance even though the actual optimization is only performed on

small sub-problems. What the most effective implementation of the Search Module looks like is an open problem and something we do not contemplate in this paper.

In the experiments we used bins of size 10×10 and multi-stage search spaces as defined by Table 2, where b', t'_r, t'_u refer to the best values found in the first stage. Figure 6 shows the execution times (dots) of 500 executions for a matrix of order $n = 1000$. The curve shows a moving average of 50 consecutive measurements. The results indicate that in general the performance is indeed improving over time.

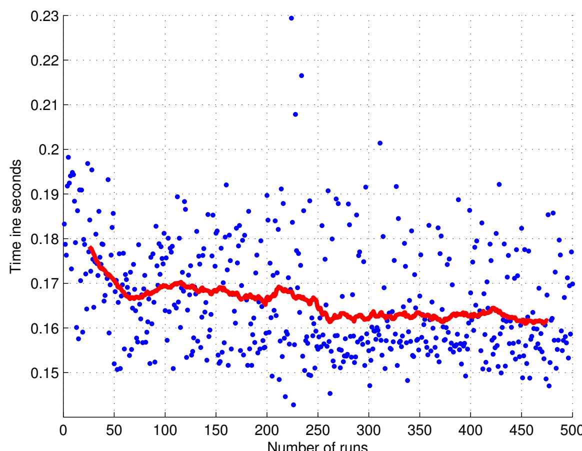

Figure 6: Execution time of 500 runs of the new Hessenberg reduction algorithm for $n = 1000$ using the framework. The red curve represents the moving average for a window of size 50.

Table 2: The search spaces used in multi-stage search.

Parameter symbol	*1st stage domain*	*2nd stage domain*
b	$10 : 10 : 100$	$b' - 9 : b' + 10$
s	$\{\text{FULL}, \text{PARTIAL}\}$	$\{\text{FULL}, \text{PARTIAL}\}$
t_r	$6 : 6 : 48$	$t'_r - 5 : t'_r + 6$
t_u	$6 : 6 : 48$	$t'_u - 5 : t'_u + 6$

6 Summary

In this paper we propose a modular auto-tuning framework that helps with tuning the parameters of a recently developed Hessenberg reduction algorithm. A brief description of the new algorithm and its parameters are

presented. The algorithm's parameters interact with each other and span a huge search space which makes using generic tuning methods and tools like [6, 1, 8] not directly applicable. Such tools, despite been successfully used in solving other problems, can not deal with a problem which has a variable number of tunable parameters (like the one we have). Specially when this number depends on the value chosen for some of the tuned parameters them selves.

In contrast, the proposed framework facilitate that. The framework applies several techniques which allow searching the huge search space efficiently. Specifically, the framework decomposes the search space into smaller subspaces revealing standard auto-tuning sub-problems which can be tuned independently and concurrently. In addition, the framework groups similar sub-problems together in a single bin and tune them as one problem, which reduces the total number of sub-problems that need to be tuned, and propagate the progress made in tuning one sub-problem to other similar sub-problems. The framework also applies a multi-stage search, which, in one stage, allows for fast discovery of a promising region, where, in a later stage, the search is localized.

Besides solving the complex problem of the huge search space, the framework defines an abstract module with clear interface which can encapsulate any standard optimization methods or generic tuning tools, including [6, 1, 8], to expand its capabilities. This abstract module allows the experimentation with different tuning algorithms.

For testing the framework's ability to improve the overall performance of the new Hessenberg reduction algorithm, we used the Nelder-Mead algorithm in the search module. The results show that the performance of the new Hessenberg reduction algorithm is gradually improving over time.

Future work includes experimenting with both generic and specialized tuning algorithms in the search module and apply the idea underlying the framework to other linear algebra algorithms besides Hessenberg reduction.

Acknowledgements.

We thank the High Performance Computing Center North (HPC2N) at Umeå University for providing computational resources and valuable support during test and performance runs. Financial support has been received from the European Unions Horizon 2020 research and innovation programme under the NLAFET grant agreement No 671633, and by eSSENCE, a strategic collaborative e-Science programme funded by the Swedish Government via VR.

References

[1] Jason Ansel, Shoaib Kamil, Kalyan Veeramachaneni, Jonathan Ragan-Kelley, Jeffrey Bosboom, Una-May O'Reilly, and Saman Amarasinghe. Opentuner: An extensible framework for program autotuning. In *International Conference on Parallel Architectures and Compilation Techniques*, Edmonton, Canada, August 2014.

[2] K. Braman, R. Byers, and R. Mathias. The multishift QR algorithm. part I: Maintaining well-focused shifts and level 3 performance. *SIMAX*, 23(4):929–947, 2002.

[3] K. Braman, R. Byers, and R. Mathias. The multishift QR algorithm. Part II: Aggressive early deflation. *SIMAX*, 23(4):948–973, 2002.

[4] A. M. Castaldo and R. C. Whaley. Achieving scalable parallelization for the Hessenberg factorization. In *Cluster Computing (CLUSTER), 2011 IEEE International Conference on*, pages 65–73. IEEE, 2011.

[5] A. M. Castaldo, R. C. Whaley, and S. Samuel. Scaling LAPACK panel operations using parallel cache assignment. *ACM Transactions on Mathematical Software*, 39(4), 2013.

[6] Cristian Ţăpuş, I-Hsin Chung, and Jeffrey K. Hollingsworth. Active harmony: Towards automated performance tuning. In *Proceedings of the 2002 ACM/IEEE Conference on Supercomputing*, SC '02, pages 1–11. IEEE Computer Society Press, 2002.

[7] M. Eljammaly, L. Karlsson, and B. Kågström. Evaluation of the Tunability of a New NUMA-Aware Hessenberg Reduction Algorithm. *NLAFET Working Note 8*, December, 2016. Also as Report UMINF 16.22, Dept. of Computing Science, Umeå University, SE-901 87 Umeå, Sweden.

[8] Michael Gerndt, Siegfried Benkner, Eduardo César, Carmen Navarrete, Enes Bajrovic, Jiri Dokulil, Carla Guillén, Robert Mijakovic, and Anna Sikora. A multi-aspect online tuning framework for hpc applications. *Software Quality Journal*, May 2017.

[9] Robert Granat, Bo Kågström, Daniel Kressner, and Meiyue Shao. Algorithm 953: Parallel library software for the multishift QR algorithm with aggressive early deflation. *ACM Trans. Math. Softw.*, 41(4):29:1–29:23, October 2015.

[10] M. R. Hasan and R. C. Whaley. Effectively exploiting parallel scale for all problem sizes in LU factorization. In *Parallel and Distributed Processing Symposium, 2014 IEEE 28th International*, pages 1039–1048. IEEE, 2014.

[11] J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[12] G. Quintana-Ortí and R. van de Geijn. Improving the performance of reduction to Hessenberg form. *ACM Transactions on Mathematical Software*, 32(2):180–194, 2006.