Blocking Techniques in Symmetric Eigensolvers *

Wilfried N. Gansterer† Dieter F. Kvasnicka‡ Karlheinz Schwarz‡
Christoph W. Ueberhuber†

Abstract

Blocking the tridiagonalization part of symmetric eigensolvers enables a significant efficiency enhancement. Various steps of a two-sweep tridiagonalization can be blocked independently of each other. Thus, a great deal of parallelism becomes available and Level 3 BLAS can be utilized in the reduction phase. A successful application of blocked eigensolvers in material science calculations is presented.

1 Problem Specification

Symmetric eigensolvers play an important role in many scientific applications (see, for example, Schwarz, Blaha [27]). General information about the numerical solution of eigenproblems is found, for example, in Golub, Van Loan [20], Parlett [26], or Wilkinson [29].

Allan and Bush [1] present an overview over parallel diagonalization routines and compare advantages and disadvantages of the implemented methods. The majority of routines is available for real dense symmetric matrices only, although also Hermitian or sparse matrices are considered. The algorithms are of one of three kinds.

- Reduction based methods, which perform a Householder reduction to tridiagonal form and then solve the tridiagonal eigenproblem, for example using QL or QR iteration, bisection and inverse iteration, or divide and conquer methods;
- One-sided or two-sided Jacobi methods (see, for example, Hegland [22]);
- Lanczos methods (Cullum, Willoughby [12]), Arnoldi methods (Lehoucq et al. [24]), and subspace decomposition (Auslander, Tsao [3]).

In terms of arithmetic complexity, i.e., the number of floating-point operations required, reduction based methods are usually most efficient for dense matrices. In a parallel environment, however, they show limited scalability. Therefore the following recommendations are made in Allan and Bush [1].

- For an $n \times n$ matrix on P processors, where $n/P < 8$, they achieved best results with a one-sided Jacobi method.
- For matrices with $n \leq 3000$ and $n/P > 8$, reduction based methods performed best.

*The work described in this paper was supported by the Special Research Program SFB F011 "AURORA" of the Austrian Science Fund.

†Institute for Applied and Numerical Mathematics, Vienna University of Technology, Austria.
‡Institute for Physical and Theoretical Chemistry, Vienna University of Technology, Austria.

- For matrices with $n > 3000$ and $n/P > 8$, they recommend a two-sided Jacobi method.

This paper discusses techniques which make reduction based methods applicable for a wider range of parameters (i) by improving sequential execution time and (ii) by improving scalability.

1.1 The Symmetric Eigenproblem

This paper concentrates on the efficient computation of a subset of the eigenpairs (eigenvalues and eigenvectors) of a dense symmetric (Hermitian) matrix $A \in \mathbb{R}^{n\times n}$ ($\mathbb{C}^{n\times n}$), i. e., the solution of the symmetric (Hermitian) eigenproblem $AX = X\Lambda$. The efficiency of the method described strongly depends on the number m of the eigenvectors computed. If only eigenvalues have to be computed, efficiency is very high. If only some eigenvectors have to be computed the method is still competitive. If all eigenvectors have to be computed, efficiency may decrease significantly, but the method presented may still be a good choice on certain computers (see Gansterer et al. [19]).

The method used has three fundamental stages.

- Reduce matrix A to a band matrix $B = Q_b^\top A Q_b$. Matrix B has typically a small bandwidth b (for example, $b = 8$).
- Solve the reduced eigenproblem $BX_b = X_b\Lambda$.
- Transform the eigenvector matrix X_b of the reduced problem back to the eigenvector matrix $X = Q_b X_b$ of the original problem.

The banded eigenproblem is solved in another three stages. Stage 1: Reduce $BX_b = X_b\Lambda$ to a tridiagonal eigenproblem $TX_t = X_t\Lambda$. Stage 2: Compute the eigenvalues of the tridiagonal eigenproblem using bisection. Stage 3: Using these eigenvalues, compute the eigenvectors X_b of matrix B using inverse iteration.

Compared to eigensolvers using a one-sweep tridiagonal reduction (for example, the routine `LAPACK/dsyevx`), which perform 50 % of their floating point operations in Level 2 Blas, the first band reduction of two-sweep algorithms can be implemented utilizing Level 3 Blas exclusively. High level Blas (i) are advantageous to exploit memory hierarchies on sequential processors, and (ii) lead to better scaling behavior in parallel implementations.

2 Existing Work

If all eigenvalues are required, the solution of the tridiagonal eigenproblem often dominates execution time. The tridiagonal eigenproblem has been worked on extensively (see, for example, Dhillon [13] and references therein). Dongarra et al. [16] and Bischof et al. [6] have significantly contributed to improving the efficiency of the reduction step.

2.1 LAPACK

Lapack (Anderson et al. [2]) makes use of high level Blas (Dongarra et al. [14, 15]) to exploit features of modern computers. In particular, Level 3 Blas exploit memory hierarchies to the highest degree. This is why most Lapack routines are based on Level 3 Blas. Unfortunately, one-sweep tridiagonalization can utilize Level 3 Blas for only 50 % of all floating-point operations due to intricate data dependencies (see Dongarra et al. [16], Gansterer et al. [17]).

2.2 SBR Toolbox

Bischof et al. [5, 6, 8] developed algorithms for the band reduction of a matrix and implemented them in the SBR (successive band reduction) toolbox (Bischof et al. [7]). They showed that it is possible to tridiagonalize a matrix in several sweeps in a blocked fashion, relying on Level 3 BLAS for the majority of floating-point operations.

Using the SBR toolbox they developed multi-sweep eigensolvers. If eigenvectors are to be computed from the tridiagonal matrix T, these multi-sweep methods are not competitive.

2.3 Computation of Eigenvectors from an Intermediate Band Matrix

Back-transformation of eigenvectors is much faster if the eigenvectors are computed from the intermediate band matrix B (see Gansterer et al. [17, 18, 19] and Fig. 1). The band width b of B has to be small to enable a fast inverse iteration. This, unfortunately, limits the block size of the first reduction sweep.

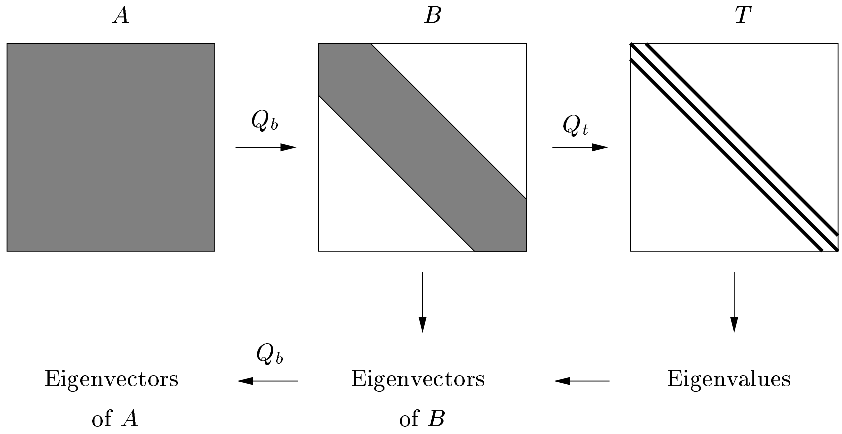

FIG. 1. *Structure of a two-sweep symmetric eigensolver.*

2.4 Parallel Eigensolvers

Parallel eigensolvers can be found, for instance, in SCALAPACK (Blackford et al. [9]), or in PLAPACK (van de Geijn [28]). The HJS algorithm is described in Hendrickson et al. [23]. One-sided Jacobi methods can be found, for example, in Hegland [22]. A parallel implementation of the Arnoldi method is found in Maschhoff, Sorensen [25]. Invariant subspace decomposition has been investigated in the PRISM project (Auslander, Tsao [3], Bischof et al. [4]). The SBR toolbox (see Section 2.2) was developed as part of PRISM (Parallel Research on Invariant Subspace Methods). A prototype parallel implementation of the SBR toolbox has been developed (Wu et al. [30]).

3 Independently Blocked Eigensolver

Two-sweep tridiagonal reduction, utilizing the SBR toolbox (Bischof et al. [7]), is an efficient scheme if eigenvectors are to be computed from an intermediate band matrix B (see the performance curve denoted "two-sweep blocked" in Fig. 2 and Gansterer et al. [17, 18, 19]).

3.1 Improved Two-Sweep Reduction

To further improve the performance of two-sweep eigensolvers *independently blocked* algorithms have been developed (Wu et al. [30]). To explain the improved two-sweep technique, one-sweep reduction is discussed first.

One-Sweep Reduction. To reduce the symmetric matrix A to a symmetric tridiagonal matrix $T = Q^\top AQ$ orthogonal similarity transformations are used. Normally, Q is not generated explicitly. Instead of calculating $Q = H_2 H_3 \dots H_{n-2}$ the Householder matrices $H_i = I - \tau u_i u_i^\top$ are applied to A iteratively for $i = 2, 3, \dots, n-1$:

$$A_i = H_i^\top A_{i-1} H_i, \qquad \text{with} \quad A_1 \equiv A, \;\; A_{n-1} \equiv T.$$

The application of each Householder matrix is done in three steps:

Step 1: $y_i = A_{i-1} u_i$,
Step 2: $v_i = y_i - \frac{1}{2} u_i (y_i^\top u_i)$,
Step 3: $A_i = A_{i-1} - u_i v_i^\top - v_i u_i^\top$.

In the routine `LAPACK/dsytrd` (see Section 2.1 and Dongarra et al. [16]), Step 3 is recast as a block operation by replacing k consecutive rank-2 updates by one rank-$2k$ update. This restructuring enables the utilization of Level 3 Blas which reuse data as much as possible and therefore maximize the ratio of floating-point operations to memory references. The resulting one-sweep algorithm still has an efficiency bottleneck. Step 1 involves a matrix-vector product which cannot be blocked. This Level 2 operation requires approximately 50 % of the floating-point operations of the tridiagonalization process, but it is responsible for up to 80 % of the execution time in a sequential solver (see Haunschmid, Kvasnicka [21]). This means, if the 80 % share were reduced to 50 %, a speedup by a factor of 2.5 could be achieved.

Two-Sweep Reduction. In *two-sweep* algorithms (see Section 2.2 and Bischof et al. [5, 6, 7]) all vectors occurring in the tridiagonalization phase can be replaced by matrix panels. Fig. 1 illustrates the sequence of operations in a two-sweep symmetric eigensolver. Two-sweep eigensolvers can be implemented using Level 3 Blas in Step 3 *and* Step 1. However, the block size has to be chosen carefully. Large block sizes help to reduce the communication overhead on parallel machines. Moreover, they are often appropriate to match the memory hierarchy. On the other hand, the block size chosen for the first sweep, i. e., in the reduction of A to the band matrix B, is a lower bound for the band width of B. Large block sizes and thus large band widths require much extra work to be done in the second sweep and in the computation of the eigenvectors.

Due to these conflicting factors an optimum block size exists for each computer and for each problem size. The execution time can be even further reduced by introducing and optimizing two different block sizes instead of one single block size: b_1 in the matrix-matrix product corresponding to Step 1 and b_3 in the rank-$2k$ update corresponding to Step 3. Thus an *independently blocked* algorithm results. This algorithm subsumes one-sweep and two-sweep block algorithms: (i) one-sweep blocking (see Section 2.1) corresponds to setting $b_1 = 1$, and (ii) two-sweep blocking as mentioned in Section 2.2 corresponds to setting $b_1 = b_3$.

In principle, both the eigenvalues and the eigenvectors can be computed from the band matrix B. Execution time can be improved, however, if the eigenvalues are computed from the tridiagonal matrix. The eigenvectors are computed from the band matrix, because this helps to significantly speed up the back-transformation. The accuracy of the eigenvectors is not affected, provided good approximations of the respective eigenvalues are available.

The floating-point performance of various sequential eigensolvers—the independently blocked algorithm, the one-sweep algorithm, the two-sweep algorithm, and the unblocked algorithm—is illustrated in Fig. 2.

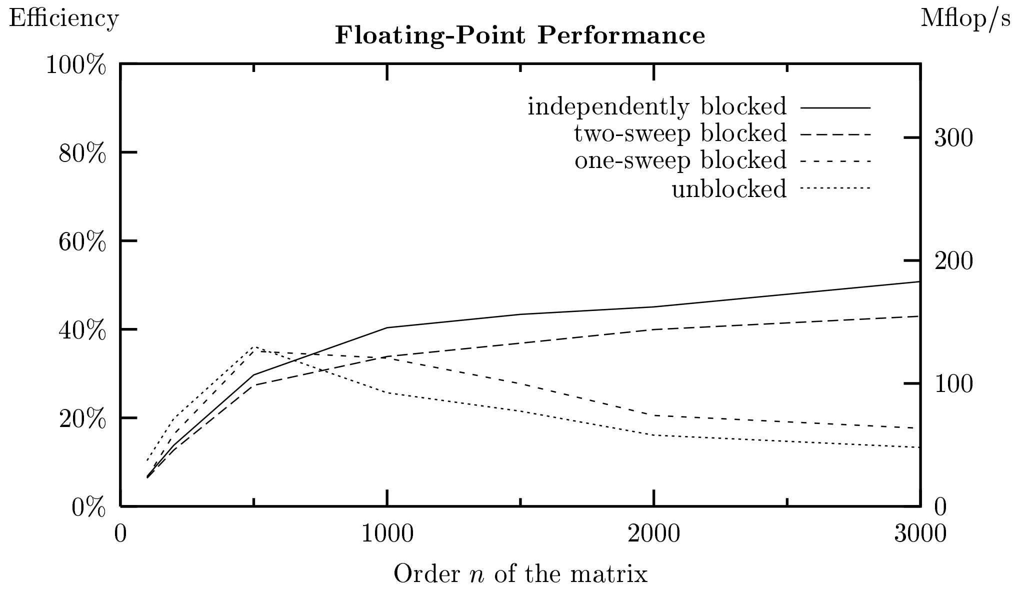

FIG. 2. *Performance of symmetric eigensolvers on one processor of an SGI Power Challenge. 10 % of the eigenvalues and eigenvectors are computed. Efficiency is given in terms of the theoretical peak performance of 360 Mflop/s.*

The independently blocked band reduction algorithm (see Algorithm 1) makes it possible to choose b_3 large enough to achieve satisfactory floating-point performance in the rank-$2k$ update. Independently, b_1 can be chosen (i) to meet the requirements of the communication network and (ii) to keep the computational overhead of the second sweep and of the eigenvector computation small. As a result, the number of accesses to matrix A is reduced: $\lceil n/b_1 \rceil$ symmetric matrix multiplications are necessary to generate the update matrices in the inner loop and $\lceil n/b_3 \rceil$ symmetric rank-$2k$ updates are required to apply the update matrices in the outer loop.

3.2 Parallelization

Not all code sections of symmetric eigensolvers have to be parallelized. If only a few (or no) eigenvectors are to be computed, a parallelization of the first reduction sweep leads to a sufficient performance improvement, because the first reduction sweep is the only section of the algorithm with an algorithmic complexity of $O(n^3)$.

Figure 3 and Tables 1 and 2 illustrate the performance of various parallel eigensolvers, when 10 % of the eigenvalues and eigenvectors are computed. The one-sweep blocked eigensolver of LAPACK, the two-sweep blocked eigensolver, based on the SBR band reduction, and the independently blocked eigensolver, which combines these two blocking methods, have been parallelized by calling the parallel BLAS routines of the optimized SGI/Cray Scientific Library (SCSL). These codes are compared with the SCALAPACK implementation of the one-sweep blocked eigensolver, i. e., a parallelization of the respective LAPACK eigensolver. This SCALAPACK code shows a satisfactory performance behavior for large matrices, where the explicit exchange of messages is an advantage. For small matrices

ALGORITHM 1

n ... order of the matrix A
p ... block size b_3 of the outer loop (for example, $p = 60$)
$\bar{p}$ ... block size b_1 of the inner loop (for example, $\bar{p} = 8$)

$$
\begin{array}{l}
N = (n-2)/p \\
\bar{N} = p/\bar{p} \\
A_0 = A \\
\textbf{do } k = 1, 2, \ldots, N \\
\quad s = (k-1)p + 1 \\
\quad \textbf{do } j = 1, 2, \ldots, \bar{N} \\
\quad\quad \bar{s} = (k-1)p + (j-1)\bar{p} + 1 \\
\quad\quad \textbf{do for } m \in \{\bar{s}, \bar{s}+1, \ldots, \bar{s}+\bar{p}-1\} \\
\quad\quad\quad a_m = a_m - \sum_{l=\bar{s}}^{m-1} (v_l^{(m)} u_l + u_l^{(m)} v_l) \\
\quad\quad \textbf{end do} \\
\quad\quad \text{find } (u_{\bar{s}}, u_{\bar{s}+1}, \ldots, u_{\bar{s}+\bar{p}+1}) \text{ by QR factorizing } (a_{\bar{s}}, a_{\bar{s}+1}, \ldots, a_{\bar{s}+\bar{p}+1}) \\
\quad\quad \textbf{do for } m \in \{\bar{s}, \bar{s}+1, \ldots, \bar{s}+\bar{p}-1\} \\
\quad\quad\quad x_m = A_{k-1} u_m \\
\quad\quad \textbf{end do} \\
\quad\quad \textbf{do } m = \bar{s}, \bar{s}+1, \ldots, \bar{s}+\bar{p}-1 \\
\quad\quad\quad y_m = x_m - \sum_{l=\bar{s}}^{m-1} \left[u_l(v_l^T u_m) + v_l(u_l^T u_m)\right] \\
\quad\quad\quad v_m = y_m - \frac{1}{2} u_m (y_m^T u_m) \\
\quad\quad \textbf{end do} \\
\quad \textbf{end do} \\
\quad A_k = A_{k-1} - \sum_{l=s}^{s+p-1} (u_l v_l^T + v_l u_l^T) \\
\textbf{end do} \\
B = A_N
\end{array}
$$

the lower overhead of a parallelization using parallel (SCSL) BLAS leads to better overall performance.

In the comparison of the one-sweep, the two-sweep, and the independently blocked eigensolvers, all parallelized by calling parallel (SCSL) BLAS, the independently blocked eigensolver shows the best performance.

3.3 Application in Material Sciences

The material sciences package WIEN 97 (see Blaha et al. [10], Schwarz, Blaha [27]), used by more than 200 research groups worldwide, allows to compute the electronic properties of solids by using quantum mechanics. One of its computationally most intensive parts is the calculation of about 10 % of the smallest eigenvalues and the corresponding eigenvectors of the real (or complex) Hamiltonian.

Preliminary tests carried out with a fine-grained parallel version of WIEN 97 showed that the symmetric eigensolver was the performance critical part. Due to intricate data dependencies in the tridiagonalization step (see Haunschmid, Kvasnicka [21]) it did not scale well with the number of processors.

Using the independently blocked eigensolvers discussed in this paper, significant performance improvements of the parallel version of WIEN 97 can be achieved.

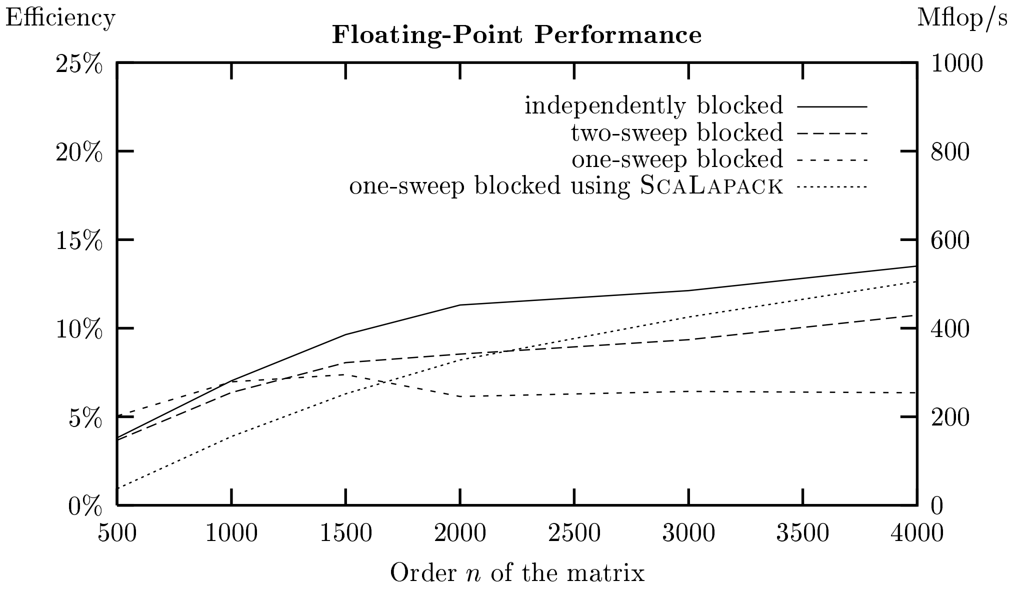

FIG. 3. *Performance of symmetric eigensolvers on 8 processors of an SGI Origin 2000. 10 % of the eigenvalues and eigenvectors are computed. Efficiency is given in terms of the theoretical peak performance of 4 Gflop/s.*

TABLE 1
Run time (in seconds) of symmetric eigensolvers on one processor of an SGI Origin 2000. 10 % of the eigenvalues and eigenvectors are computed.

$\mathbf{P = 1}$	n	independently blocked	two-sweep blocked	one-sweep blocked	ScaLapack
	1000	7.1 s	7.8 s	7.0 s	10.5 s
	1500	21.3 s	27.2 s	23.9 s	35.3 s
	2000	60.4 s	65.8 s	67.8 s	83.6 s
	3000	210.5 s	245.6 s	249.6 s	296.3 s
	4000	721.9 s	896.3 s	586.4 s	734.0 s

TABLE 2
Run time (in seconds) of symmetric eigensolvers on eight processors of an SGI Origin 2000. 10 % of the eigenvalues and eigenvectors are computed.

$\mathbf{P = 8}$	n	independently blocked	two-sweep blocked	one-sweep blocked	ScaLapack
	1000	5.7 s	6.3 s	5.7 s	10.3 s
	1500	14.0 s	16.8 s	18.3 s	21.5 s
	2000	28.3 s	37.5 s	52.1 s	39.0 s
	3000	89.0 s	115.4 s	167.8 s	101.6 s
	4000	189.4 s	238.6 s	402.4 s	202.6 s

4 Conclusion

In this paper methods for solving symmetric eigenproblems are discussed, which are efficient in cases where not all eigenvalues and eigenvectors are required. Independently blocked tridiagonalization enables a better exploitation of modern hierarchical memories, and improves scalability on parallel computers.

References

[1] R. J. Allan and I. J. Bush, *Parallel application software on high performance computers. Parallel diagonalisation routines.*, tech. rep., The CCLRC HPCI Centre at Daresbury Laboratory, 1996.

[2] E. Anderson et al., LAPACK *Users' Guide, 2nd ed.*, SIAM Press, Philadelphia, 1995.

[3] L. Auslander and A. Tsao, *On parallelizable eigensolvers*, Adv. Appl. Math., 13 (1992), pp. 253–261.

[4] C. H. Bischof, S. Huss-Ledermann, X. Sun, A. Tsao, and T. Turnbull, *Parallel studies of the invariant subspace decomposition approach for banded symmetric matrices*, in Proceedings of Seventh SIAM Conference of Parallel Processing for Scientific Computing, Philadelphia, 1995, SIAM Press, pp. 516–521.

[5] C. H. Bischof, B. Lang, and X. Sun, *Parallel tridiagonalization through two-step band reduction*, in Proceedings of the Scalable High-Performance Computing Conference, Washington, D. C., 1994, IEEE Press, pp. 23–27.

[6] ——, *A framework for symmetric band reduction*, Tech. Rep. ANL/MCS-P586-0496, Argonne National Laboratory, 1996.

[7] ——, *The* SBR *toolbox—software for successive band reduction*, Tech. Rep. ANL/MCS-P587-0496, Argonne National Laboratory, 1996.

[8] C. H. Bischof, M. Marques, and X. Sun, *Parallel bandreduction and tridiagonalization*, in Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing, R. Sincovec, ed., Philadelphia, 1993, SIAM Press, pp. 383–390.

[9] L. S. Blackford et al., SCALAPACK *Users' Guide*, SIAM Press, Philadelphia, 1997.

[10] P. Blaha, K. Schwarz, and J. Luitz, WIEN 97*. Improved and updated* UNIX *version of the original copyrighted* WIEN *code [11]*, tech. rep., Vienna University of Technology, 1997.

[11] P. Blaha, K. Schwarz, P. Sorantin, and S. B. Trickey, *Full-potential, linearized augmented plane wave programs for crystalline systems*, Comp. Phys. Commun., 59 (1990), pp. 399–415.

[12] J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations, Volume 1: Theory, Volume 2: Programs*, Birkhäuser, Basel, 1985.

[13] I. S. Dhillon, *A New $O(n^2)$ Algorithm for the Symmetric Tridiagonal Eigenvalue / Eigenvector Problem*, PhD thesis, University of California, Berkeley, 1997.

[14] J. J. Dongarra, J. Du Croz, S. Hammarling, and I. Duff, *A set of Level 3* BLAS, ACM Trans. Math. Software, 16 (1990), pp. 1–17.

[15] J. J. Dongarra, J. Du Croz, S. Hammarling, and R. J. Hanson, *An extended set of* BLAS, ACM Trans. Math. Software, 14 (1988), pp. 18–32.

[16] J. J. Dongarra, S. J. Hammarling, and D. C. Sorensen, *Block reduction of matrices to condensed forms for eigenvalue computations*, J. Comput. Appl. Math., 27 (1989), pp. 215–227.

[17] W. N. Gansterer, D. F. Kvasnicka, and C. W. Ueberhuber, *High performance computing in material sciences. Higher Level* BLAS *in symmetric eigensolvers*, Tech. Rep. AURORA TR1998-18, Vienna University of Technology, 1998.

[18] ——, *High performance computing in material sciences. Numerical experiments with symmetric eigensolvers*, Tech. Rep. AURORA TR1998-19, Vienna University of Technology, 1998.

[19] ——, *A Level 3 algorithm for the symmetric eigenproblem*, in Proceedings of the Third International Meeting on Vector and Parallel Processing (VECPAR '98), to appear, Berlin Heidelberg New York Tokio, 1998, Springer-Verlag.

[20] G. Golub and C. Van Loan, *Matrix Computations, 3rd ed.*, The Johns Hopkins University Press, Baltimore, Maryland, 1996.

[21] E. Haunschmid and D. Kvasnicka, *High performance computing in material sciences. Maximizing cache utilization without increasing memory requirements*, Tech. Rep. AURORA TR1998-17, Vienna University of Technology, 1998.

[22] M. Hegland, *A distribution independent algorithm for the reduction to tridiagonal form using one-sided rotations*, in First IEEE International Conference on Algorithms and Architectures for Parallel Processing, IEEE Press, 1995, pp. 286–269.

[23] B. Hendrickson, E. Jessup, and C. Smith, *Towards an efficient parallel eigensolver for dense symmetric matrices*, tech. rep., available from http://www.cs.sandia.gov/~bahendr/-papers.html.

[24] R. B. Lehoucq, D. C. Sorensen, and C. Yang, ARPACK *Users' Guide—Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM Press, Philadelphia, 1998.

[25] K. J. Maschhoff and D. C. Sorensen, PARPACK*: an efficient portable large scale eigenvalue package for distributed memory parallel architectures*, in Applied Parallel Computing. Industrial Computation and Optimization. Third International Workshop, PARA '96 Proceedings, J. Wasniewski, J. J. Dongarra, K. Madsen, and D. Olesen, eds., Berlin Heidelberg New York Tokyo, 1996, Springer-Verlag, pp. 478–486.

[26] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM Press, Philadelphia, 1998.

[27] K. Schwarz and P. Blaha, *Description of an* LAPW DF *program (*WIEN 95*)*, in Quantum-Mechanical Ab-Initio Calculation of Properties of Crystalline Materials, C. Pisani, ed., vol. 67 of Lecture Notes in Chemistry, Berlin Heidelberg New York Tokyo, 1996, Springer-Verlag, pp. 139–154.

[28] R. A. van de Geijn, *Using* PLAPACK*: Parallel Linear Algebra Package*, MIT Press, 1997.

[29] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

[30] Y.-J. J. Wu, P. A. Alpatov, C. H. Bischof, and R. A. van de Geijn, *A parallel implementation of symmetric band reduction using* PLAPACK, in Proceedings of the Scalable Parallel Libraries Conference, 1996.