ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

MATHICSE

Mathematics Institute of Computational Science and Engineering

School of Basic Sciences - Section of Mathematics

MATHICSE Technical Report

Nr. 31.2013

October 2013 (NEW 03.06.2014)

Parallel library software for the multishift QR algorithm with aggressive early deflation

R. Granat, B. Kagström, D. Kressner, M. Shao

http://mathicse.epfl.ch

Phone: +41 21 69 37648

Address:
EPFL - SB - MATHICSE (Bâtiment MA)
Station 8 - CH-1015 - Lausanne - Switzerland

Fax: +41 21 69 32545

Parallel Library Software for the Multishift QR Algorithm with Aggressive Early Deflation*

Robert Granat†, Bo Kågström†, Daniel Kressner‡, and Meiyue Shao†,‡

June 3, 2014

Abstract

Library software implementing a parallel small-bulge multishift QR algorithm with aggressive early deflation (AED) targeting distributed memory high-performance computing systems is presented. Starting from recent developments of the parallel multishift QR algorithm [Granat et al., SIAM J. Sci. Comput. 32(4), 2010], we describe a number of algorithmic and implementation improvements. These include communication avoiding algorithms via data redistribution and a refined strategy for balancing between multishift QR sweeps and AED. Guidelines concerning several important tunable algorithmic parameters are provided. As a result of these improvements, a computational bottleneck within AED has been removed in the parallel multishift QR algorithm. A performance model is established to explain the scalability behavior of the new parallel multishift QR algorithm. Numerous computational experiments confirm that our new implementation significantly outperforms previous parallel implementations of the QR algorithm.

Keywords: Multishift QR algorithm, aggressive early deflation, parallel algorithms, distributed memory architectures

1 Introduction

The QR algorithm is the method of choice for computing all eigenvalues of a nonsymmetric matrix $A \in \mathbb{R}^{n\times n}$. This paper describes a novel parallel implementation of the multishift QR algorithm for distributed memory architectures. While our implementation is largely based on the algorithms described in [21], a number of additional algorithmic improvements have been made, leading to significantly reduced execution times and higher robustness.

In the following, we give a brief history of serial and parallel implementations of the QR algorithm. The algol procedure `hqr` by Martin, Petersen, and Wilkinson [32] was among the first computer implementations of the QR algorithm. A Fortran translation of this procedure was included in EISPACK [35] as routine `HQR`. The initial version of the LAPACK routine `DHSEQR` was based on work by Bai and Demmel [5]; the most notable difference to `HQR` was the use of multishift techniques to improve data locality. This routine had only seen a few minor modifications [2] until LAPACK version 3.1, when it was replaced by an implementation incorporating pipelined bulges and aggressive early deflation techniques from the works by Braman, Byers, and Mathias [12, 13]. This implementation is described in more detail in [14]. While there has been a lot of early work on parallelizing the QR algorithm, for example in [23, 33, 37, 38, 39, 40], the first publicly available parallel implementation was released only 1997 in ScaLAPACK [11] version 1.5 as routine `PDLAHQR`, based on work by Henry, Watkins, and Dongarra [24]. A complex version `PZLAHQR` of this routine was included later on [16]. In this work, we describe

*Report UMINF-12.06.The work is supported by the Swedish Research Council under grant A0581501, UMIT Research Lab via an EU Mål 2 project, and eSSENCE, a strategic collaborative e-Science programme funded by the Swedish Research Council.

†Department of Computing Science and HPC2N, Umeå University, SE-901 87 Umeå, Sweden. Email: {granat, bokg, myshao}@cs.umu.se.

‡MATHICSE, EPF Lausanne, CH-1015 Lausanne, Switzerland. Email: daniel.kressner@epfl.ch.

a new parallel implementation of the QR algorithm that aims to replace PDLAHQR. It might be interesting to note that all recently released high-performance linear algebra packages, such as MAGMA and PLASMA [1], ELPA [3], FLAME [10] lack adapted implementations of the QR algorithm or other nonsymmetric eigenvalue solvers.

Given a *nonsymmetric* matrix A, the parallel implementation of the eigenvalue solver in ScaLAPACK consists of the following steps. In the first optional step, the matrix is balanced, that is, an invertible diagonal matrix D is computed to make the rows and columns of $A \leftarrow D^{-1}AD$ as close as possible. In the second step, A is reduced to Hessenberg form: $H = Q_0^T A Q_0$ with an orthogonal matrix Q_0 and $h_{ij} = 0$ for $i \geq j + 2$. In the third step, the QR algorithm iteratively reduces H further to real Schur form, eventually resulting in an orthogonal matrix Q such that

$$T = Q^T H Q \tag{1}$$

is quasi-upper triangular. This means that T is block upper triangular with 1×1 blocks (corresponding to real eigenvalues) and 2×2 blocks (corresponding to complex conjugate eigenvalue pairs) on the diagonal. Therefore, the Schur decomposition of A is $A = ZTZ^T$, where $Z = Q_0 Q$. The last optional step consists of computing the eigenvectors of T and performing a back transformation to obtain the eigenvectors of the original matrix A. This paper is only concerned with the reduction to real Schur form (1). In particular, we will not discuss the implementation of Hessenberg reduction, see [31, 29, 36] for recent developments in this direction.

The rest of this paper is organized as follows. Section 2 provides a summary of our implementation and the underlying algorithm, emphasizing improvements over [21]. In Section 3, we present a performance model that provides insights into the cost of computations and communication. The derivation of this model is given in the electronic appendix. The performance model is then used to guide the choice of the parameters in Section 4. Finally, in Section 5, we evaluate the performance of our parallel library software by a large set of numerical experiments.

2 Algorithms and Implementation

Modern variants of the QR algorithm usually consist of two major components—multishift QR sweeps and aggressive early deflation (AED). A typical structure of the modern QR algorithm in a sequential or parallel setting is provided in Algorithm 1. Although our parallel multishift QR algorithm also builds on Algorithm 1, there are several issues to be considered for reaching high performance. Figure 1 shows the software hierarchy of our implementation of the parallel multishift QR algorithm. Details of the algorithm and some implementation issues are discussed in the successive subsections.

Algorithm 1 Multishift QR algorithm with AED

Input: $H \in \mathbb{R}^{n \times n}$, H is upper Hessenberg.
Output: A real Schur form of H.

1: **while** not converged **do**
2: Perform AED on the $n_{\text{AED}} \times n_{\text{AED}}$ trailing principle submatrix.
3: Apply the accumulated orthogonal transformation to the corresponding off-diagonal blocks.
4: **if** a large fraction of eigenvalues has been deflated in Step 2 **then**
5: **goto** Step 2.
6: **end if**
7: Perform a small-bulge multishift QR sweep with n_{shift} undeflatable eigenvalues obtained from Step 2 as shifts.
8: Check for negligible subdiagonal elements.
9: **end while**

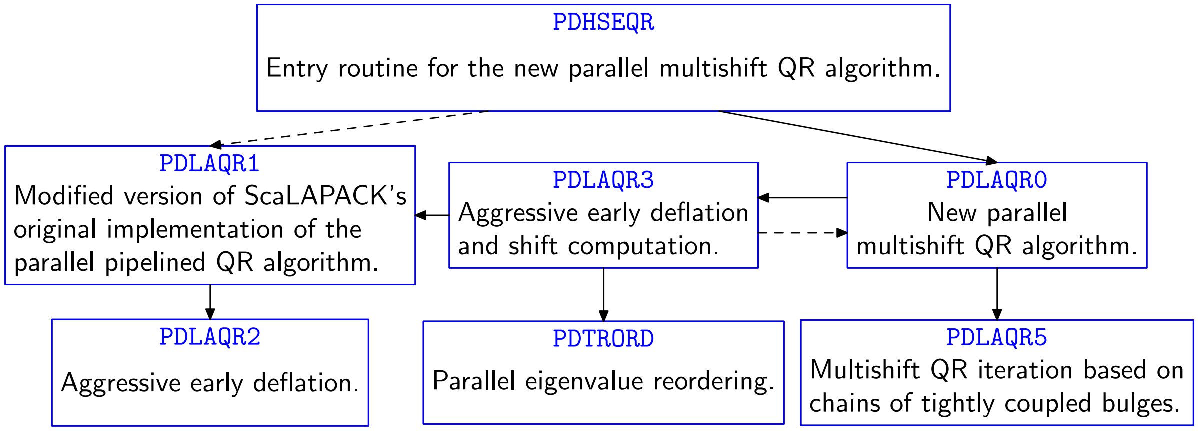


Figure 1: Software hierarchy of the parallel multishift QR algorithm with AED

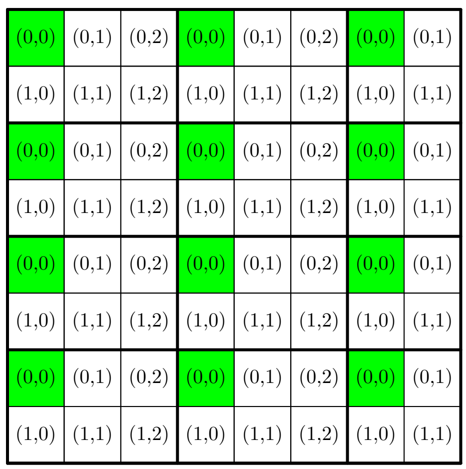


Figure 2: The 2D block-cyclic data layout across a 2×3 processor grid. For example, processor $(0, 0)$ owns all highlighted blocks.

2.1 Data layout convention in ScaLAPACK

In ScaLAPACK, the $p = p_r p_c$ processors are usually arranged into a $p_r \times p_c$ grid. Matrices are distributed over the rectangular processor grid in a *2D block-cyclic layout* with block size $m_b \times n_b$ (see an example in Figure 2). The information regarding the data layout is stored in an *array descriptor* so that the mapping between entries of the global matrix and their corresponding locations in the memory hierarchy can be established. We adopt ScaLAPACK's data layout convention and require that the $n \times n$ input matrices H and Z have identical data layout with square data blocks (i.e., $m_b = n_b$). However, the processor grid need not to be square unless explicitly specified.

2.2 Multishift QR sweep

The multishift QR sweep is a bulge chasing process that involves several shifts. The QR sweep applied to a Hessenberg matrix H with k shifts $\sigma_1, \sigma_2, \ldots, \sigma_k$ yields another Hessenberg matrix $Q^T HQ$ where Q is determined by the QR decomposition of the shift polynomial:

$$(H - \sigma_1 I)(H - \sigma_2 I) \cdots (H - \sigma_k I) = QR.$$

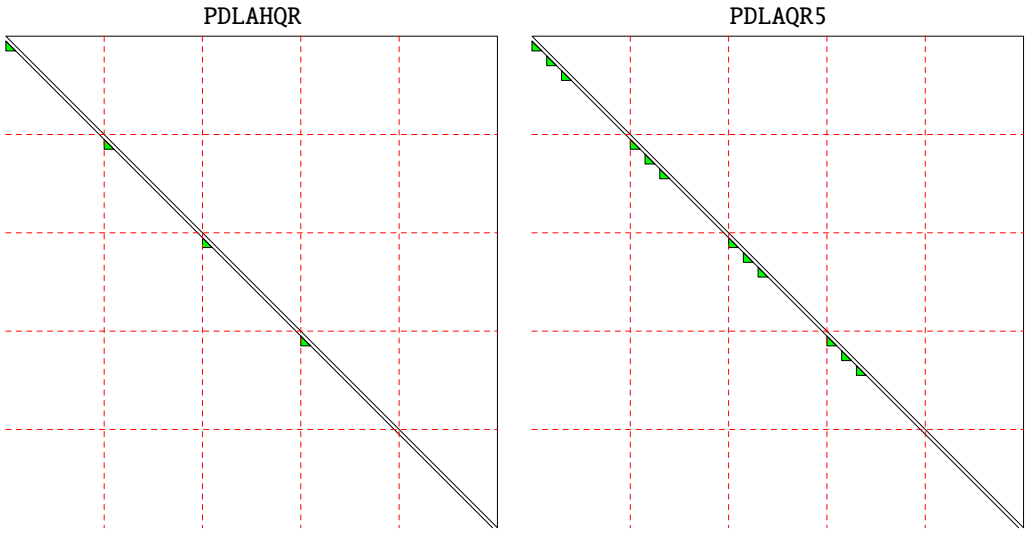


Figure 3: Loosely coupled shifts v.s. tightly coupled shifts. The dashed lines represent borders of the processor grid. Only parts of the matrix are displayed. Intrablock bulge chasings can be performed independently and in parallel.

Thanks to the *Implicit Q theorem* (e.g., see [17, 18]), there is a lot of freedom regarding how to perform the QR sweep. In ScaLAPACK v1.8.0 and earlier versions, `PDLAHQR` uses a pipelined approach that chases a chain of loosely coupled bulges, see Figure 3(a). While this approach clearly gives potential for parallelism, it comes with the disadvantage that its computational intensity is concentrated at level 1 and level 2 BLAS. In addition, frequent communication between processors is required. To avoid these shortcomings of the pipelined QR algorithm, we use several chains of tightly coupled bulges, see Figure 3(b), to improve node performance and reduce communication. For the new bulge chasing routine `PDLAQR5`, the total number of shifts (n_{shift}) used in a single QR sweep is shown in Table 1 and is usually much larger compared to the pipelined approach. These shifts are divided into several chains of tightly coupled bulges with up to $\lfloor n_b/3 \rfloor$ shifts per chain so that the length of each chain does not exceed $n_b/2$. The chains are placed on different diagonal blocks, such that $\Theta(\sqrt{p})$ chains can be chased locally and simultaneously. The corresponding off-diagonal blocks are updated by explicit multiplication with the orthogonal matrix accumulated in the diagonal chasing step. This delay-and-accumulate technique leads to level 3 computational intensity. When the chains are passing through the processor border, they are chased in an odd-even manner (see Figure 4) to avoid conflicts between different tightly coupled chains. We refer to [21] for a detailed description of this technique.

2.3 Aggressive Early Deflation (AED)

The AED technique has been proposed by Braman, Byers, and Mathias [13] and proceeds by partitioning the current upper Hessenberg matrix $H \in \mathbb{R}^{n \times n}$ as

$$H = \begin{array}{c} \\ n-n_{\text{AED}}-1 \\ 1 \\ n_{\text{AED}} \end{array} \begin{array}{c} \begin{array}{ccc} n-n_{\text{AED}}-1 & 1 & n_{\text{AED}} \end{array} \\ \begin{pmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ 0 & H_{32} & H_{33} \end{pmatrix} \end{array},$$

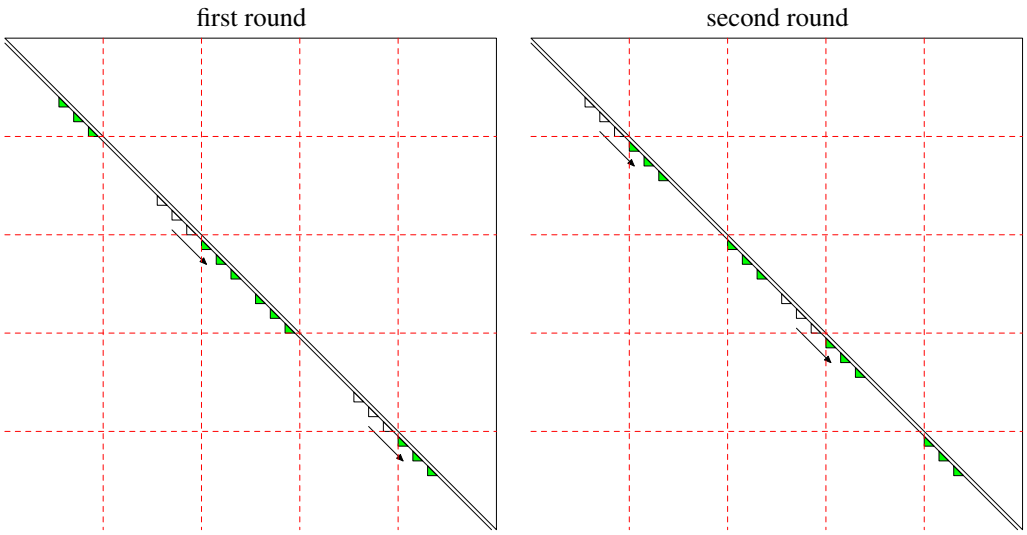


Figure 4: Interblock bulge chasing. Odd-numbered chains (left) and even-numbered chains (right) are chased separately in two rounds.

Table 1: Recommended values for n_{shift} and n_{AED}. Values taken from Table 2.1 in [21] for $n \leq 96K$, values for $n > 96$K are extrapolations. These values can be tuned by the user. However, such a tuning would not only need to take the computer architecture into account but also the balance between multishift QR iterations and AED, which depends on the particular matrix under consideration.

matrix size (n)	n_{shift}	n_{AED}
<6K	see [14]	
6K–12K	256	384
12K–24K	512	768
24K–48K	1024	1536
48K–96K	2048	3072
96K–192K	4096	6144
192K–384K	8192	12288
384K–768K	16384	24576
768K–1000K	32768	49152
> 1M	$\lceil n/25 \rceil$	$3n_{\text{shift}}/2$

where $H_{33} \in \mathbb{R}^{n_{\mathrm{AED}} \times n_{\mathrm{AED}}}$ is the so called *AED window*. By computing the (real) Schur decomposition $H_{33} = VTV^T$, and applying the corresponding similarity transformation to H, we obtain

$$U^T H U = \begin{pmatrix} H_{11} & H_{12} & H_{13}V \\ H_{21} & H_{22} & H_{23}V \\ 0 & s & T \end{pmatrix},$$

where

$$U = \begin{array}{c} \\ n-n_{\mathrm{AED}}-1 \\ 1 \\ n_{\mathrm{AED}} \end{array} \begin{array}{c} \begin{array}{ccc} n-n_{\mathrm{AED}}-1 & 1 & n_{\mathrm{AED}} \end{array} \\ \begin{pmatrix} I & & \\ & 1 & \\ & & V \end{pmatrix} \end{array}.$$

The vector $s \in \mathbb{R}^{n_{\mathrm{AED}}}$ is the so called *spike*, created from the first entry of the vector H_{32}. The last diagonal entry (or 2×2 diagonal block) of T can be deflated if the magnitude of the last component (or the last two components) of the spike is negligible. Undeflatable eigenvalues are moved to the top left corner of T by a swapping algorithm [6, 20]. The orthogonal transformations for reordering eigenvalues in the Schur form of the AED window are accumulated in an $n_{\mathrm{AED}} \times n_{\mathrm{AED}}$ orthogonal matrix. By repeating the same procedure to all diagonal entries (or 2×2 blocks) of T, the eigenvalues of T are checked subsequently and possibly deflated. Then the entire matrix is reduced back to upper Hessenberg form and the off-diagonal blocks H_{13} and H_{23} are multiplied by $\widetilde{V}$, the product of all involved orthogonal transformations. Typically, the size of the AED window is recommended to be somewhat larger, e.g., by 50%, than the number of shifts in the multishift QR sweeps [12, 14], see also Table 1.

In principle, aggressive early deflation can be incorporated into any variant of the QR algorithm. Therefore both the new multishift QR algorithm and the pipelined QR algorithm benefit from performing AED. As the efficient implementation of AED requires some care, we will now discuss these two settings in more detail.

2.3.1 AED within the new multishift QR algorithm

As this setting will be used for targeting large-scale problems, the AED window can be expected to become quite large. It is therefore not reasonable to expect that executing AED locally and sequentially yields good performance. Hence, the corresponding routine `PDLAQR3` for performing AED requires a parallel approach.

The first and most costly step of AED is to calculate the Schur decomposition of H_{33}, the $n_{\mathrm{AED}} \times n_{\mathrm{AED}}$ trailing principal submatrix of H. This eigenvalue problem can be solved by either recursively using the new multishift QR algorithm (`PDLAQR0`) or using the pipelined QR algorithm (`PDLAQR1`). The choice of the solver is determined by the size of the AED window as well as the number of processors used. Since n_{AED} is relatively small compared to n, the number of available processors may be too large to facilitate all of them without causing significant communication overhead. In this case, we only use a subset of the processors to reduce the overhead and minimizing the execution time. See Section 2.6 for a more detailed discussion.

In the deflation checking phase, the reordering algorithm is arranged in a blocked manner, to reduce memory transfers and communication. Unlike the procedure described in [12], undeflatable eigenvalues are not moved immediately and individually towards the top left corner of the AED window. Instead, they are first reordered within an $n_b \times n_b$ computational window. Only after all eigenvalues in this $n_b \times n_b$ window are checked, the group of undeflatable eigenvalues is moved simultaneously to the top left corner of the AED window. This blocked approach increases the computational intensity and avoids the frequent communication needed when reordering each eigenvalue individually. The procedure is repeated until all eigenvalues in the AED window are checked.

The last step is to eliminate the spike and reduce the AED window back to the upper Hessenberg form. This task is performed with the ScaLAPACK routine `PDGEHRD`. The corresponding off-diagonal blocks are updated by explicitly multiplying the accumulated orthogonal matrix using the PBLAS routine `PDGEMM`.

2.3.2 AED within the pipelined QR algorithm

Within the AED stage in the new parallel multishift QR algorithm, the pipelined QR algorithm is often used as the eigensolver since the AED window is relatively small compared to the whole matrix. However, the original

ScaLAPACK v1.8.0 implementation PDLAHQR of the pipelined QR algorithm is not equipped with the AED strategy and is hence not efficient. When adding AED, we have taken into account that we will only use this routine for small- to medium-size (sub)matrices. In particular, we can expect the AED window to be sufficiently small such that AED can be performed on one processor efficiently, by using the LAPACK implementation of AED.

Apart from AED, our new routine PDLAQR1 incorporates further modifications to PDLAHQR, making it both faster and more robust. The following list summarizes the most important aspects; additional details can be found in [30, 34].

- Aggressive early deflation: AED is implemented in an auxiliary routine PDLAQR2 which copies the AED window to local memory and calls the LAPACK routine DLAQR3 to solve the problem sequentially. To determine the parameters of AED, we use the settings of the LAPACK installation determined by ILAENV. However, a notable difference is that we do *not* use the undeflatable eigenvalues from the AED step as shifts in the subsequent pipelined QR sweep. Instead, we recompute the eigenvalues of the trailing submatrix to improve the quality of the shifts. We have observed that this shifting strategy accelerates the convergence of the pipelined QR algorithm.

- Conventional deflation: In PDLAHQR, pipelined QR sweeps are performed until the very end, that is, the remaining diagonal blocks are all of size 1×1 or 2×2. In PDLAQR1, we use a different strategy: Once the active block is sufficiently small (say, not larger than 385×385), we copy this block to local memory and call the LAPACK routines DLAHQR/DLAQR4 to solve the problem sequentially. This strategy significantly reduces communication overhead in the latter stages and is implemented in an auxiliary routine PDLAQR4.

- Avoidance of anomalies: The original ScaLAPACK routine PDLAHQR suffered from two anomalies, which have been removed. First, the routine sometimes returned 2×2 diagonal blocks containing real eigenvalues, which is not in accordance with the specification of the interface. In PDLAQR1, each 2×2 diagonal block contains a pair of complex conjugate eigenvalues. The second issue is concerned with a strategy already proposed by Francis [17] to potentially benefit from two consecutive small sub-diagonal entries. In the rare case when it is successful, this strategy allows to introduce bulges below the top left corner of the active submatrix. However, this turns out to be difficult to implement in a safe manner in pipelined or multishift QR sweeps. Indeed, when using the ScaLAPACK routine PDLACONSB which implements this strategy, we have observed large relative residuals of norm up to 10^{-5}, indicating numerical instabilities. As the performance improvements gained from this strategy are usually negligible, we have decided to remove it. In return, the numerical stability is improved.

2.4 Switching between Multishift QR and AED

In the LAPACK implementation of the QR algorithm, there are rules for balancing the cost between multishift QR sweeps and AED, see Step 4 in Algorithm 1. The precise meaning of this step is characterized by a threshold called NIBBLE. If we let n_{undflt} denote the number of undeflatable shifts in an AED step, the multishift QR sweep is skipped if

$$\frac{n_{\text{AED}} - n_{\text{undflt}}}{n_{\text{AED}}} \geq \frac{\texttt{NIBBLE}}{100}.$$

Since AED behaves differently for different matrices, this strategy automatically adjusts the choice between AED and QR sweeps based on the properties of the matrix.

The default value of NIBBLE in the sequential LAPACK implementation is 14, which provides a good balance between multishift QR sweeps and AED. The same default value is used in the pipelined QR algorithm. However, in the new parallel multishift QR algorithm, the parallel AED process becomes substantially more expensive than the sequential AED process due to communication. As explained above, the AED process only involves a smaller trailing submatrix, leading to decreased parallel efficiency. To account for this, NIBBLE should be set larger to avoid performing AED too frequently. A good choice of this threshold depends both on the size of the matrix H and the number of processors involved. We use the model $\texttt{NIBBLE} = a \cdot n^b p^c$ for this purpose, where a, b, and c

are machine-dependent constants. An appropriate choice of these constants can be gained from repeated runs of the program with different thresholds. It turns out that the right choice of `NIBBLE` becomes rather sensitive when communication is slow. (In our numerical experiments, see Section 5, the default values on our computing architectures are chosen as $(a, b, c) = (335, -0.44, 0.5)$.)

A complication arises when using `NIBBLE` > 33. Such a choice may lead to situations where the number of undeflatable eigenvalues is less than the desired number of shifts, that is $n_{\text{undflt}} < n_{\text{shift}}$, due to the fact that $n_{\text{AED}} = 3n_{\text{shift}}/2$. The solution in the software is that as long as $n_{\text{undflt}} \geq n_{\text{shift}}/2$, we only use these n_{undflt} undeflatable eigenvalues as shifts in the next QR sweep. However, the condition $n_{\text{undflt}} \geq n_{\text{shift}}/2$ may also fail when `NIBBLE` ≥ 67. In this case we calculate the eigenvalues of the $n_{\text{shift}} \times n_{\text{shift}}$ trailing principal submatrix of H and use them as shifts. The calculation can be performed by either `PDLAQR0` or `PDLAQR1`, just like computing the Schur decomposition in the AED step.

2.5 Task duplication—efficient but hazardous

Task duplication is a common technique in parallel computing to reduce communication and potentially improve performance, see, e.g., [19]. This technique has already been employed in previous implementations of the parallel QR algorithm, such as `PDLAHQR` and software developed in [21, 30]. However, a crucial assumption is made when applying this technique: all involved processors need to produce identical outputs for identical tasks. Due to the effect of roundoff error, this assumption is not always satisfied, especially on heterogeneous architectures.

This lack of numerical reproducibility is potentially harmful to the robustness of the parallel QR algorithm. As discussed in Section 2.3.2, all computations within the AED window are performed sequentially for the pipelined QR algorithm. In [30], we have proposed to duplicate these sequential parts on all involved processors. The parallel update of the off-diagonal blocks can then be performed with the local copy of the orthogonal transformation matrix resulting from AED, without any extra communication. However, even the slightest change in finite-precision arithmetic may lead to very different outputs produced by AED. In particular, the ordering of the eigenvalues in the Schur decomposition computed within AED is very sensitive to such changes. In turn, the off-diagonal blocks are updated using completely different local copies of the orthogonal transformation matrices, leading to meaningless results. We have observed similar problems in crossborder bulge chasing and eigenvalue reordering. To avoid this, we use explicit communication rather than task duplication in the new implementation. For a moderate number of processors (e.g., $p \leq 100$), the change in performance is negligible; while for a large number of processors, the performance can drop. For example, for computing the Schur decomposition of a $100,000 \times 100,000$ matrix using 40×40 processors, up to 25% performance drop has been observed by replacing task duplication with explicit communication.

2.6 Avoiding communication via data redistribution

As observed in [30], the parallel QR algorithm is not efficient at solving relatively small eigenvalue problem on many processors, due to excessive communication, to the point that the execution time actually increases when increasing the number of processors. Such a situation is regularly encountered when calculating the Schur decomposition of the AED window. Therefore, an efficient parallel solver for this relatively small eigenvalue problem is of great interest. One possible approach to this problem is to use alternative algorithms, e.g., the spectral divide-and-conquer algorithm [7, 8]. In the following we propose a solution within the framework of the QR algorithm—a data redistribution strategy that reduces the communication overhead.

Recent work on communication avoiding algorithms [9, 8, 25, 28] usually focuses on the design of algorithms that can attain the theoretical lower bounds of the communication cost. A basic assumption in these theoretical analyses is that the data are nearly evenly distributed over the processors. Here we propose an alternative approach, which does not rely on this assumption and is especially useful for operations involving smaller submatrices.

We first consider a simple and extreme case. Suppose there is one processor which has a large amount of local memory and very high clock speed. Then by gathering all data to this processor, the problem can be solved

without further communication. Once the computation is completed, the data are scattered to their original owners. The total amount of communication does not exceed the cost of scattering and gathering regardless of the complexity of computational work. Although this simple idea does not work for large problems that cannot be stored on a single processor, it is still useful for smaller problems. For example, the AED process in the pipelined QR algorithm is implemented in such a manner since we know in advance that the AED window is always sufficiently small, such that the associated Schur decomposition can be efficiently solved sequentially. By introducing the overhead of data redistribution, the total amount of communication as well as the execution time are reduced.

For larger problems, it is not feasible to solve them sequentially via data redistribution. Specifically, the AED window within the new parallel multishift QR algorithm usually becomes quite large, although much smaller compared to the whole matrix. In this case, we choose a subset of processors instead of a single processor to perform AED. The data redistribution is performed using the routine `PDGEMR2D` in ScaLAPACK; its overhead has been observed to be negligible relative to the AED process as a whole.

The tunable parameter $p_{\min}$ = `PILAENVX(ISPEC = 23)` determines our heuristic strategy for choosing the number of processors for the redistribution. If $\min(p_r, p_c) > p_{\min} + 1$, we redistribute the AED window to a $p_{\min} \times p_{\min}$ processor grid and perform the calculations on this subset of processors. The same strategy is also applied if we need to compute shifts after an AED step. The default value for this parameter is $p_{\min} = \lceil n_{\text{AED}}/(n_b\lceil 384/n_b\rceil)\rceil$, implying that each processor needs to own at least 384 columns of the AED window. The constant 384 has been obtained via extensive numerical experiments on one of our target architectures. It certainly needs adjustment for optimal performance on other architectures.

2.7 Summary

Finally, we present pseudocode to summarize the discussion on our new parallel multishift QR algorithm. Algorithm 2 represents the parallel AED procedure used within the new multishift QR algorithm. Algorithm 3 provides pseudocode for the parallel multishift QR sweeps. For simplicity, the start-up and ending stages, i.e., bulge introduction and bulge annihilation, are not considered in the pseudocode. These algorithms are to be used within Algorithm 1 on distributed memory architectures.

3 Performance model

In this section, we briefly discuss the cost of computation and communication of the new parallel multishift QR algorithm for reducing a Hessenberg matrix to Schur form. The cost of accumulating orthogonal transformations are taken into account. The derivation of the results presented here will be given in the electronic appendix. For simplicity, we consider a square processor grid, that is, $p_r = p_c = \sqrt{p}$. In addition, we assume that each processor contains reasonably many data blocks of the matrices, i.e., $\sqrt{p}\, n_b \ll n$, so that the work load is balanced. The parallel execution time consists of two main components:

$$T_p = T_a + T_c,$$

where T_a and T_c are the times for arithmetic operations and communication, respectively. The possibility of overlapping communication with computations is not taken into account. By neglecting the communication between memory and cache lines inside one core, the serial runtime can be approximated by

$$T_a = \frac{\#(\text{flops})}{f(p)}\gamma,$$

where γ is the average time for performing one floating point operation and $f(p)$ is the degree of concurrency. For the communication between two processors, we define α and β as the start-up time (or communication latency) and the time for transferring one word without latency (or reciprocal of bandwidth), respectively. The time for a single point-to-point communication is modelled as $\alpha + L\beta$ where L is the message size in words. A one-to-all broadcast or an all-to-one reduction within a scope of p processors is assumed to take $\Theta(\log p)$ steps.

Algorithm 2 Parallel aggressive early deflation within the new multishift QR algorithm

Input: $H \in \mathbb{R}^{n\times n}$ is upper Hessenberg, with the AED window contained in $H(i:j, i:j)$; $Q \in \mathbb{R}^{n\times n}$ is orthogonal.

Output: Updated Hessenberg matrix $H \leftarrow U^T HU$ obtained after performing AED, $Q \leftarrow QU$, with U orthogonal.

1: Estimate the optimal process grid size, $p_{\text{AED}} = p_{\min}^2$, based on $n_{\text{AED}} = j - i + 1$ and n_b.
2: **if** $\min(p_r, p_c) \le p_{\min} + 1$ **then**
3: $\quad \hat{p}_r \leftarrow p_{\min}$, $\hat{p}_c \leftarrow p_{\min}$.
4: **else**
5: $\quad \hat{p}_r \leftarrow p_r$, $\hat{p}_c \leftarrow p_c$.
6: **end if**
7: Redistribute $H_0 \leftarrow H(i:j, i:j)$ to a $\hat{p}_r \times \hat{p}_c$ process subgrid if necessary.
8: Compute the Schur decomposition $H_0 = V_0 T_0 V_0^T$ on the $\hat{p}_r \times \hat{p}_c$ process subgrid.
9: Redistribute $T \leftarrow T_0$, $V \leftarrow V_0$ back to the original $p_r \times p_c$ process grid.
10: $s \leftarrow H(i, i-1)V(1,:)^T$.
11: **repeat**
12: $\quad$ Check deflation for the bottommost n_b unchecked eigenvalues; undeflatable eigenvalues are moved to the top-left corner of this $n_b \times n_b$ block.
13: $\quad$ Move all undeflatable eigenvalues within this group of n_b eigenvalues to the top-left corner of the AED window.
14: $\quad$ Update $s \leftarrow V_1^T s$, $T \leftarrow V_1^T T V_1$ in parallel (on the original $p_r \times p_c$ process grid), where V_1 is the accumulated orthogonal matrix from Steps 12 and 13.
15: **until** all eigenvalues of T are tested
16: Eliminate the spike: $V_2^T s = \eta e_1$; Update $T \leftarrow V_2^T T V_2$ in parallel.
17: Reduce T to an upper Hessenberg matrix $H_1 \leftarrow V_3^T T V_3$ in parallel.
18: Set $H(i:i-1) \leftarrow \eta$, $H(i:j, i:j) \leftarrow T$; Update $V \leftarrow VV_3$ in parallel.
19: Update $H(i:j, j+1:n) \leftarrow V^T H(i:j, j+1:n)$ in parallel.
20: Update $H(1:i-1, i:j) \leftarrow H(1:i-1, i:j)V$ in parallel.
21: Update $Q(1:n, i:j) \leftarrow Q(1:n, i:j)V$ in parallel.

Let k_{AED} and k_{sweep} denote the number of AED steps and QR sweeps, respectively, performed by the new parallel multishift QR algorithm. We have $k_{\text{sweep}} \le k_{\text{AED}}$, since some QR sweeps are skipped when the percentage of deflated eigenvalues in the AED step is larger than the threshold (`NIBBLE`). When the number of undeflatable eigenvalues from AED is not sufficient for performing the next QR sweep, we need to calculate shifts from the trailing submatrix. The number of extra calls to the parallel Schur decomposition solver in this case is denoted by k_{shift}, which of course satisfies $k_{\text{shift}} \le k_{\text{sweep}}$. Given the constants k_{AED}, k_{sweep}, and k_{shift}, the execution time of the new parallel multishift QR algorithm is modelled as the sum of the corresponding phases:

$$T_{\text{new}}(n,p) = k_{\text{AED}} T_{\text{AED}}(n, n_{\text{AED}}, p) + k_{\text{sweep}} T_{\text{sweep}}(n, n_{\text{shift}}, p) + k_{\text{shift}} T_{\text{shift}}(n, n_{\text{shift}}, p),$$

where T_{AED}, T_{sweep}, and T_{shift} are the runtimes for performing each phase once. For simplicity, it is assumed that n_{AED} and n_{shift} remain constant throughout the entire QR algorithm, and all QR sweeps act on the entire matrix. We further assume that AED is always performed on a $\sqrt{p_{\text{AED}}} \times \sqrt{p_{\text{AED}}}$ processor grid, so that the property $\sqrt{p_{\text{AED}}}\, n_b \ll n_{\text{AED}}$ is also valid inside the AED window. The same assumption is made for the shift calculation phase.

Typically, we have

$$n_{\text{shift}} \approx \frac{2}{3} n_{\text{AED}} \approx \frac{1}{C_1} n \quad \text{and} \quad \frac{n_{\text{AED}}}{\sqrt{p_{\text{AED}}}} \approx \frac{n_{\text{shift}}}{\sqrt{p_{\text{shift}}}} \ge C_2,$$

where C_1 and C_2 are constants (e.g., $C_1 = 24$, $C_2 = 384$). In practice k_{AED}, k_{sweep}, k_{shift} can vary a lot for different matrices. We assume $k_{\text{AED}} = \Theta(n/n_{\text{AED}}) = \Theta(C_1)$, which appears to be reasonable.

In the following we present performance models based on the assumptions above. Tiny terms, especially lower order terms with reasonably sized constants, are omitted. The derivation of these models can be found in

Algorithm 3 Parallel multishift QR sweep (bulge chasing process)

Input: $H \in \mathbb{R}^{n\times n}$ is upper Hessenberg except for several tightly coupled bulge chains in the top left corner; $Q \in \mathbb{R}^{n\times n}$ is orthogonal.

Output: Updated matrix $H \leftarrow U^T HU$ has the tightly coupled bulge chains moved to the bottom right corner, $Q \leftarrow QU$, where U is orthogonal and the new H is upper Hessenberg.

for each window $w = (i : i + n_b - 1)$ in parallel **do**
  if $(myrow, mycol)$ owns parts of $(:, w)$ or $(w, :)$ **then**
    if $(myrow, mycol)$ owns (w, w) **then**
      Chase the bulge chain inside w down $\lfloor n_b/2 \rfloor$ rows.
      Broadcast the local orthogonal matrix V in process row $myrow$.
      Broadcast the local orthogonal matrix V in process column $mycol$.
    else
      Receive V.
    end if
    Update $H(w, i + n_b : n) \leftarrow V^T H(w, i + n_b : n)$ in parallel.
    Update $H(1 : i - 1, w) \leftarrow H(1 : i - 1, w)V$ in parallel.
    Update $Q(1 : n, w) \leftarrow Q(1 : n, w)V$ in parallel.
  end if
end for
for each odd-numbered window $w = (j : j + n_b - 1)$ in parallel **do**
  if $(myrow, mycol)$ owns parts of $(:, w)$ or $(w, :)$ **then**
    if $(myrow, mycol)$ owns parts of (w, w) **then**
      Form a process subgrid $G_w = \{(0, 0)_w, (0, 1)_w, (1, 0)_w, (1, 1)_w\}$.
      Exchange data in G_w to build $H(w, w)$ at $(0, 0)_w$.
      if $(myrow, mycol) = (0, 0)_w$ **then**
        Chase the bulge chain inside w down $\lceil n_b/2 \rceil$ rows.
        Send $H(w, w)$ and the local orthogonal matrix V to other processes in G_w.
      else
        Receive $H(w, w)$ and V from $(0, 0)_w$.
      end if
      if $(myrow, mycol) = (0, 0)_w$ **or** $(myrow, mycol) = (1, 1)_w$ **then**
        Broadcast V in process row $myrow$.
        Broadcast V in process column $mycol$.
      end if
    else
      Receive V.
    end if
    Exchange local parts of $H(w, j + n_b : n)$, $H(1 : j - 1, w)$, and $Q(1 : n, w)$ with neighboring processes in parallel.
    Update $H(w, j + n_b : n) \leftarrow V^T H(w, j + n_b : n)$ in parallel.
    Update $H(1 : j - 1, w) \leftarrow H(1 : j - 1, w)V$ in parallel.
    Update $Q(1 : n, w) \leftarrow Q(1 : n, w)V$ in parallel.
  end if
end for
for each even-numbered window $w = (j : j + n_b - 1)$ in parallel **do**
  % Analogous procedure as described for the odd case above.
end for

the electronic appendix. By [24], the execution time of the pipelined QR algorithm is

$$T_{\text{pipe}}(n, p) = \Theta\Big(\frac{n^3}{p}\Big)\gamma + \Theta\Big(\frac{n^2 \log p}{\sqrt{p}\, n_b}\Big)\alpha + \Theta\Big(\frac{n^3}{p n_b}\Big)\beta, \tag{2}$$

provided that the average number of shifts required for deflating each eigenvalue is $\Theta(1)$. One QR sweep in our new parallel multishift QR algorithm roughly requires

$$T_{\text{sweep}}(n, n_{\text{shift}}, p) \approx \frac{36n^2 n_{\text{shift}}}{p}\gamma + \frac{9n\, n_{\text{shift}}}{\sqrt{p}\, n_b^2}(\log_2 p + 4)\alpha + \frac{9n^2 n_{\text{shift}}}{p n_b}\beta$$

execution time. Therefore, under the same assumption of convergence rate (i.e., $k_{\text{sweep}} n_{\text{shift}} = \Theta(n)$), it can be shown that the execution time of the new parallel multishift QR algorithm *without* AED is

$$T_{\text{new}}(n, p) = k_{\text{sweep}} T_{\text{sweep}}(n, n_{\text{shift}}, p) = \Theta\Big(\frac{n^3}{p}\Big)\gamma + \Theta\Big(\frac{n^2 \log p}{\sqrt{p}\, n_b^2}\Big)\alpha + \Theta\Big(\frac{n^3}{p n_b}\Big)\beta. \tag{3}$$

It is interesting to make a comparison between (2) and (3). Both solvers have an ideal degree of concurrency. However, the tightly coupled shift strategy is superior to loosely coupled shifts, because it yields less frequent communication. The number of messages is reduced by a factor of $\Theta(n_b)$; in return the average message length increases correspondingly. Another important observation is that the serial term in T_{pipe} assumes level 3 performance, which is actually not the case. This already explains why the pipelined QR algorithm is usually much slower than the new parallel multishift QR algorithm for larger matrices, even when neglecting the effects of AED.

Taking AED into account makes the model significantly more complicated. The execution times for AED and computing shifts are estimated by

$$\begin{aligned} T_{\text{AED}}(n, n_{\text{AED}}, p) \approx{}& \left[\frac{30C_2^2 n}{C_1} + \frac{9n^2 n_b}{C_1^2\sqrt{p}} + \frac{9(C_1+6)n^3}{2C_1^3 p}\right]\gamma \\ &+ \left(\frac{9C_2 n}{C_1 n_b}\log_2\frac{3n}{2C_1C_2} + \frac{3n}{2C_1}\log_2 p\right)\alpha \\ &+ \left[\frac{9C_2 n}{C_1}\log_2\frac{3n}{2C_1C_2} + \frac{12C_2^2 n}{C_1 n_b} + \frac{3n^2(18 + C_1 + 51\log_2 p)}{16C_1^2\sqrt{p}}\right]\beta \end{aligned}$$

and

$$T_{\text{shift}}(n, n_{\text{shift}}, p) \approx \frac{10C_2^2 n}{C_1}\gamma + \frac{6C_2 n}{C_1 n_b}\log_2\frac{n}{C_1C_2}\alpha + \Big(\frac{6C_2 n}{C_1}\log_2\frac{n}{C_1C_2} + \frac{8C_2^2 n}{C_1 n_b}\Big)\beta,$$

respectively. To be able to provide some intuition, we assign concrete values to most parameters. For example, let us set $C_1 = 24$, $C_2 = 384$, and assume $k_{\text{AED}} = 2k_{\text{sweep}} = 16k_{\text{shift}} = 64$. Then

$$\begin{aligned} T_{\text{sweep}}(n, n_{\text{shift}}, p) &\approx \frac{3n^3}{2p}\gamma + \frac{3n^2}{8\sqrt{p}\, n_b^2}(\log_2 p + 4)\alpha + \frac{3n^3}{8pn_b}\beta, \\ T_{\text{AED}}(n, n_{\text{AED}}, p) &\approx \Big(184320n + \frac{n^2 n_b}{64\sqrt{p}} + \frac{5n^3}{256p}\Big)\gamma \\ &\quad + \Big[\frac{144n}{n_b}(\log_2 n - 14) + \frac{n\log_2 p}{8}\Big]\alpha \\ &\quad + \Big[144n\Big(\log_2 n - 14 + \frac{512}{n_b}\Big) + \frac{(51\log_2 p + 42)n^2}{3072\sqrt{p}}\Big]\beta, \\ T_{\text{shift}}(n, n_{\text{shift}}, p) &\approx 61440n\gamma + \frac{96n}{n_b}(\log_2 n - 13)\alpha + 96n\Big(\log_2 n - 13 + \frac{512}{n_b}\Big)\beta. \end{aligned} \tag{4}$$

This yields the following overall estimate for the new parallel QR algorithm with AED:

$$T_{\text{new}}(n, p) \approx \Big(\frac{48n^3}{p} + 1.2\times 10^7 n\Big)\gamma + \frac{12n^2\log_2 p}{\sqrt{p}\, n_b^2}\alpha + \Big(\frac{12n^3}{pn_b} + \frac{17n^2\log_2 p}{16\sqrt{p}}\Big)\beta, \tag{5}$$

$$= \Theta\Big(\frac{n^3}{p}\Big)\gamma + \Theta\Big(\frac{n^2\log p}{\sqrt{p}\, n_b^2}\Big)\alpha + \Theta\Big(\frac{n^3}{pn_b}\Big)\beta, \tag{6}$$

Figure 5: Calling sequences of the newly developed routine PDHSEQR, the corresponding LAPACK routine DHSEQR, and the ScaLAPACK routine PDLAHQR.

```
      SUBROUTINE PDHSEQR( JOB, COMPZ, N, ILO, IHI, H, DESCH, WR, WI, Z,
     $                    DESCZ, WORK, LWORK, IWORK, LIWORK, INFO )
*
*     .. Scalar Arguments ..
      INTEGER            IHI, ILO, INFO, LWORK, LIWORK, N
      CHARACTER          COMPZ, JOB
*     ..
*     .. Array Arguments ..
      INTEGER            DESCH( * ) , DESCZ( * ), IWORK( * )
      DOUBLE PRECISION   H( * ), WI( N ), WORK( * ), WR( N ), Z( * )
```

```
      SUBROUTINE DHSEQR( JOB, COMPZ, N, ILO, IHI, H, LDH, WR, WI, Z,
     $                   LDZ, WORK, LWORK, INFO )
```

```
      SUBROUTINE PDLAHQR( WANTT, WANTZ, N, ILO, IHI, A, DESCA, WR, WI,
     $                    ILOZ, IHIZ, Z, DESCZ, WORK, LWORK, IWORK,
     $                    ILWORK, INFO )
```

where most small-order terms are neglected. It turns out that both QR sweeps and AED have significant serial runtime when n is not very large. However, QR sweeps usually dominate the communication cost. As a consequence, the models (2), (3), and (6) nearly have the same asymptotic behavior. AED is asymptotically not more expensive compared to QR sweeps, and hence it does not represent a computational bottleneck for larger matrices. Combined with the convergence acceleration often observed when using AED (and not fully attributed in the model above), this contributes to the superior performance of the new parallel multishift QR algorithm.

4 Other Implementation Issues

4.1 Calling sequence

The calling sequence of the newly developed routine PDHSEQR is nearly identical with the LAPACK routine DHSEQR, see Figure 5. Apart from the need of a descriptor for each globally distributed matrix and the leading dimension for each local matrix, the only difference is that PDHSEQR requires an extra integer workspace. The calling sequence of the ScaLAPACK routine PDLAHQR is also similar, hopefully allowing to easily switch from PDLAHQR and DHSEQR in existing software making use of ScaLAPACK. In practice, it is advisable to call PDHSEQR twice—one call for the workspace query (by setting LWORK=-1) and another call for actually doing the computation. This follows the convention of many LAPACK/ScaLAPACK routines that make use of workspace.

4.2 Tuning parameters

In the new software for the parallel multishift QR algorithm, tunable parameters are defined in the routine PILAENVX. They are available via the function call PILAENVX(ICTXT, ISPEC, ...) with $12 \leq$ ISPEC ≤ 23. A complete list of these parameters is provided in Table 2. Some of them require fine tuning to attain nearly optimal performance across different architectures.

Although a reasonable choice of n_b, the data layout block size, is important, we have observed the performance to be not overly sensitive to this choice. On the one hand, n_b should be large enough so that the local computations can achieve level 3 performance. On the other hand, it is advisable to avoid n_b being too large. A large value of n_b harms load balance and increases the overhead in the start-up and ending stages of the bulge chasing process, especially when computing the Schur decomposition of the AED window. In our performance

Table 2: List of tunable parameters

ISPEC	Name	Description	Recommended value
12	$n_{\min}$	Crossover point between PDLAQR0 and PDLAQR1	$220\min(p_r, p_c)$
13	n_{AED}	Size of the AED window	See Table 1
14	NIBBLE	Threshold for skipping a multishift QR sweep	See Section 2.4
15	n_{shift}	Number of simultaneous shifts	See Table 1
16	KACC22	Specification of how to update off-diagonal blocks in the multishift QR sweep	Use GEMM/TRMM
17	NUMWIN	Maximum number of concurrent computational windows (for both QR sweep and eigenvalue reordering)	$\min(p_r, p_c, \lceil n/n_b \rceil)$
18	WINEIG	Number of eigenvalues in each window (for eigenvalue reordering)	$\min(n_b/2, 40)$
19	WINSIZE	Computational window size (for both bulge-chasing and eigenvalue reordering)	$\min(n_b, 80)$
20	MMULT	Minimal percentage of flops for performing GEMM instead of pipelined Householder reflections when updating the off-diagonal blocks in the eigenvalue reordering routine	50
21	NCB	Width of block column slabs for rowwise update of Householder reflections in factorized form	$\min(n_b, 32)$
22	WNEICR	Maximum number of eigenvalues to move over a block border in the eigenvalue reordering routine	Identical to WINEIG
23	$p_{\min}$	Size of processor grid involving AED	See Section 2.6

model, we always assume $n_b \ll n/\sqrt{p}$ to avoid such kind of overhead. For many architectures, $n_b \in [32, 128]$ will offer a good choice.

We expect that most of the recommended values in Table 2 yield reasonable performance on existing architectures. However, the parameters $n_{\min}$, $p_{\min}$ and NIBBLE require some extra care, as the performance of AED crucially relies on them. The values of these parameters need to be determined by performing a bunch of test runs: To determine $n_{\min}$ and $p_{\min}$, it is advisable to use the typical sizes of the AED windows (see Table 1) and run tests on different processor grids. Then the optimal values for both $n_{\min}$ and $p_{\min}$ can be chosen via examining the number of columns of H owned by each processor. NIBBLE should be tuned lastly, once all other parameters are fixed. Auto-tuning tools have been provided in the software for tuning these three parameters. We refer to the User's Guide [22] for details of the tuning procedure. Tuning NIBBLE is time-consuming but highly recommended, especially on older architectures with relatively slow communication. As discussed in Section 2.4, NIBBLE $= a \cdot n^b p^c$ is a reasonably good model that takes into account both n and p. We have observed reasonable performance behavior of this model for large problems up to size $n = 100,000$ and $p = 40 \times 40$.

5 Computational Experiments

We have performed a large set of computational experiments on *Akka* [27] and *Abisko* [26] hosted by the High Performance Computing Center North (HPC2N), and on *Bellatrix* [15] hosted by École Polytechnique Fédérale de Lausanne. In this section, we present a subset from these computational experiments to confirm and demonstrate the improvements we have made in the parallel QR algorithm. The computational environments are summarized in Table 3.

We compare the following implementations:

Table 3: Computational environments.

Akka	64-bit Intel Xeon (Harpertown) Linux cluster 672 dual socket nodes with L5420 quad-core 2.5GHz processors and 16GB RAM per node Cisco Infiniband and Gigabit Ethernet, 10Gbps bandwidth PathScale compiler version 4.0.13 OpenMPI 1.4.4, LAPACK 3.4.0, GotoBLAS2 1.13
Abisko	64-bit AMD Opteron (Interlagos) Linux cluster 322 nodes with four Opteron 6238 12-core 2.6GHz processors and 128GB RAM per node Mellanox 4X QSFP Infiniband connectivity, 40Gbps bandwidth PathScale compiler version 4.0.13 OpenMPI 1.6.4, LAPACK 3.4.0, OpenBLAS 0.1 alpha 2.4
Bellatrix	64-bit Intel Xeon (Sandy Bridge) Linux cluster 424 dual socket nodes with E5-2660 octa-core 2.2GHz processors and 32GB RAM per node Qlogic Infiniband QDR 2:1 connectivity, 40Gbps bandwidth Intel compiler version 13.0.1 Intel MPI version 4.1.0, Intel Math Kernel Library version 11.0

S-v180	Pipelined QR algorithm in ScaLAPACK v1.8.0.
SISC	Previous implementation of parallel multishift QR algorithm with AED, as described in [21].
NEW	New implementation of parallel multishift QR algorithm with AED, as described in this paper.

The implementation NEW improves upon the implementation SISC in terms of robustness and performance, in particular because of the proposed modifications to the AED step.

The data layout block size $n_b = 50$ is used for all experiments. No multithreaded features (such as OpenMP or threaded BLAS) are used. Therefore the number of processors ($p = p_r \cdot p_c$) means the number of cores involved in the computation.

5.1 Random matrices

First, we consider two types of random matrices—`fullrand` and `hessrand` [21].

Matrices of the type `fullrand` are dense square matrices with all entries randomly generated from a uniform distribution in $[0, 1]$. We call the ScaLAPACK routine `PDGEHRD` to reduce them to upper Hessenberg form before applying the QR algorithm. Only the time for the QR algorithm (i.e., reducing the upper Hessenberg matrix to Schur form) is measured. These matrices usually have well-conditioned eigenvalues and exhibit "regular" convergence behavior.

Matrices of the type `hessrand` are upper Hessenberg matrices whose nonzero entries are randomly generated from a uniform distribution in $[0, 1]$. The eigenvalues of these matrices are extremely ill-conditioned for larger n, affecting the convergence behavior of the QR sweeps [21]. On the other hand, AED often deflates a high fraction of eigenvalues in the AED window for such matrices. These properties sometimes cause erratic convergence rates.

Tables 4 and 5 show the parallel execution times of the three solvers on *Akka*. Both the real Schur form T and the orthogonal transformation matrix Z are calculated. We limit the total execution time (including the Hessenberg reduction) by 10 hours for each individual problem, to avoid excessive use of the computational resources. An entry ∞ corresponds to an execution time larger than 10 hours. These tables reveal that our new version of `PDHSEQR` (i.e., NEW) always improves the performance compared to the SISC version. On average, the improvement is 31% for matrices of type `fullrand` and 14 times for matrices of type `hessrand`. Not surprisingly, the improvements compared to `PDLAHQR` in ScaLAPACK v1.8.0 are even more significant.

The convergence rates for `fullrand` are sufficiently regular, so that we can analyze the scalability of the

Table 4: Execution time (in seconds) on *Akka* for `fullrand` matrices.

$p =$	$n = 4000$			$n = 8000$			$n = 16000$			$n = 32000$		
$p_r \times p_c$	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW
1×1	834	178	115	10730	939	628						
2×2	317	87	56	2780	533	292						
4×4	136	50	35	764	205	170	6671	1220	710			
6×6	112	50	43	576	142	116	3508	754	446	∞	3163	2200
8×8	100	45	37	464	127	104	2536	506	339	∞	2979	1470
10×10	97	50	36	417	159	119	2142	457	320	∞	2401	1321

Table 5: Execution time (in seconds) on *Akka* for `hessrand` matrices.

$p =$	$n = 4000$			$n = 8000$			$n = 16000$			$n = 32000$		
$p_r \times p_c$	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW
1×1	685	317	14	6981	2050	78						
2×2	322	200	8	2464	1904	27						
4×4	163	112	36	1066	679	71	8653	2439	65			
6×6	137	84	31	768	412	113	4475	1254	71	∞	373	252
8×8	121	68	25	634	321	107	3613	719	71	∞	919	228
10×10	131	83	23	559	313	111	3549	667	76	∞	943	267

parallel multishift QR algorithm. If we fix the memory load per core to $n/\sqrt{p} = 4000$, the execution times in Table 4 satisfy

$$2T(n,p) \leq T(2n,4p) < 4T(n,p),$$

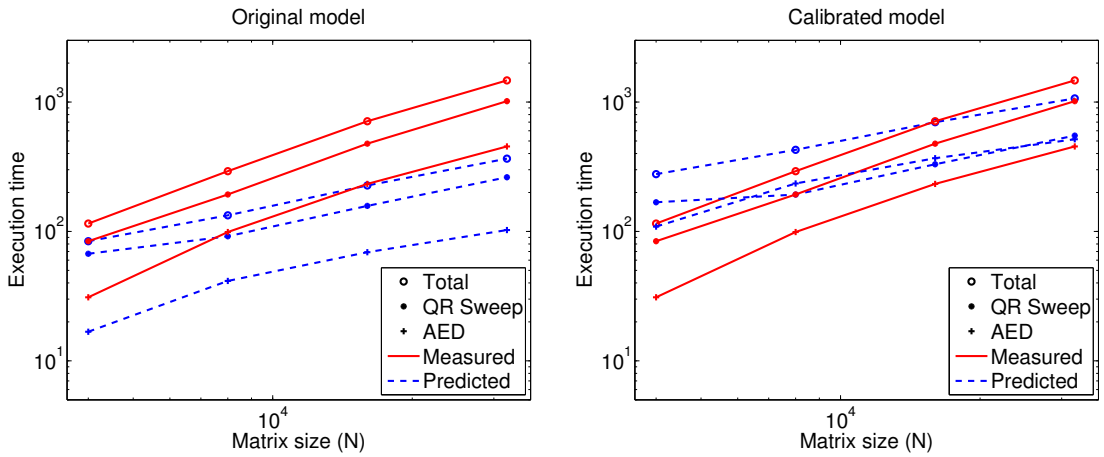
indicating that the parallel multishift QR algorithm scales reasonably well but not perfectly. To verify the performance models we have derived in Section 3, we use (4) together with the measured values of k_{AED}, k_{sweep}, and k_{shift} to predict the execution time. Since $k_{\text{shift}} = 0$ is observed for all these examples, there are only two components in the total execution time (i.e., $T = k_{\text{sweep}}T_{\text{sweep}} + k_{\text{AED}}T_{\text{AED}}$). Figure 6(a) illustrates that the predicted execution times underestimate the measured ones (especially, for AED), mainly due to too optimistic choices of the parameters (α, β, γ). If we assume that the program executes at 40% and 7.5% of the peak core performance for level 3 and level 1–2 BLAS operations, respectively, the calibrated model fits the actual execution time quite well for large matrices (see Figure 6(b)). Since the model(s) are asymptotic, the results are very satisfactory.

For `hessrand`, it is observed that most eigenvalues are deflated with very few (or even no) QR sweeps. Considering that the main difference of `PDHSEQR` between versions NEW and SISC is in the AED process, it is not surprising to see the great convergence acceleration for `hessrand`, where AED dominates the calculation. In Table 5, sometimes the execution time for the new parallel multishift QR algorithm does not change too much when increasing the number of processors. This is mainly because the Schur decomposition of the AED window, which is the most expensive part of the algorithm, is performed by a constant number of processors ($p_{\min} \cdot p_{\min} \leq p$) after data redistribution.

In Tables 6–9 we list the execution times received from *Abisko* and *Bellatrix*. The observations are similar to those obtained from *Akka*. Therefore, in the rest of this section we only present experiments on *Akka* for economical consideration.

5.2 $100,000 \times 100,000$ matrices

The modifications proposed and presented result into dramatic improvements for settings with very large matrices and many processors. To demonstrate this, we present the obtained execution times for $100,000 \times 100,000$ matrices in Table 10. Although the QR algorithm does not scale as well as Hessenberg reduction when fixing the problem size and increasing the number of processors, the execution times of these two reduction steps are still on the same order of magnitude. With the help of the dynamic `NIBBLE` strategy, the fraction of the execution time spent on AED for `fullrand` matrices is under control. In contrast to our earlier implementation, AED is not a



(a) (b)

Figure 6: Comparison between the measured execution times and the predicted times using (4) (`fullrand`, $n/\sqrt{p} = 4000$). The original model (left) uses theoretical values of (α, β, γ) according to the hardware information, while the calibrated one (right) adjusts γ according to different computational intensities (level 1, 2, and 3).

Table 6: Execution time (in seconds) on *Abisko* for `fullrand` matrices.

$p =$	$n = 4000$			$n = 8000$			$n = 16000$			$n = 32000$		
$p_r \times p_c$	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW
1×1	764	139	97	7373	694	471						
2×2	302	77	49	2479	417	240						
4×4	91	40	31	781	156	132	5507	1040	548			
6×6	76	36	28	541	101	91	2799	591	374	∞	2641	1706
8×8	52	34	29	276	88	98	1881	383	294	∞	2506	1245
10×10	52	30	18	234	99	92	1455	317	257	∞	1909	1118

Table 7: Execution time (in seconds) on *Abisko* for `hessrand` matrices.

$p =$	$n = 4000$			$n = 8000$			$n = 16000$			$n = 32000$		
$p_r \times p_c$	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW
1×1	611	307	12	5021	2064	63						
2×2	302	202	7	1966	1458	29						
4×4	110	77	18	881	516	48	6671	1603	53			
6×6	96	61	21	578	339	70	4006	1034	52	∞	231	176
8×8	84	53	18	423	249	98	2822	605	53	∞	737	166
10×10	73	58	17	360	214	79	2456	553	56	∞	670	166

Table 8: Execution time (in seconds) on *Bellatrix* for `fullrand` matrices.

$p =$	$n = 4000$			$n = 8000$			$n = 16000$			$n = 32000$		
$p_r \times p_c$	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW
1×1	637	73	50	5377	441	252						
2×2	192	24	21	1594	137	91						
4×4	68	16	13	498	79	63	4505	552	271			
6×6	47	12	11	294	44	39	1886	247	165	∞	1267	901
8×8	36	16	12	204	42	37	1347	184	129	∞	1362	714
10×10	37	14	9	181	39	40	961	140	110	∞	726	525

Table 9: Execution time (in seconds) on *Bellatrix* for `hessrand` matrices.

$p = p_r \times p_c$	$n = 4000$			$n = 8000$			$n = 16000$			$n = 32000$		
	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW
1×1	510	174	6	4353	1102	26						
2×2	205	66	4	1590	530	11						
4×4	87	42	7	657	259	19	5416	808	22			
6×6	57	23	9	428	134	37	3054	380	22	∞	102	77
8×8	54	37	9	340	125	32	2227	365	22	∞	342	72
10×10	46	22	8	280	92	27	1846	214	24	∞	214	115

Table 10: Execution time (in seconds) on *Akka* of the new parallel multishift QR algorithm (NEW) for $100,000 \times 100,000$ matrices.

	$p = 16 \times 16$		$p = 24 \times 24$		$p = 32 \times 32$		$p = 40 \times 40$	
	`fullrand`	`hessrand`	`fullrand`	`hessrand`	`fullrand`	`hessrand`	`fullrand`	`hessrand`
Balancing	876	–	881	–	886	–	912	–
Hess. reduction	10084	–	6441	–	3868	–	2751	–
QR algorithm	13797	922	8055	1268	6646	5799	8631	1091
k_{AED}	35	19	31	19	27	18	23	18
k_{sweep}	5	0	6	0	13	0	12	0
#(shifts)/n	0.20	0	0.23	0	0.35	0	0.49	0
AED% in the QR alg.	48%	100%	43%	100%	39%	100%	54%	100%

bottleneck of the whole QR algorithm now. As reported in [21], it took 7 hours for our earlier implementation PDHSEQR to solve the $100,000 \times 100,000$ `fullrand` problem with 32×32 processors; 80% execution time of the QR algorithm was spent on AED. Our new version of PDHSEQR is able to solve the same problem in roughly 1.85 hours, which is about four times faster. The time fraction spent on AED is reduced to 39%.

5.3 Benchmark examples

Besides random matrices, we also report performance results for some commonly used benchmark matrices. For comparison, we have tested the same matrices as in [21], see Table 11. The execution times for the three solvers are listed in Tables 12–19. The conclusions are similar to those we have made for random matrices: our earlier version of PDHSEQR outperforms the ScaLAPACK 1.8.0 routine PDLAHQR by a large extent; the new PDHSEQR is usually even faster, especially for BBMSN and GRCAR.

In [21], it was observed that the accuracy for AF23560 is not fully satisfactory; the relative residuals $R_r = \|Q^T A Q - T\|_F / \|A\|_F$ were large for both PDLAHQR and PDHSEQR. It turns out that these large residuals are caused by an anomaly in PDLAHQR, which has been fixed by avoiding the use of PDLACONSB, see Section 2.3.2. As a result, the new PDHSEQR always produce $R_r \in [10^{-15}, 10^{-13}]$ for all test matrices.

6 Conclusions and Future Work

We have presented a new parallel implementation of the multishift QR algorithm with aggressive early deflation. The new routine PDHSEQR combines a number of techniques to improve serial performance and reduce communication. These include performing multiple levels of AED, reducing communication overhead by data redistribution, and refining the strategy for balancing between multishift QR sweeps and AED. Our numerical experiments provide compelling evidence that PDHSEQR significantly outperforms not only the original ScaLAPACK routine PDLAHQR but also an earlier version of PDHSEQR presented in [21]. In particular, our new implementation removes a bottleneck in the aggressive early deflation strategy by reducing communication and tuning algorithmic parameters. As a result, our new version is both faster and more robust. An intermediate version of the software presented in this paper is available in ScaLAPACK version 2.0.

Table 11: Benchmark matrices.

ID	Name	Dimension (n)	Type/Structure
1	BBMSN [13]	n	$S_n = \begin{bmatrix} n & n-1 & n-2 & \cdots & 2 & 1 \\ 10^{-3} & 1 & 0 & & 0 & 0 \\ & 10^{-3} & 2 & & 0 & 0 \\ & & 10^{-3} & & 0 & 0 \\ & & & \ddots & n-2 & 0 \\ & & & & 10^{-3} & n-1 \end{bmatrix}$
2	AF23560 [4]	23560	Computational fluid dynamics
3	CRY10000 [4]	10000	Material science
4	OLM5000 [4]	5000	Computational fluid dynamics
5	DW8192 [4]	8192	Electrical engineering
6	MATRAN [4]	n	Sparse random matrix
7	MATPDE [4]	n	Partial differential equations
8	GRCAR [4]	n	$G_n = \begin{bmatrix} 1 & 1 & 1 & 1 & & & & \\ -1 & 1 & 1 & 1 & 1 & & & \\ & -1 & 1 & 1 & 1 & 1 & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & -1 & 1 & 1 & 1 & 1 \\ & & & & -1 & 1 & 1 & 1 \\ & & & & & -1 & 1 & 1 \\ & & & & & & -1 & 1 \end{bmatrix}$

Table 12: Execution time (in seconds) for BBMSN.

n	$p = p_r \times p_c$	S-v180	SISC	NEW
5000	1×1	523	6	3
10000	2×2	1401	30	9
15000	3×3	1489	62	13

Table 13: Execution time (in seconds) for AF23560.

$p = p_r \times p_c$	S-v180	SISC	NEW
4×4	15486	2651	1375
6×6	9088	1279	826
8×8	6808	793	563
10×10	5694	662	475
12×12	5422	578	404

Table 14: Execution time (in seconds) for CRY10000.

$p = p_r \times p_c$	S-v180	SISC	NEW
1×1	6394	1331	1251
2×2	2123	580	495
4×4	979	236	209
6×6	731	161	132
8×8	545	128	95
10×10	496	144	90

Table 15: Execution time (in seconds) for OLM5000.

$p = p_r \times p_c$	S-v180	SISC	NEW
1×1	426	206	189
2×2	167	98	108
4×4	76	54	53
6×6	58	52	42
8×8	46	49	29
10×10	48	51	47

Table 16: Execution time (in seconds) for DW8192.

$p = p_r \times p_c$	S-v180	SISC	NEW
1×1	10307	1329	1297
2×2	1187	572	524
4×4	635	225	236
6×6	357	152	138
8×8	302	129	104
10×10	275	121	93

Table 17: Execution time (in seconds) for MATRAN.

$p =$	$n = 5000$			$n = 10000$			$n = 15000$		
$p_r \times p_c$	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW
1×1	1617	332	218	∞	1756	1137			
2×2	579	152	122	4495	931	471			
4×4	247	74	60	1555	321	268	5122	937	575
6×6	178	64	57	1035	207	170	3046	535	382
8×8	147	58	50	746	153	157	2166	390	315
10×10	149	59	43	615	169	140	1669	362	264

Table 18: Execution time (in seconds) for MATPDE.

$p =$	$n = 10000$			$n = 14400$			$n = 19600$		
$p_r \times p_c$	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW
1×1	12429	2600	1699						
2×2	3531	1081	966						
4×4	1565	415	361	4446	1207	1027	9915	2844	2406
6×6	1118	256	225	3069	654	573	6130	1426	1284
8×8	871	189	156	2259	449	384	4615	912	802
10×10	789	189	137	1955	431	313	4046	743	628
12×12	719	194	126	1736	367	260	3483	648	504

Table 19: Execution time (in seconds) for GRCAR.

$p =$	$n = 6000$			$n = 12000$			$n = 18000$		
$p_r \times p_c$	S-v180	SISC	NEW	S-v180	SISC	NEW	S-v180	SISC	NEW
1×1	2738	1340	69						
2×2	850	645	40	8199	2734	132			
4×4	363	258	226	2499	1336	114	8171	4037	182
6×6	244	190	173	1471	849	112	4385	2172	187
8×8	217	150	145	1107	515	142	3342	1345	175
10×10	207	161	126	923	538	338	2675	1104	276

Concerning future work, we believe to have come to a point, where it will be difficult to attain further dramatic performance improvements for parallel nonsymmetric eigensolvers, without leaving the classical framework of QR algorithms. Considering the fact that the execution times spent on Hessenberg reduction and on QR iterations are now nearly on the same level, any further improvement of the iterative part will only have a limited impact on the total execution time. The situation is quite different when shared memory many-core processors with accelerators, such as GPUs, are considered. Although efficient implementations of the Hessenberg reduction on such architectures have recently been proposed [31, 29, 36], the iterative part remains to be done. Another future challenge is to combine the message passing paradigm used in this new implementation of the multishift QR algorithm and dynamic and static scheduling on many-core nodes using multithreading.

Acknowledgements

We are grateful to Björn Adlerborn and Lars Karlsson for constructive discussions and comments on the subject, and to Åke Sandgren for support at HPC2N. We also thank David Guerrero, Rodney James, Julien Langou, Jack Poulson, Jose Roman, as well as anonymous users from IBM for helpful feedback.

References

[1] E. Agullo, J. W. Demmel, J. J. Dongarra, B. Hadri, J. Kurzak, J. Langou, H. Ltaief, P. Luszczek, and S. Tomov. Numerical linear algebra on emerging architectures: The PLASMA and MAGMA projects. *Journal of Physics: Conference Series*, 180(1):012037, 2009.

[2] M. Ahues and F. Tisseur. A new deflation criterion for the QR algorithm. LAPACK Working Note 122, 1997.

[3] T. Auckenthaler, V. Blum, H.-J. Bungartz, T. Huckle, R. Johanni, L. Kraemer, B. Lang, H. Lederer, and P. R. Willems. Parallel solution of partial symmetric eigenvalue problems from electronic structure calculations. *Parallel Comput.*, 37(12):783–794, 2011.

[4] Z. Bai, D. Day, J. W. Demmel, and J. J. Dongarra. A test matrix collection for non-Hermitian eigenvalue problems (release 1.0). Technical Report CS-97-355, Department of Computer Science, University of Tennessee, 1997. Also available online from http://math.nist.gov/MatrixMarket.

[5] Z. Bai and J. W. Demmel. On a block implementation of Hessenberg multishift QR iteration. *Intl. J. High Speed Comput.*, 1:97–112, 1989.

[6] Z. Bai and J. W. Demmel. On swapping diagonal blocks in real Schur form. *Linear Algebra Appl.*, 186:73–95, 1993.

[7] Z. Bai, J. W. Demmel, J. J. Dongarra, A. Petitet, H. Robison, and K. Stanley. The spectral decomposition of nonsymmetric matrices on distributed memory parallel computers. *SIAM J. Sci. Comput.*, 18(5):1446–1461, 1997.

[8] G. Ballard, J. W. Demmel, and I. Dumitriu. Minimizing communication for eigenproblems and the singular value decomposition. Technical Report UCB/EECS-2010-136, EECS Department, University of California, Berkeley, 2010. Also as LAPACK Working Note 237.

[9] G. Ballard, J. W. Demmel, O. Holtz, and O. Schwartz. Minimizing communication in numerical linear algebra. *SIAM J. Matrix Anal. Appl.*, 32(3):866–901, 2011.

[10] P. Bientinesi, E. S. Quintana-Ortí, and R. A. van de Geijn. Representing linear algebra algorithms in code: The FLAME application program interfaces. *ACM Trans. Math. Software*, 31(1):27–59, 2005.

[11] L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. W. Demmel, I. Dhillon, J. J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley. *ScaLAPACK User's Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.

[12] K. Braman, R. Byers, and R. Mathias. The multishift QR algorithm. Part I: Maintaining well-focused shifts and level 3 performance. *SIAM J. Matrix Anal. Appl.*, 23(4):929–947, 2002.

[13] K. Braman, R. Byers, and R. Mathias. The multishift QR algorithm. Part II: Aggressive early deflation. *SIAM J. Matrix Anal. Appl.*, 23(4):948–973, 2002.

[14] R. Byers. LAPACK 3.1 xHSEQR: Tuning and implementation notes on the small bulge multi-shift QR algorithm with aggressive early deflation, 2007. LAPACK Working Note 187.

[15] EPFL. Bellatrix, 2013. URL: `http://hpc.epfl.ch/clusters/bellatrix/`.

[16] M. R. Fahey. Algorithm 826: A parallel eigenvalue routine for complex Hessenberg matrices. *ACM Trans. Math. Software*, 29(3):326–336, 2003.

[17] J. G. F. Francis. The QR transformation: A unitary analogue to the LR transformation — Part 2. *Comput. J.*, 4(4):332–345, 1962.

[18] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

[19] A. Grama, A. Gupta, G. Karypis, and V. Kumar. *Introduction to Parallel Computing*. Addison-Wesley, Boston, MA, USA, second edition, 2003.

[20] R. Granat, B. Kågström, and D. Kressner. Parallel eigenvalue reordering in real Schur forms. *Concurrency and Computat.: Pract. Exper.*, 21(9):1225–1250, 2009.

[21] R. Granat, B. Kågström, and D. Kressner. A novel parallel QR algorithm for hybrid distributed memory HPC systems. *SIAM J. Sci. Comput.*, 32(4):2345–2378, 2010.

[22] R. Granat, B. Kågström, D. Kressner, and M. Shao. PDHSEQR User's Guide, 2014. Available from `http://www8.cs.umu.se/~myshao/software/pdhseqr/ug.pdf`.

[23] G. Henry and R. A. van de Geijn. Parallelizing the QR algorithm for the unsymmetric algebraic eigenvalue problem: Myths and reality. *SIAM J. Sci. Comput.*, 17(4):870–883, 1996.

[24] G. Henry, D. S. Watkins, and J. J. Dongarra. A parallel implementation of the nonsymmetric QR algorithm for distributed memory architectures. *SIAM J. Sci. Comput.*, 24(1):284–311, 2002.

[25] M. Hoemmen. *Communication-Avoiding Krylov Subspace Methods*. PhD thesis, University of California, Berkeley, 2010.

[26] HPC2N. Abisko, 2013. URL: `http://www.hpc2n.umu.se/resources/abisko/`.

[27] HPC2N. Akka, 2013. URL: `http://www.hpc2n.umu.se/resources/akka/`.

[28] D. Irony, S. Toledo, and A. Tiskin. Communication lower bounds for distributed-memory matrix multiplication. *J. Parallel Distr. Comput.*, 64(9):1017–1026, 2004.

[29] B. Kågström, D. Kressner, E. S. Quintana-Ortí, and G. Quintana-Ortí. Blocked algorithms for the reduction to Hessenberg-triangular form revisited. *BIT*, 48(3):563–584, 2008.

[30] B. Kågström, D. Kressner, and M. Shao. On aggressive early deflation in parallel variants of the QR algorithm. In K. Jónasson, editor, *Applied Parallel and Scientific Computing (PARA 2010)*, volume 7133 of *Lecture Notes in Comput. Sci.*, pages 1–10, Berlin, Germany, 2012. Springer-Verlag.

[31] L. Karlsson and B. Kågström. Parallel two-stage reduction to Hessenberg form using dynamic scheduling on shared-memory architectures. *Parallel Comput.*, 37(12):771–782, 2011.

[32] R. S. Martin, G. Peters, and J. H. Wilkinson. Handbook Series Linear Algebra: The QR algorithm for real Hessenberg matrices. *Numer. Math.*, 14(3):219–231, 1970.

[33] T. Schreiber, P. Otto, and F. Hofmann. A new efficient parallelization strategy for the QR algorithm. *Parallel Comput.*, 20(1):63–75, 1994.

[34] M. Shao. `PDLAQR1`: An improved version of the ScaLAPACK routine `PDLAHQR`. Technical Report UMINF-11.22, Department of Computing Science, Umeå University, 2011. Available from `http://www8.cs.umu.se/research/uminf/`.

[35] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler. *Matrix Eigensystem Routines — EISPACK Guide*, volume 6 of *Lecture Notes in Comput. Sci.* Springer-Verlag, New York, second edition, 1976.

[36] S. Tomov, R. Nath, and J. J. Dongarra. Accelerating the reduction to upper Hessenberg, tridiagonal, and bidiagonal forms through hybrid GPU-based computing. *Parallel Comput.*, 36(12):645–654, 2010.

[37] R. A. van de Geijn. Storage schemes for parallel eigenvalue algorithms. In G. H. Golub and P. Van Dooren, editors, *Numerical Linear Algebra Digital Signal Processing and Parallel Algorithms*, pages 639–648. Springer-Verlag, 1988.

[38] R. A. van de Geijn. Deferred shifting schemes for parallel QR methods. *SIAM J. Matrix Anal. Appl.*, 14:180–194, 1993.

[39] R. A. van de Geijn and D. G. Hudson. An efficient parallel implementation of the nonsymmetric QR algorithm. In *Fourth Conference on Hypercube Concurrent Computers and Applications, Monterey, CA*, pages 697–700, 1989.

[40] D. S. Watkins. Shifting strategies for the parallel QR algorithm. *SIAM J. Sci. Comput.*, 15(4):953–958, 1994.

Parallel Library Software for the Multishift QR Algorithm with Aggressive Early Deflation —Electronic Appendix: Derivation of the Performance Model

Robert Granat*, Bo Kågström†, Daniel Kressner†, and Meiyue Shao†,‡

1 Estimating T_{sweep}

The QR sweep is relatively simple because the computation and communication cost is well-determined by n, n_{shift}, and p. Usually there are up to $\sqrt{p}$ simultaneous computational windows, one at each diagonal processor in the grid, with at most $n_b/3$ shifts in each window. If $n_{\text{shift}} > \sqrt{p}\, n_b/3$, these shifts are chased in several rounds. So we use a rough approximation $n^*_{\text{shift}} = \sqrt{p}\, n_b/3$ to represent the total amount of shifts which can be chased simultaneously in the QR sweep. Based on the assumption $\sqrt{p}\, n_b \ll n$, the overhead for the start-up and ending phases of the bulge chasing are not important. Therefore the cost of one QR sweep is roughly

$$T_{\text{sweep}}(n, n_{\text{shift}}, p) = \frac{n_{\text{shift}} n}{n^*_{\text{shift}} n_b}(T_{\text{local}} + T_{\text{cross}}),$$

where T_{local} and T_{cross} represent the runtime for local and crossborder bulge chasing, respectively. Both parts require chasing the chain of bulges with $n_b/2$ steps inside the computational window, as well as updating the corresponding off-diagonal blocks. Hence the runtime for arithmetic operations is $\left(4n_b^3 + 4n\, n_b^2/\sqrt{p}\right)\gamma$, half of which is for accumulating the orthogonal matrix Q. The only communication cost in the local chasing phase is broadcasting the accumulated orthogonal matrix rowwise and columnwise in the processor grid, which requires $\log_2 p\left(\alpha + n_b^2\beta\right)$ runtime, i.e.,

$$T_{\text{local}} = \left(4n_b^3 + \frac{4n\, n_b^2}{\sqrt{p}}\right)\gamma + \log_2 p\left(\alpha + n_b^2\beta\right) \approx \frac{4n\, n_b^2}{\sqrt{p}}\gamma + \log_2 p\left(\alpha + n_b^2\beta\right).$$

One round crossborder chasing requires at least the same amount of communication as in one local chasing step, with some extra cost for explicitly forming the $n_b \times n_b$ computational window and exchanging data with processor neighbours for updating the off-diagonal blocks. Notice that usually there are two rounds for a crossborder chasing step, therefore we have

$$T_{\text{cross}} = 2\left[T_{\text{local}} + 3\left(\alpha + \frac{n_b^2}{4}\beta\right) + 3\left(\alpha + \frac{n n_b}{2\sqrt{p}}\beta\right)\right],$$

and then

$$\begin{aligned} T_{\text{sweep}}(n, n_{\text{shift}}, p) &\approx \frac{12n^2 n_{\text{shift}} n_b}{\sqrt{p}\, n^*_{\text{shift}}}\gamma + \frac{3n\, n_{\text{shift}}}{n_b n^*_{\text{shift}}}(\log_2 p + 4)\alpha + \frac{3n^2 n_{\text{shift}}}{\sqrt{p}\, n^*_{\text{shift}}}\beta \\ &= \frac{36n^2 n_{\text{shift}}}{p}\gamma + \frac{9n\, n_{\text{shift}}}{\sqrt{p}\, n_b^2}(\log_2 p + 4)\alpha + \frac{9n^2 n_{\text{shift}}}{p n_b}\beta. \end{aligned}$$

*Department of Computing Science and HPC2N, Umeå University, SE-901 87 Umeå, Sweden. Email: {granat, bokg, myshao}@cs.umu.se.

†MATHICSE, EPF Lausanne, CH-1015 Lausanne, Switzerland. Email: daniel.kressner@epfl.ch.

From this model, we can see that the cost for updating the off-diagonal blocks dominates in both the computation and communication parts, under the assumption that $\sqrt{p}\,n_b \ll n$ (or equivalently $n^*_{\text{shift}} \ll n$). As a byproduct, the performance model of a plain multishift QR algorithm without AED can also be obtained. By assuming the convergence rate as $\Theta(1)$ shifts per eigenvalue, i.e. $k_{\text{sweep}} = \Theta(n/n_{\text{shift}})$ and neglecting the cost for generating shifts, the total execution time of a plain multishift QR algorithm is

$$T_{\text{new}}(n,p) = \Theta\Big(\frac{n^3}{p}\Big)\gamma + \Theta\Big(\frac{n^2\log p}{\sqrt{p}\,n_b^2}\Big)\alpha + \Theta\Big(\frac{n^3}{pn_b}\Big)\beta.$$

Fixing the memory load per processor (i.e., $n/\sqrt{p}$ = constant) yields

$$T_{\text{new}}(n,p) = \Theta(n)\,\gamma + \Theta(n\log n)\,\alpha + \Theta(n)\,\beta.$$

2 Estimating T_{AED} and T_{shift}

The execution time for one step AED is modelled as

$$\begin{aligned} T_{\text{AED}}(n, n_{\text{AED}}, p) = {} & T_{\text{redist}}(n_{\text{AED}}, p, p_{\text{AED}}) + T_{\text{pipe}}(n_{\text{AED}}, p_{\text{AED}}) \\ & + T_{\text{reorder}}(n_{\text{AED}}, p) + T_{\text{Hess}}(n_{\text{AED}}, p) + T_{\text{update}}(n, n_{\text{AED}}, p), \end{aligned}$$

where the terms in the right-hand-side represent the runtime for data redistribution, Schur decomposition of the AED window, deflation checking and reordering of eigenvalues, Hessenberg reduction, and updating the off-diagonal blocks corresponding to the AED window, respectively. We estimate these terms one by one using the hierarchical approach in [3].

- T_{redist}: The general-purpose data redistribution routine `PDGEMR2D` in ScaLAPACK uses the algorithm described in [7]. Since the scheduling part is tiny compared to the communication part, the complexity of data redistribution is provided [7] as

$$T_{\text{redist}}(n_{\text{AED}}, p, p_{\text{AED}}) = \Theta(p)\,\alpha + \Theta\Big(\frac{n_{\text{AED}}^2}{\sqrt{p\,p_{\text{AED}}}}\Big)\beta.$$

- T_{pipe}: The complexity of the Schur decomposition performed by `PDLAQR1` largely depends on the property of the matrix, since AED affects the convergence rate significantly. To obtain an estimate of the complexity, we assume that AED roughly reduces the number of pipelined QR sweeps by half. According to the experimental results presented in [6], this assumption usually provides a reasonable upper bound of the runtime, although it can be overestimated. Using the model in [5], we obtain an approximate execution time

$$\begin{aligned} T_{\text{pipe}}(n,p) &= \frac{20n^3}{p}\gamma + \frac{3n^2}{\sqrt{p}\,n_b}(\log_2 p + 2)\alpha + \Big(\frac{3n^2\log_2 p}{\sqrt{p}} + \frac{8n^3}{pn_b}\Big)\beta \\ &= \Theta\Big(\frac{n^3}{p}\Big)\gamma + \Theta\Big(\frac{n^2\log p}{\sqrt{p}\,n_b}\Big)\alpha + \Theta\Big(\frac{n^3}{pn_b}\Big)\beta. \end{aligned} \tag{1}$$

  If the orthogonal matrix Q is not accumulated in the calculation, the arithmetic operations are roughly halved, i.e.,

$$\widetilde{T}_{\text{pipe}}(n,p) = \frac{10n^3}{p}\gamma + \frac{3n^2}{\sqrt{p}\,n_b}(\log_2 p + 2)\alpha + \Big(\frac{3n^2\log_2 p}{\sqrt{p}} + \frac{8n^3}{pn_b}\Big)\beta.$$

  The model provided in [1] is similar, but with slightly different coefficients.

- T_{reorder}: Obviously, the cost for eigenvalue reordering depends on the deflation ratio. However, we can evaluate an upper bound for the cost—all eigenvalues are involved in the reordering. Then the performance model is almost the same as that of QR sweeps, since updating the off-diagonal blocks is the dominant operation. Notice that each eigenvalue needs to move $n_{\text{AED}}/2$ steps in average, so the overall cost for eigenvalue reordering inside the AED window is bounded by

$$T_{\text{reorder}}(n_{\text{AED}}, p) \approx \frac{4n_{\text{AED}}^2 n_b}{\sqrt{p}}\gamma + \frac{2n_{\text{AED}}}{n_b}(\log_2 p + 3)\alpha + \frac{3n_{\text{AED}}^2}{2\sqrt{p}}\beta.$$

  As a different feature compared to QR sweeps or the performance model in [4] for parallel eigenvalue reordering, the degree of concurrency here is $\Theta(\sqrt{p})$ instead of $\Theta(p)$ since usually there are at most two computational windows for the reordering phase inside the AED window.

- T_{Hess}: The Hessenberg reduction routine `PDGEHRD` uses the parallel algorithm described in [2]. Almost all computations and communication are performed on matrix-vector and matrix-matrix multiplications. Therefore we need to model these PBLAS operations first. The level 2 operations GEMV and GER require

$$T_{\text{GEMV}}(m, n, p) \approx T_{\text{GER}}(m, n, p) \approx \frac{2mn}{p}\gamma + \log_2 p\left(\alpha + \frac{m+n}{2\sqrt{p}}\beta\right),$$

  where $m \times n$ is the size of the matrix. This model can be directly generalized to multiplying two $m \times k$ and $k \times n$ matrices as long as $\min\{m, n, k\} \le n_b$ since it is merely a "fat" level 2 operation. In the Hessenberg reduction algorithm, all level 3 operations are "fat" level 2 operations, so the cost for one GEMM operation can be modelled as

$$T_{\text{GEMM}}(m, n, n_b, p) \approx T_{\text{GEMM}}(m, n_b, n, p) \approx \frac{2mnn_b}{p}\gamma + \log_2 p\left(\alpha + \frac{(m+n)n_b}{2\sqrt{p}}\beta\right). \tag{2}$$

  Using these simple models of PBLAS operations, we are able to establish a model for T_{Hess}. The level 2 part consists roughly of n matrix-vector multiplications of dimension $n \times (n - j)$ (for $j = 1, 2, \ldots, n$). Therefore the cost is

$$T_{\text{level2}} = \sum_{j=1}^{n}\left[\frac{2n(n-j)}{p}\gamma + \log_2 p\left(\alpha + \frac{2n-j}{2\sqrt{p}}\right)\right] \approx \frac{n^3}{p}\gamma + \log_2 p\left(n\alpha + \frac{3n^2}{4\sqrt{p}}\beta\right).$$

  The level 3 part contains roughly n/n_b iterations with one `PDGEMM` and one `PDLARFB` per iteration. Within the jth iteration ($j = 1, 2, \ldots, n/n_b$), `PDGEMM` involves matrices of dimension $n \times n_b$ and $n_b \times (n - jn_b - n_b)$; `PDLARFB` mainly performs two parallel GEMM operations, with $\{\, n_b \times (n - jn_b), (n - jn_b) \times (n - jn_b) \,\}$ and $\{\, (n - jn_b) \times n_b, n_b \times (n - jn_b) \,\}$ matrices involved. Another sequential TRMM operation in `PDLARFB` is neglected since it only contributes lower order terms in both arithmetic and communication costs. So the cost for level 3 part is

$$\begin{aligned} T_{\text{level3}} &= \sum_{j=1}^{n/n_b}\left[\frac{2jn_b + 6(n - jn_b)}{p}n_b(n - jn_b)\gamma + \log_2 p\left(3\alpha + \frac{6n - 5jn_b}{2\sqrt{p}}\beta\right)\right] \\ &\approx \frac{7n^3}{3p}\gamma + \frac{3n\log_2 p}{n_b}\alpha + \frac{7n^2\log_2 p}{4\sqrt{p}}\beta, \end{aligned}$$

  and hence the execution time for Hessenberg reduction (without explicitly forming the orthogonal matrix) is

$$\widetilde{T}_{\text{Hess}}(n, p) = T_{\text{level2}} + T_{\text{level3}} \approx \frac{10n^3}{3p}\gamma + n\log_2 p\,\alpha + \frac{5n^2\log_2 p}{2\sqrt{p}}\beta. \tag{3}$$

  Even if the proportion of level 3 operations is improved to 80% as suggested in [8] but not implemented in the current `PDGEHRD` yet, the estimate in (3) would not change too much since the number of messages in the level 2 part is not reduced.

Since the Householder reflections are stored in a compact form in the lower triangular part of the upper Hessenberg matrix, formulating the orthogonal matrix after Hessenberg reduction is another necessary step. This step is done by the ScaLAPACK routine `PDORMHR`, which is mainly a series of calls to `PDLARFB`. Similar to the discussion above, we obtain

$$T_{\text{ORMHR}} \approx \frac{2n^3}{p}\gamma + \frac{3n\log_2 p}{n_b}\alpha + \frac{7n^2}{4\sqrt{p}}\beta.$$

Therefore the total runtime for the Hessenberg reduction process including formulating the orthogonal matrix is

$$T_{\text{Hess}}(n,p) = \widetilde{T}_{\text{Hess}} + T_{\text{ORMHR}} \approx \frac{16n^3}{3p}\gamma + n\log_2 p\,\alpha + \frac{17n^2\log_2 p}{4\sqrt{p}}\beta. \tag{4}$$

- T_{update}: The cost for updating the off-diagonal blocks with respect to the AED window is simple to analyze since it merely contains three GEMM operations. Since these GEMM operations are not "fat" level 2 operations, we need to use a model different to (2). According to [9], the execution time for a GEMM operation on a $\sqrt{p}\times\sqrt{p}$ processor grid with $m\times k$ and $k\times n$ matrices involved is

$$T_{\text{GEMM}}(m,n,k,p) \approx \frac{2mnk}{p}\gamma + \left(\frac{k}{n_b} + 2\sqrt{p}\right)\left(2\alpha + \frac{(m+n)n_b}{\sqrt{p}}\beta\right)$$

if $\min\{m,n,k\} = k \gg n_b$. Then we conclude that

$$T_{\text{update}}(n,n_{\text{AED}},p) \approx \frac{2nn_{\text{AED}}^2}{p}\gamma + \frac{n_{\text{AED}}}{n_b}\left(6\alpha + \frac{2nn_b}{\sqrt{p}}\beta\right).$$

Now we are ready to estimate the overall runtime T_{AED} by substituting n with n_{AED} in (1) and (4). We can see that T_{redist} is always negligible compared to other components. Reordering contributes with only marginal communication costs also. By merging all these estimates together, we eventually obtain

$$\begin{aligned}
T_{\text{AED}}(n,n_{\text{AED}},p) &\approx T_{\text{pipe}}(n_{\text{AED}},p_{\text{AED}}) + T_{\text{reorder}}(n_{\text{AED}},p) + T_{\text{Hess}}(n_{\text{AED}},p) + T_{\text{update}}(n,n_{\text{AED}},p)\\
&\approx \left(\frac{20n_{\text{AED}}}{p_{\text{AED}}} + \frac{4\sqrt{p}\,n_b + 16n_{\text{AED}} + 2n}{p}\right)n_{\text{AED}}^2\gamma\\
&\quad + \frac{n_{\text{AED}}^2}{n_b}\left(\frac{3(\log_2 p_{\text{AED}} + 2)}{\sqrt{p_{\text{AED}}}} + \frac{n_b\log_2 p}{n_{\text{AED}}}\right)\alpha\\
&\quad + \frac{n_{\text{AED}}^2}{n_b}\left(\frac{3n_b\log_2 p_{\text{AED}}}{\sqrt{p_{\text{AED}}}} + \frac{8n_{\text{AED}}}{p_{\text{AED}}} + \frac{3n_b}{2\sqrt{p}} + \frac{17n_b\log_2 p}{4\sqrt{p}} + \frac{2nn_b}{n_{\text{AED}}\sqrt{p}}\right)\beta\\
&\approx \left[\frac{30C_2^2 n}{C_1} + \frac{9n^2 n_b}{C_1^2\sqrt{p}} + \frac{9(C_1+6)n^3}{2C_1^3 p}\right]\gamma\\
&\quad + \left(\frac{9C_2 n}{C_1 n_b}\log_2\frac{3n}{2C_1C_2} + \frac{3n}{2C_1}\log_2 p\right)\alpha\\
&\quad + \left[\frac{9C_2 n}{C_1}\log_2\frac{3n}{2C_1C_2} + \frac{12C_2^2 n}{C_1 n_b} + \frac{3n^2(18 + C_1 + 51\log_2 p)}{16C_1^2\sqrt{p}}\right]\beta.
\end{aligned}$$

When n is extremely large (i.e., C_1, C_2 and n_b are all tiny enough compared to n) and $n/\sqrt{p}$ = constant, we have

$$\begin{aligned}
T_{\text{AED}} &= \Theta\left(n + \frac{n^2}{\sqrt{p}} + \frac{n^3}{p}\right)\gamma + \Theta(n\log n + n\log p)\,\alpha + \Theta\left(n\log n + \frac{n^2}{\sqrt{p}}\log p\right)\beta\\
&= \Theta(n)\,\gamma + \Theta(n\log n)\,\alpha + \Theta(n\log n)\,\beta.
\end{aligned}$$

Asymptotically AED only has slightly larger message sizes by a $\Theta(\log n)$ factor compared to QR sweeps and is hence not much more expensive. However, in practice we still need to handle AED very carefully since large leading factors in lower order terms have significant impact on the performance if the matrix is not large enough. Similar to the analysis for T_{AED}, the cost for computing shifts can be estimated by

$$\begin{aligned}
T_{\text{shift}}(n, n_{\text{shift}}, p) &\approx \widetilde{T}_{\text{pipe}}(n_{\text{shift}}, p_{\text{shift}}) \\
&\approx \frac{10n_{\text{shift}}^3}{p_{\text{shift}}}\gamma + \frac{3n_{\text{shift}}^2}{\sqrt{p_{\text{shift}}}\,n_b}(\log_2 p_{\text{shift}} + 2)\,\alpha + \Big(\frac{3n_{\text{shift}}^2 \log_2 p_{\text{shift}}}{\sqrt{p_{\text{shift}}}} + \frac{8n_{\text{shift}}^3}{p_{\text{shift}} n_b}\Big)\beta \\
&\approx \frac{10C_2^2 n}{C_1}\gamma + \frac{6C_2 n}{C_1 n_b}\log_2 \frac{n}{C_1 C_2}\,\alpha + \Big(\frac{6C_2 n}{C_1}\log_2 \frac{n}{C_1 C_2} + \frac{8C_2^2 n}{C_1 n_b}\Big)\beta.
\end{aligned}$$

Asymptotically T_{shift} is not so important in the scalability analysis since it can never be larger than T_{AED}.

References

[1] L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. W. Demmel, I. Dhillon, J. J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley. *ScaLAPACK User's Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.

[2] J. Choi, J. J. Dongarra, and D. W. Walker. The design of a parallel dense linear algebra software library: Reduction to Hessenberg, tridiagonal, and bidiagonal form. *Numer. Algorithms*, 10(2):379–399, 1995.

[3] K. Dackland and B. Kågström. An hierarchical approach for performance analysis of ScaLAPACK-based routines using the distributed linear algebra machine. In J. Waśniewski, J. J. Dongarra, K. Madsen, and D. Olesen, editors, *Applied Parallel Computing Industrial Computation and Optimization (PARA 1996)*, volume 1184 of *Lecture Notes in Comput. Sci.*, pages 186–195, Heidelberg, Germany, 1996. Springer-Verlag.

[4] R. Granat, B. Kågström, and D. Kressner. Parallel eigenvalue reordering in real Schur forms. *Concurrency and Computat.: Pract. Exper.*, 21(9):1225–1250, 2009.

[5] G. Henry, D. S. Watkins, and J. J. Dongarra. A parallel implementation of the nonsymmetric QR algorithm for distributed memory architectures. *SIAM J. Sci. Comput.*, 24(1):284–311, 2002.

[6] B. Kågström, D. Kressner, and M. Shao. On aggressive early deflation in parallel variants of the QR algorithm. In K. Jónasson, editor, *Applied Parallel and Scientific Computing (PARA 2010)*, volume 7133 of *Lecture Notes in Comput. Sci.*, pages 1–10, Berlin, Germany, 2012. Springer-Verlag.

[7] L. Prylli and B. Tourancheau. Fast runtime block cyclic data redistribution on multiprocessors. *J. Parallel Distr. Comput.*, 45(1):63–72, 1997.

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

[9] R. A. van de Geijn and J. Watts. SUMMA: Scalable universal matrix multiplication algorithm. *Concurrency and Computat.: Pract. Exper.*, 9(4):255–274, 1997. Also as LAPACK Working Note 96.

Recent publications:

MATHEMATICS INSTITUTE OF COMPUTATIONAL SCIENCE AND ENGINEERING
Section of Mathematics
Ecole Polytechnique Fédérale
CH-1015 Lausanne

18.2013 A. ABDULLE, Y. BAI, G. VILMART:
An online-offline homogenization strategy to solve quasilinear two-scale problems at the cost of one-scale problems

19.2013 C.M. COLCIAGO, S. DEPARIS, A. QUARTERONI:
Comparison between reduced order models and full 3D models for fluid-structure interaction problems in haemodynamics

20.2013 D. KRESSNER, M. STEINLECHNER, B. VANDEREYCKEN:
Low-rank tensor completion by Riemannian optimization

21.2013 M. KAROW, D. KRESSNER, E. MENGI:
Nonlinear eigenvalue problems with specified eigenvalues

22.2013 T. LASSILA, A. MANZONI, A. QUARTERONI, G. ROZZA:
Model order reduction in fluid dynamics: challenges and perspectives

23.2013 M. DISCACCIATI, P. GERVASIO, A. QUARTERONI:
The interface control domain decomposition (ICDD) method for the Stokes problem

24.2013 V. LEVER, G. PORTA, L. TAMELLINI, M. RIVA:
Characterization of basin-scale systems under mechanical and geochemical compaction

25.2013 D. DEVAUD, A. MANZONI, G. ROZZA:
A combination between the reduced basis method and the ANOVA expansion: on the computation of sensitivity indices

26.2013 M. SHAO:
On the finite section method for computing exponentials of doubly-infinite skew-Hermitian matrices

27.2013 A. ABDULLE, G. VILMART, K. C. ZYGALAKIS:
High order numerical approximation of the invariant measure of ergodic SDEs

28.2013 S. ROSSI, T. LASSILA, R. RUIZ-BAIER, A. SEQUEIRA, A. QUARTERONI:
Thermodynamically consistent orthotropic activation model capturing ventricular systolic wall thickening in cardiac electromechanics

29.2013 F. BONIZZONI, F. NOBILE:
Perturbation analysis for the Darcy problem with log-normal permeability

30.2013 Z. LI, A. USCHMAJEW, S. ZHANG:
On convergence of the maximum block improvement method

31.2013 NEW R. GRANAT, B. KAGSTRÖM, D. KRESSNER, M. SHAO:
Parallel library software for the multishift QR algorithm with aggressive early deflation