Sparse approximate multifrontal factorization with composite compression methods

LISA CLAUS, National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, USA

PIETER GHYSELS, Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, USA

YANG LIU, Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, USA

THÁI ANH NHAN, Department of Mathematics and Computer Science, Santa Clara University, USA

RAMAKRISHNAN THIRUMALAISAMY, Department of Mechanical Engineering, San Diego State University, USA

AMNEET PAL SINGH BHALLA, Department of Mechanical Engineering, San Diego State University, USA

SHERRY LI, Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, USA

This paper presents a fast and approximate multifrontal solver for large sparse linear systems. In a recent paper by Liu et al. we showed 21 22 the efficiency of a multifrontal solver leveraging the butterfly algorithm and its hierarchical matrix extension, Hierarchically Off-23 Diagonal Butterfly compression (HODBF), to compress large frontal matrices. The resulting multifrontal solver can attain quasi-linear 24 computation and memory complexity when applied to sparse linear systems arising from spatial discretization of high-frequency 25 wave equations. To further reduce the overall number of operations and especially the factorization memory usage in order to scale to 26 larger problem sizes, in this paper, we develop a composite multifrontal solver that employs the HODBF format for large sized fronts, 27 a reduced-memory version of the non-hierarchical Block Low-Rank format for medium sized fronts and a lossy compression format 28 for small sized fronts. This allows us to solve sparse linear systems of dimension up to 2.7× larger than before and leads to a memory 29 30 consumption that is reduced by 70 percent while ensuring the same execution time. The code is made publicly available in github.

Additional Key Words and Phrases: Sparse direct solver, multifrontal method, butterfly algorithm, block low-rank compression

ACM Reference Format:

Authors' addresses: Lisa Claus, lclaus@lbl.gov, National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley,
 CA, USA; Pieter Ghysels, pghysels@lbl.gov, Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory,
 Berkeley, CA, USA; Yang Liu, liuyang@lbl.gov, Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory,
 Berkeley, CA, USA; Thái Anh Nhan, anhan@scu.edu, Department of Mathematics and Computer Science, Santa Clara University, Santa Clara, CA, USA;
 Ramakrishnan Thirumalaisamy, rthirumalaisam1857@sdsu.edu, Department of Mechanical Engineering, San Diego State University, San Diego, CA,
 USA; Amneet Pal Singh Bhalla, asbhalla@sdsu.edu, Department of Mechanical Engineering, San Diego State University, San Diego, CA, USA; Sherry Li,
 xsli@lbl.gov, Applied Mathematics and Computational Research Division, Lawrence Berkeley, CA, USA.

- ⁴⁹ © 2018 Association for Computing Machinery.
- 50 Manuscript submitted to ACM
- 51

1 2

3

8

9 10

11

12 13

14

15 16

17

18 19

20

31

32 33

34

35

36

 <sup>45
 46
 46
 47
 47
 48
 48
 49
 49
 49
 41
 41
 42
 43
 44
 44
 45
 46
 47
 47
 48
 48
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 40
 41
 42
 43
 44
 44
 45
 46
 47
 48
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 40
 41
 41
 42
 43
 44
 44
 44
 44
 45
 46
 47
 48
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 49
 4</sup>

1 INTRODUCTION

Efficiently computing the solution of large sparse linear systems arising from finite element, finite difference, or finite volume discretizations of partial differential equations (PDEs) is an important requirement for many scientific and engineering applications. The multifrontal method is a fast solution method that can be implemented very efficiently on modern hardware, since it arranges the computations in such a way that most of the computational work is done on smaller dense submatrices, so-called frontal matrices. Unfortunately, the overall amount of dense linear algebra operations needed to complete the multifrontal method sums up to $O(N^2)$ for typical three dimensional PDEs, where N is the matrix dimension corresponding to the sparse linear system.

63 For many applications arising from wide classes of PDEs, this complexity can be reduced up to $O(N \log^{\alpha} N)$ for some 64 α by leveraging algebraic compression formats to exploit rank structures in off-diagonal blocks of the matrix. Examples 65 of low-rank-based compression methods include \mathcal{H} matrices [12, 15], hierarchically off-diagonal low-rank (HODLR) 66 67 formats [1], hierarchically semiseparable (HSS) formats [43], and block low-rank (BLR) formats [2, 4, 42]. Available 68 software packages that couple these rank-structured formats with multifrontal methods include STRUMPACK [10] 69 and MUMPS [4]. PaStiX [17, 35] is an additional software package that couples low-rank compression methods with 70 supernodal methods instead of multifrontal methods. 71

72 In addition to the above mentioned algorithms, we consider another low-rank-based compression tool called 73 butterfly [21, 22, 29, 32, 38], a multilevel matrix decomposition algorithm well-suited for representing highly oscillatory 74 operators such as Fourier transforms and integral operators and special function transforms. When combined with 75 hierarchical matrix techniques, butterfly can also serve as the building block for accelerating iterative methods, direct 76 77 solvers and preconditioners for boundary element methods for high-frequency wave equations. These techniques 78 essentially replace low-rank products in the $\mathcal H$ and HODLR formats with butterflies and leverage fast and randomized 79 butterfly algebra to compute the matrix inverse (for direct solvers and preconditioners). An open-source software 80 package that provides an implementation of the butterfly algorithm is available at https://github.com/liuyangzhuan/ 81 82 ButterflyPACK.

83 In a related work [27], we presented a fast multifrontal sparse solver for high-frequency wave equations. The 84 solver leverages the butterfly algorithm and its hierarchical matrix extension, Hierarchically Off-Diagonal Butterfly 85 compression (HODBF), to compress large frontal matrices. The resulting solver can attain quasi-linear computation 86 87 and memory complexity when applied to high-frequency Helmholtz and Maxwell equations. Similar complexities 88 have been analyzed and observed for Poisson equations as well. Nevertheless, to further reduce the overall number 89 of operations and to enable solving larger problem sizes, in this paper, we presents a composite multifrontal solver 90 which employs the HODBF format for compressing the large frontal matrices in the multifrontal method and leverage 91 92 additional compression methods for the remaining fronts. To be more specific, HODBF serves as a good compression 93 method as shown in [27]. However, due to its more complex data structures it only pays off to use HODBF for larger 94 fronts to yield a high compression ratio. For medium sized fronts we employ block low-rank (BLR) compression. 95 We combine a left-looking and a right-looking versions of BLR to a hybrid method which decreases the memory 96 97 consumption significantly. For the remaining small sized fronts, we make use of lossy compression enabled through the 98 zfp software [24] to further decrease the memory consumption while maintaining accuracy of the preconditioner. 99

Our contributions in this paper are the development of a composite multifrontal solver that employs three compression methods: HODBF, BLR and floating point compression, and an implementation of BLR with a reduced memory footprint which makes use of a column-wise construction of matrix tiles. The sparse approximate multifrontal solver is used as a

102 103

100

101

104 Manuscript submitted to ACM

53

preconditioner for restarted GMRES(30) with modified Gram-Schmidt and a zero-vector initial guess. These updates
 lead to a significant decrease of memory consumption and it allows us to solve sparse linear systems of dimension
 up to 2.7× larger compared to a multifrontal solver with HODBF compression only [27]. The code is made publicly
 available in the sparse solver package STRUMPACK [9].

The rest of the paper is organized as follows. The multifrontal factorization method is presented in Section 2. The zfp, hierarchically off-diagonal butterfly compression and block low-rank algorithms are described in Section 3, including the classical left- and right-looking BLR versions as well as our proposed hybrid and memory optimized BLR algorithm. The proposed composite rank-structured multifrontal method is detailed in Section 4. Numerical results demonstrating the efficiency and applicability of the proposed solver for the 3D Helmholtz, reaction-diffusion, and Navier-Stokes equations are presented in Section 5, followed by conclusions in Section 6.

2 MULTIFRONTAL FACTORIZATION

118

119 120

121

122 123

124

125

126

127 128

129

130

131

132

133 134

135

136

137 138

139 140

141

142

143

144 145

146

147

148 149

150 151

152

This section briefly recalls the main ingredients of the multifrontal method for general invertible sparse matrices. For a more detailed discussion of multifrontal methods, see [20, 25]. The method separates the factorization of a sparse matrix (A = LU) into a series of partial factorizations of many smaller dense matrices, which correspond to the separators from a nested dissection ordering. After each factorization step a Schur complement is formed and carried along temporarily, and its scattering to the global Schur complement is delayed until that part of panel factorization is about to start.

As a preprocessing step, the system matrix A is first scaled and permuted for numerical stability: $A \leftarrow D_r A D_c Q_c$, where D_r and D_c are diagonal matrices that scale the rows and columns of A and Q_c is a column permutation that places large entries on the diagonal. We use the MC64 code by Duff and Koster [19] or the parallel method-without the diagonal scaling-described in [6] to perform the scaling and column permutation. After that, a fill-reducing permutation $A \leftarrow PAP^T$ is applied, i.e. the number of nonzero entries in the sparse factors L and U is minimized. The permutation matrix P is typically computed using nested dissection applied to the adjacency graph of $A + A^T$, as implemented in Scotch [39] or METIS [11].

The multifrontal method relies on a structure called the assembly tree. Each node τ of the assembly tree is represented by a dense frontal matrix F_{τ} , with the following 2×2 block structure: $F_{\tau} = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}$. The rows and columns corresponding to the F_{11} block are called the fully summed variables, these variables have received all their Schur complement updates, when the front is constructed. We denote the dimension of F_{11} by $\#I_{\tau}^s$ and the dimension of F_{22} by $\#I_{\tau}^u$. The I_{τ}^u index sets define the temporary Schur complement update blocks. Let $n_{\tau} = \#I_{\tau}^s + \#I_{\tau}^u$ denote the dimension of F_{τ} . Note that the frontal matrices tend to get bigger toward the root of the assembly tree. Furthermore, if ν is a child of τ in the assembly tree, then $I_{\nu}^u \subset \{I_{\tau}^s \cup I_{\tau}^u\}$. For the root node $t, I_t^u = \emptyset$. When considering a single front, we will omit the τ subscript.

The multifrontal method consists of a bottom-up traversal of the assembly tree following a topological ordering. Processing a node consists of four steps making up the numerical factorization of that node:

(1) Assembling the frontal matrix F_{τ} , i.e., combining elements from the sparse matrix A with the children's (v_1 and v_2) contribution blocks. This involves a scatter operation and is called extend-add, denoted by +:



- 157 (2) Elimination of the fully summed variables in the F_{11} block, i.e., dense LU factorization with partial pivoting of 158 F_{11} .
 - (3) Updating the off-diagonal blocks F_{12} and F_{21} .
 - (4) Computing the contribution block from the Schur complement update of $F_{22} \leftarrow F_{22} F_{21}F_{11}^{-1}F_{12}$. F_{22} is temporary storage and can be released as soon as it has been used in the front assembly (step 1) of the parent node.

After the numerical factorization, the lower triangular sparse factor is available in the F_{21} and F_{11} blocks and the upper triangular factor in the F_{11} and F_{12} blocks. These can then be used to efficiently solve linear systems, using forward and backward substitution. A high-level overview is given in Algorithm 1.

169 Algorithm 1 Sparse multifrontal factorization and solve. 170 **Input:** $A \in \mathbb{R}^{N \times N}$, $b \in \mathbb{R}^N$ 171 **Output:** $x \approx A^{-1}b$ 172 1: $A \leftarrow D_r A D_c Q_c$ ▷ (optional) col perm & scaling 173 $2: A \leftarrow PAP^{\top}$ 174 ▶ symm fill-reducing reordering 3: Build assembly tree: define I_{τ}^{s} and I_{τ}^{u} for every frontal matrix F_{τ} 175 4: for nodes τ in assembly tree in topological order do 176 ▶ sparse with the children updates extended and added 177 5:
$$\begin{split} F_{\tau} &\leftarrow \begin{bmatrix} A(I_{\tau}^{s}, I_{\tau}^{s}) & A(I_{\tau}^{s}, I_{\tau}^{u}) \\ A(I_{\tau}^{u}, I_{\tau}^{z}) & 0 \end{bmatrix} \div F_{22;\nu_{1}} \div F_{22;\nu_{2}} \\ P_{\tau}L_{\tau}U_{\tau} &\leftarrow F_{11} \\ F_{12} &\leftarrow L_{\tau}^{-1}P_{\tau}^{\top}F_{12} \\ F_{21} &\leftarrow F_{21}U_{\tau}^{-1} \\ F_{22} &\leftarrow F_{22} - F_{21}F_{12} \\ d \text{ for } \end{split}$$
178 6 179 ▶ LU with partial pivoting 7: 180 8: 181 9: 182 ▶ Schur update 10: 183 11: end for 184 12: $x \leftarrow D_c Q_c P^\top$ bwd-solve (fwd-solve (PD_rb)) 185 186 187 S⁰ 188 S^0 189 190 191 S_{1}^{1} I_{τ}^{s} 192 193 I^{u} 194 195 196 197 198 199 (a) (b) 200 201

- Fig. 1. (a) The top three levels of nested dissection for an 11^2 mesh. The root separator S^0 is a vertical 11 point line. The next level separators are S_0^1 and S_1^1 . The root separator corresponds to the top level front in (b), and similarly for the next level down in the assembly/frontal tree. Note that the fronts in (b) typically get smaller lower in the tree.
- 204 205 206

207

208

Figure 1 illustrates the multifrontal algorithm for a sparse matrix resulting from the discretization of a partial differential equation using a 5-point finite difference stencil on a regular two-dimensional 11 × 11 mesh. Figure 1a Manuscript submitted to ACM

4

160

161

162 163

shows the mesh and the top 3 levels of the nested dissection ordering. Nested dissection is a heuristic algorithm for the 209 210 ordering of a sparse matrix to reduce the fill-in in the sparse factors. It is based on recursively finding vertex separators. 211 The vertical line marked S^0 is the root separator and this separator corresponds to the root of the assembly tree, see 212 Fig. 1b. The next level separators, S_0^1 and S_1^1 correspond to the F_{11} blocks of the next lower level in the assembly tree. 213 214 Typically the larger frontal matrices are found near the root of the assembly tree since the separators tend to get smaller 215 further in the nested dissection recursion. 216

2.1 Parallel Traversal of the Assembly Tree

217

218

235 236

237

238 239

240

241

242

243

244 245

246

247 248

249

251

252

259

260

219 Following the assembly tree, the distributed algorithm creates nested MPI subcommunicators to facilitate the com-220 putation at each node and its subtree. At the root of the tree, we create a two-dimensional process grid using the 221 available processes in the root MPI communicator, and distribute the frontal matrix over this grid using the ScaLAPACK 222 2D block-cyclic data layout. Next, the root MPI communicator is split in two communicators proportionally to the 223 224 memory required by the subtrees rooted at the children of the root node. Each child constructs a 2D process grid and 225 distributes the child's frontal matrix over this subgrid. This is repeated recursively until the MPI communicator has 226 only one process in it, at which point the local subtree is traversed within a single OpenMP region using OpenMP task 227 parallelism. While moving up the distributed part of the assembly tree, communication between fronts is required for 228 the extend-add operation. This is implemented using an MPI Alltoally on the MPI subcommunicator of the parent node. 229 230 Note that this paragraph discusses the parallel traversal of the assembly tree with a focus on dense frontal matrices. 231 In Section 4 we will discuss details of the procedure for compressed frontal matrices, where a 2D block-cyclic layout 232 is used only for dense, Block Low-Rank and lossy compressed fronts while the Hierarchically Off-Diagonal Butterfly 233 234 compression is based on a 1D block layout, see Section 3.3.1.

We implemented the multifrontal method in the STRUMPACK library [9], using C++, message passing interface (MPI), and OpenMP. STRUMPACK supports real/complex arithmetic, single/double precision and 32/64-bit integers. STRUMPACK has recently ported the sparse direct multifrontal solver to GPU, targeting both NVIDIA and AMD hardware. In this paper, we focus on the CPU implementation only.

In what follows, we leverage multiple compression methods, namely lossy compression enabled through the zfp software [24], see Section 3.1, Block Low-Rank (BLR) compression (Section 3.2), and the hierarchical matrix extension of the butterfly algorithm (HODBF), see Section 3.3. These methods are used to represent frontal matrices and to construct fast sparse direct solvers, particularly for large matrix systems resulting from for example high-frequency wave equations.

3 COMPRESSION METHODS

250 In this paper, we make use of compression methods within the multifrontal solver which maintains the solver's robustness and reliability and reduces the computational complexity. The three compression methods of interest are the HODBF format, the BLR format and lossy compression. The HODBF format can be described as a hierarchical 253 254 compression format with a multilevel matrix decomposition algorithm. HODBF is available as an effort to integrate the 255 dense solver package ButterflyPACK [26] into the sparse solver package STRUMPACK and can be used to compress 256 frontal matrices within the multifrontal solver. BLR is based on a flat low-rank based compression format which exploits 257 rank structures in off-diagonal blocks of the frontal matrix. BLR is implemented in STRUMPACK and can also be used 258 to compress fronts. The lossy floating point compression method is available through the zfp package [24] and is Manuscript submitted to ACM

integrated into STRUMPACK. Since the multifrontal method relies on dense factorizations, all three approximations can
 be easily incorporated into the multifrontal factorization by representing the frontal matrices as zfp, BLR, and HODBF
 matrices respectively, as will be described in Section 4. All three compression formats are described in detail in the
 following subsections.

3.1 Lossy Compression with zfp

zfp is an open-source library for compressed floating-point data. zfp was designed to achieve high compression ratios and therefore uses lossy compression. zfp is often more accurate and faster than other lossy compressors. For more details on the zfp software library, see [24].

In contrast to low-rank formats, lossy compression is a near-lossless compression scheme that maps small blocks of 4^d values with dimension d to a fixed number of bits per block, called bitplanes in Section 5. This compressor is based on an orthogonal block transform. For more details on the zfp algorithm please refer to [23].

In Section 4, we make use of the zfp software library integrated into STRUMPACK to compress small frontal matrices within a multifrontal solver.

3.2 Block Low-Rank Compression

Among the possible low-rank formats, Block Low-Rank (BLR) is the simplest. The format partitions the matrix with a flat, non-hierarchical blocking of the matrix which is defined by conveniently clustering the associated unknowns and approximates its off-diagonal blocks by low-rank submatrices. A BLR representation \tilde{B} of a dense matrix B is shown in (1), with $p \times p$ blocks.

$$\tilde{B} = \begin{bmatrix} \tilde{B}_{11} & \tilde{B}_{12} & \dots & \tilde{B}_{1p} \\ \vdots & \ddots & \ddots & \vdots \\ \tilde{B}_{p1} & \dots & \dots & \tilde{B}_{pp} \end{bmatrix}$$
(1)

Assuming $I = \{1, ..., n\}$ is the set of row and column indices of B, we can define the blocking of the matrix as follows: We call a set of indices $\sigma \subseteq I$ a cluster. Then, a clustering of *I* is a disjoint union of clusters which equals *I*. $b = \sigma \times \tau \in I \times I$ is called a block cluster based on clusters σ and τ . A block clustering of $I \times I$ is then defined as a disjoint union of block clusters which equals $I \times I$. A block $B_{\sigma\tau}$ corresponds to an interaction between two subdomains σ and τ , where σ contains the row indices of $B_{\sigma\tau}$ while τ contains its column indices. The rank of a given block $B_{\sigma\tau}$ depends on the interaction it represents. Indeed, if $B_{\sigma\tau}$ is a diagonal block, i.e. $\sigma = \tau$, it represents a self- interaction and is thus full-rank. However, if $B_{\sigma\tau}$ is an off-diagonal block, it may be either full-rank or low-rank depending on the interaction it represents: the weaker the interaction, the lower the rank. The admissibility condition determines whether a block $\sigma \times \tau$ is admissible for low-rank compression. We support both weak admissibility, where every off-diagonal block is compressed, and strong admissibility, where only matrix blocks corresponding to well separated clusters are compressed. The block clustering, and the admissibility condition, are typically formulated in terms of the geometry of the physical system being modeled. However, in this paper we discuss BLR matrices and the corresponding block clustering in the context of an algebraic multifrontal solver, where no geometry information is available. The block clustering and the strong admissibility condition in the context of the multifrontal solver are discussed in Section 4. Based on the choice of the admissibility condition, the off-diagonal blocks $B_{\sigma\tau}(\sigma \neq \tau)$ of size $m_{\sigma} \times m_{\tau}$ and numerical rank $r_{\sigma\tau}$ are approximated by a low-rank matrix $\tilde{B}_{\sigma\tau} = X_{\sigma\tau}Y_{\sigma\tau}^T$ at accuracy ε . $X_{\sigma\tau}$ is a $m_{\sigma} \times r_{\sigma\tau}$ matrix and $Y_{\sigma\tau}$ is a $m_{\tau} \times r_{\sigma\tau}$ matrix.

As part of the numerical factorization within the multifrontal solver, as discussed in Section 2, we make use of a LU factorization of each F_{11} part of a frontal matrix. In order to compute the blocked factorization of the front, four fundamental tasks must be performed:

- Factor (F): LU decomposition " $B_{\sigma\sigma} = L_{\sigma\sigma}U_{\sigma\sigma}$ " for diagonal blocks $\sigma = 1, ..., p$ with partial pivoting
 - Solve (S): Solve triangular linear system " $B_{\sigma\tau} = L_{\sigma\sigma}^{-1} B_{\sigma\tau}$ ", " $B_{\tau\sigma} = B_{\tau\sigma} U_{\sigma\sigma}^{-1}$ ", with $\sigma = 1, ..., p, \tau = \sigma + 1, ..., p$
- Update (U): Matrix-matrix multiplication " $B_{\kappa\tau} = B_{\kappa\tau} - B_{\kappa\sigma}B_{\sigma\tau}$ ", with $\sigma = 1, \dots, p, \tau = \sigma + 1, \dots, p, \kappa = \sigma + 1, \dots, p$ Compression (C): Compress off-diagonal blocks $B_{\sigma\tau} \approx \tilde{B}_{\sigma\tau} = X_{\sigma\tau}Y_{\sigma\tau}$

We refer to the computation of the low-rank approximation $\hat{B}_{\sigma\tau}$ of each block as the compression step, which can be performed in different ways. We chose a QR factorization with column pivoting (i.e., LAPACK's (Anderson et al., 1995) geqp3 routine), which is modified to stop the factorization when the diagonal coefficient of R, $r_{i,i}$, falls below a prescribed threshold ε . We use a relative tolerance (i.e. we stop the factorization after $|r_{i,i}|/|r_{0,0}| < \varepsilon$).

Depending on when the compression step is performed within the numerical factorization, several algorithm variants can be defined and implemented based on the execution order of the four tasks defined above: FSUC, FSCU, FCSU, CFSU. These acronyms indicate the order in which the tasks are performed [31].

Algorithm 2 Right-looking BLR algorithm: FCSU	Algorithm 3 Left-looking BLR algorithm: UFCS						
Input: a $p \times p$ block matrix <i>B</i> of size <i>n</i> ,	Input: a $p \times p$ block matrix <i>B</i> of size <i>n</i> ,						
$B = [B_{i,j}]_{i=1:p,j=1:p}$	$B = [B_{i,j}]_{i=1:p,j=1:p}$						
1: Construct tiles $B_{i,j}$, $\forall i, j$	1: Construct tiles $B_{i,j}$, $\forall i, j$						
2: for $i = 1$ to p do	2: for $i = 1$ to p do						
3: Factor: $B_{i,i} = L_{i,i}U_{i,i}$	3: for $j = i$ to p do						
4: for $j = i + 1$ to p do	4: for $k = 1$ to $i - 1$ do						
5: Compress: $B_{i,j} \approx X_{i,j} Y_{i,j}^T$	5: Update: $B_{i,j} \leftarrow B_{i,j} - X_{i,k} (Y_{i,k}^T X_{k,j}) Y_{k,j}^T$						
6: Solve: $B_{i,j} \leftarrow L_{i,i}^{-1} B_{i,j}$	6: if $j \neq i$ then						
7: Compress: $B_{j,i} \approx X_{j,i} Y_{j,i}^T$	7: $B_{j,i} \leftarrow B_{j,i} - X_{j,k} (Y_{i,k}^T X_{k,i}) Y_{k,i}^T$						
8: Solve: $B_{j,i} \leftarrow B_{j,i} U_{i,i}^{-1}$	8: end if						
9: end for	9: end for						
10: for $j = i + 1$ to p do	10: end for						
11: for $k = i + 1$ to p do	11: Factor: $B_{i,i} = L_{i,i}U_{i,i}$						
12: Update: $B_{k,j} \leftarrow B_{k,j} - X_{k,i}(Y_{k,i}^T X_{i,j}) Y_{i,j}^T$	12: for $j = i + 1$ to p do						
13: end for	13: Compress: $B_{i,j} \approx X_{i,j} Y_{i,j}^T$						
14: end for	14: Solve: $B_{i,j} \leftarrow L_{i,i}^{-1} B_{i,j}$						
15: end for	15: Compress: $B_{j,i} \approx X_{j,i} Y_{j,i}^T$						
	16: Solve: $B_{j,i} \leftarrow B_{j,i} U_{i,i}^{-1}$						
	17: end for						
	18: end for						

These variants are so-called right-looking versions, in the sense that at each step k, as soon as the factor and solve tasks for all blocks in row k and column k have been performed, the entire trailing submatrix (column blocks to its "right") is updated, see Fig. 2a. We make use of the FCSU (standing for Factor, Compress, Solve and Update) variant of the BLR factorization algorithm. All low-rank updates of a given block $\tilde{B}_{\sigma\tau}$ are compressed before the triangular solve of the LU factorization of a dense BLR matrix. Based on the comparative study in [31] the FCSU variant seems the Manuscript submitted to ACM

most promising out of the right-looking versions, since it provides a good balance between efficiency and accuracy. Compressing earlier influences the accuracy negatively, an effect which is quantified in [18].

The right-looking algorithms can be rewritten in a left-looking¹ form, where at each step k, blocks in row k as well as column k are updated using all the blocks already computed (those at its "left"): UFS, UFSC, UFCS, UCFS.

In this paper, we focus on the UFCS version, since this version is most promising in terms of timing due to a different memory access pattern [8] as described in [4, 31]. See Algorithm 2 for the right-looking (RL) and Algorithm 3 for the left-looking (LL) implementation and Fig. 2a and Fig. 2b for a comparison between RL and LL.



Fig. 2. First steps of a BLR compression with FCSU/UFCS version. (2a) Right-looking. (2b) Left-looking. (2c) Hybrid with col_{max} = 3.

The RL and LL variants perform the same number of operations but in a different order, which results in a different memory access pattern. The impact of using a RL or LL factorization is mainly observed on the Update step. In particular, for the RL variant, at each step k, the full rank blocks of the trailing sub-matrix are written and therefore they are loaded many times (at each step of the factorization), while the low rank blocks of the current panel are read once and

408 409

410

411

412 413

414

365 366

367

368

369 370

371

372

⁴¹⁵ ¹This algorithmic variant is also called "left-up-looking" in the literature, for brevity we use "left-looking" throughout the paper.

⁴¹⁶ Manuscript submitted to ACM

Sparse approximate multifrontal factorization with composite compression methods

never loaded again. In the LL variant, at each step k, the full rank blocks of the current panel are written for the first

and last time of the factorization, while the low rank blocks of all the previous panels are read, and therefore they are
 loaded many times during the entire factorization.

421 Algorithm 4 Hybrid BLR algorithm (Column-wise constructed) 422 **Input:** a $p \times p$ block matrix *B* of size *n*, $B = [B_{i,j}]_{i=1:p,j=1:p}$ 423 424 1: **for** b = 0 to $\lceil p/col_{\max} \rceil - 1$ **do** ▷ Advance col_{max} block columns each time 425 **for** $i = col_{\max}b + 1$ to $col_{\max}(b + 1)$ **do** \triangleright Construct *col*_{max} block columns among *P*_c process columns 2: 426 **for** *j* = 1 to *p* **do** 3: 427 Construct tiles $B_{j,i}$ 4: 428 5: end for 429 end for 6: 430 if b > 0 then ▶ See Fig. 3, Subfigures 4-6 7: 431 8: **for** j = 1 to $col_{max}b$ **do** 432 **for** $i = col_{\max}b + 1$ to $col_{\max}(b + 1)$ **do** 9: 433 Compress: $B_{j,i} \approx X_{j,i} Y_{j,i}^T$ 10: 434 Solve: $B_{j,i} \approx B_{j,i} U_{i,i}^{-1}$ 11: 435 **for** k = j + 1 to p **do** 12: 436 Update: $B_{k,i} \approx B_{k,i} - X_{k,j}(Y_{k,i}^T X_{j,i})Y_{j,i}^T$ 13: 437 end for 14: 438 end for 15: 439 end for 16: 440 end if 17: 441 **for** $i = col_{\max}b + 1$ to $col_{\max}(b + 1)$ **do** ▶ See Fig. 3, Subfigures 1-3, 7-9 18: 442 Factor: $P_i B_{i,i} = L_{i,i} U_{i,i}$ LU with partial pivoting for the diagonal block 19: 443 **for** j = i+1 to $col_{max}(b+1)$ **do** 20: 444 Compress: $B_{i,j} \approx X_{i,j} Y_{i,j}^T$ 21: 445 Solve: $B_{i,j} \leftarrow L_{i,i}^{-1} B_{i,j}$ 22: 446 447 23: end for **for** j = i+1 to p **do** 448 24: Compress: $B_{j,i} \approx X_{j,i} Y_{j,i}^T$ 449 25: 450 Solve: $B_{j,i} \approx B_{j,i} U_{i,i}^{-1}$ 26: 451 27: end for 452 **for** j = i+1 to $col_{max}(b+1)$ **do** 28: 453 **for** *k* = *i*+1 to *p* **do** 29 454 Update: $B_{k,j} \approx B_{k,j} - X_{k,i}(Y_{k,i}^T X_{i,j}) Y_{i,i}^T$ 30: 455 end for 31: 456 end for 32: 457 end for 33: 458 34: end for 459 \triangleright We assume that p is a multiple of col_{max} for simplicity, but our code can easily handle any value of p. 35: 460 461 462

As visualized in Fig. 2, all BLR tiles of the entire frontal matrix for the LL as well as the RL versions are constructed at once, as dense tiles, which results in a high peak memory consumption. This is followed by the factorization of the *i*-th diagonal tile, then the tiles of the *i*-th row and *i*-th column are compressed and solved. For the RL version, the update task is executed on all remaining tiles of the trailing sub-matrix. For the LL version, the update task is executed on the immediate neighboring row and column block columns of the trailing sub-matrix.

463

464

465

466 467

468

Based on the observation that both the LL and the RL versions result in a memory bottleneck due to an initial construction of the entire frontal matrix as dense tiles, the problem sizes that can be executed within a multifrontal solver are limited, see Section 5. Therefore, we decided to implement a memory efficient version where we combine aspects of the LL and the RL version which results in a column-wise construction of the frontal matrices, see Algorithm 4 and Fig. 2c. We call this variant the hybrid BLR version. For the hybrid version only the neighboring $#col_{max}$ block columns are created and subsequently the compression and solve steps are executed only on the neighboring $#col_{max}$ row tiles, see Algorithm 4 for more details. Algorithm 4 is based on the assumption that p is integer multiple of b for simplicity, but our code can easily handle any value of p in general. The experiments in Section 5 are based on $#col_{max} =$ number of columns in the 2D MPI process grid such that each MPI process is involved in updating one local column.

The benefit of using the hybrid version over the RL or the LL version is the reduced memory consumption which eventually enables to solve larger problem sizes, see Section 5.

3.2.1 Parallel layout of the hybrid BLR matrix. In our implementation, the hybrid BLR matrix is parallelized with a two-dimensional process grid using the available processes in the MPI communicator, and the matrix is distributed over this grid using a 2D block-cyclic data layout, similar to the ScaLAPACK layout but with non-uniform block sizes. The process grid is $P_r \times P_c$ with $P_r = |\sqrt{P}|$ and $P_c = P/P_r$, with at most $|\sqrt{P}| - 1$ idle processes, where P represents the number of available MPI processes. In a parallel setting, the block columns for the hybrid BLR algorithm are distributed among all available MPI processes in the grid, based on the 2D block-cyclic data layout with $col_{max} = P_c$. After the block columns are created locally, the factor, compression and solve steps are executed on the neighboring #col_{max} row tiles. In between each of these four tasks, the MPI processes communicate their updates to the processes in the same row communicator and in the same column communicator.

Fig. 3 presents the algorithmic steps for a parallel hybrid BLR method with twelve MPI processes arranged as four processes for each column times three processes for each row. In particular, each process constructs its local tiles within the first $col_{max} = 3$ columns. The process that owns the tile in the first row- and column executes the factorization step; this is followed by two broadcast operations to share the updated tile with all processes of row one and all processes of column one. Afterwards the compression and solve steps are executed for all tiles in the first column as well as the two tiles in row one. The update operation as described in Algorithm 4 is executed for all tiles in columns two and three. These four tasks are repeated over and over such that all tiles that have been constructed already, are updated. Afterwards these steps are executed again for the next $col_{max} = 3$ columns.

520 Manuscript submitted to ACM

Sparse approximate multifrontal factorization with composite compression methods

521		
522		
523		
524		
525		
526		
527		
528		
529		
530		
531		
532		
533		
534		
535		
536		
537		
538		
539		
540		
541		
542		
543		
544		
545		
546		
547		
548		
549		
550		
551		
552		
553		
554		
555		
556		
557		
558		
559		
560		
561		
562		
564		
565		
566		
567		
568		
569		
570		
571		
572		



Fig. 3. Hybrid BLR algorithm with 12 processes arranged as 4x3 process grid and $col_{max} = 3$. Manuscript submitted to ACM

Communication Cost Comparison for BLR Factorization. As described in Section 3.2.1 the parallel layout of the 3.2.2 BLR matrix follows a 2D block-cyclic data layout, similar to the ScaLAPACK layout but with varying block sizes. As visualized in Fig. 2 and Fig. 3 the BLR(RL), BLR(LL) and BLR(Hybrid) algorithms consist of various different execution steps which lead to a different communication pattern in a parallel setting.

The BLR(RL) algorithm consists of a repetition of the following steps:

- (1) LU factorization followed by a broadcast of the LU factored tile along the row processes and the column processes. In addition, the pivot elements are shared along the row processes with an additional broadcast.
 - (2) After the compression and solve step, the row of updated tiles as well as the column of updated tiles are broadcast along all row and column processes using two broadcast operations times the number of column/row processes each.

Step 1 consists of three broadcast operations, while step 2 consists of two broadcast operations times the number of column/row processes each. These steps are repeated for each row of tiles, i.e.:

$$(\lceil n/b \rceil - 1) \times 3 \text{ bcast_ops},$$

 $(\lceil n/b \rceil - 1) \times (2 \times (P_r + P_c)) \text{ bcast_ops},$

with n = size of F_{11} , b = size of tile, $P_r = number$ of row processes and $P_c = number$ of column processes.

The BLR(LL) algorithm consists of a repetition of the following steps:

- (1) LU factorization followed by a broadcast of the LU factored tile along the row processes and the column processes. In addition, the pivot elements are shared along the row processes with an additional broadcast.
- (2) After the compression and solve step, the row and column of updated tiles as well as all previous updated rows and columns are communicated to the necessary MPI processes using four broadcast and two send operations times the number of column/row processes each. While the broadcast operations are used for the communication of one tile along columns and one tile along rows, the send operations are used for the remaining tiles of the rows and columns, which results in two send operations times the number of column/row processes each.

Step 1 consists of three broadcast operations, while step 2 consists of four broadcast operations times the number of previous updated columns/rows and two send operations times the number of column/row processes each times the number of previous updated columns/rows, leading to a total of $3 + (i \times 4)$ broadcast and $i \times (P_r \times 2 + P_c \times 2)$ send operations, with *i* current phase of the factorization. These steps are repeated for each row of tiles $i = 1 : \lfloor n/b \rfloor - 1$, i.e.:

610
$$(\lceil n/b \rceil - 1) \times 3 \text{ bcast_ops},$$

$$|n/b|-1$$

$$\sum_{i=1}^{\lceil n/b \rceil - 1} (i \times (2 \times (P_r + P_c))) \text{ send_ops.}$$

The formula indicate that BLR(RL) uses less broadcast operations than BLR(RL). However, BLR(RL) consists of broadcast operations only, while BLR(LL) needs additional send operationswhich leads to higher communication cost overall for BLR(LL) compared to the BLR(RL) variant.

The BLR(Hybrid) algorithm consists of a repetition of the following two stages:

(1) Construct columns and if columns to the left already exist, execute the following step

Manuscript submitted to ACM

(a) After compression and solve step of the new partial row, the updated tiles are communicated to the necessary MPI processes using two broadcast operations times the number of column processes. In addition, the previously updated block columns are communicated to the necessary MPI processes using two broadcast operations times the number of row processes.

This first stage is executed for all previously constructed columns, i.e. four broadcast operations (times the number of row/column processes) times the number of previously constructed columns. The second stage consists of the following operations:

- (2) (a) LU factorization followed by a broadcast of the LU factored tile along the row processes and the column processes. In addition, the pivot elements are shared along the row processes with an additional broadcast. Even though only a part of a row is constructed in each step, all processes are involved in the update and communication steps.
 - (b) After the compression and solve step of one partial row and one column, the updated tiles are communicated to the necessary MPI processes using two broadcast operations times the number of row/column processes each.

Step (a) of stage 2 consists of three broadcast operations, while step (b) consists of two broadcast operations. For both stages the number of communication operations sum up to:

 $\lceil I/col_{\max} \rceil \times col_{\max} \times 3$ bcast_ops,

$$\sum_{i=1}^{l} (col_{\max} \times i \times (2 \times P_r + 2 \times P_c)) \text{ bcast_ops},$$

with $I = \lceil \lceil n/b \rceil / col_{\max} \rceil$.

In summary, we notice additional communication cost for the two BLR variants (LL) and (Hybrid) compared to the BLR(RL) variant. BLR(RL) and BLR(Hybrid) consist of broadcast operations only, while BLR(LL) needs additional send operations. The formula indicate that BLR(RL) uses the least amount of broadcast operations and BLR(Hybrid) uses significantly more than the other two variants, which is due to additional broadcast operations needed for each newly constructed set of columns.

Incorporating BLR fronts in the sparse multifrontal solver adds some additional challenges that we discuss in Section 4.

3.3 Hierarchically Off-Diagonal Butterfly Compression

The HODBF matrix representation [28] is the butterfly extension of the HODLR matrix, i.e., \mathcal{H} -matrix with weak admissibility condition [16], that means every off-diagonal block is compressed. The clustering into blocks as well as the weak admissibility condition in the context of the multifrontal solver are discussed in detail in Section 4. In what follows, we briefly describe the HODBF format, which is used in Section 4 to construct the quasi-linear complexity multifrontal solver.

HODBF starts with a hierarchical clustering of the row and column indices of an $n \times n$ matrix A into $L = O(\log n)$ levels. At the leaf level L, we have a partitioning $T_1^L, T_2^L, \dots, T_{2^L}^L$, the same as the BLR matrix; at the next level, we have a new partitioning $T_1^{L-1}, \dots, T_{2^{L-1}}^{L-1}$ by combining every two adjacent clusters/nodes at level L into one. For a given cluster at level l < L, T_k^l , we use \mathcal{T}_k^l to denote the subtree rooted at T_k^l and $\mathcal{T}_H = \mathcal{T}_0^0$ (see Fig. 4).



Fig. 4. Illustration of a 4-level hierarchically off-diagonal butterfly matrix. The root node is at level l = 0, all the leaf nodes are at level L = 3. The two largest off-diagonal blocks are approximated using 2-level butterfly matrices. The 4 off-diagonal blocks one level down in the hierarchy are approximated using a 1-level butterfly ($U^1B^1V^0$). Finally, the smallest off-diagonal blocks are approximated as 0-level butterfly matrices.

It follows that off-diagonal blocks of the HODBF matrix, representing interactions between two distinct level-*l* clusters with subtrees \mathcal{T}_S and \mathcal{T}_O , are compressed as butterfly representations, while the diagonal blocks at the leaf level $D_{\tau} = A(T_{\tau}^L, T_{\tau}^L)$ for cluster τ are stored as regular dense matrices (see Fig. 4). The butterfly representation of $L_b = L - l$ levels for the block $K = A(T_O, T_S)$ reads:

$$K \approx U^{L_b} R^{L_b - 1} \cdots R^h B^h (W^h)^* \cdots (W^1)^* (V^0)^*,$$
(2)

where $h = L_b/2$, U^{L_b} and V^0 are block diagonal matrices, $R^{L_b-1}, \dots, R^h, B^h, W^1, \dots, W^h$ are all block diagonal matrices after certain predefined permutations. Typically, butterfly representation of an $m \times n$ matrix K contains at most $O(n \log n)$ nonzeros. In practice, one can rapidly constructs such a representation by evaluating $O(n \log n)$ matrix entries from K. Once constructed, the HODBF matrix can be inverted efficiently with randomized algorithms [28]. For more details on butterfly decomposition and its components, please refer to [27, 29].

3.3.1 Parallel HODBF Layout . In our implementation, the HODBF matrix is distributed using a given MPI communicator. The overall 1D block layout can be summarized as follows: starting with $D_{\tau} = A$ at the root τ of \mathcal{T}_H , the given communicator sharing D_{τ} is split into two sub communicators of similar number of processes. The two sub communicators store and compute D_{τ_1}, B_{τ_1} and D_{τ_2}, B_{τ_2} , respectively. Layouts of B_{τ_1} and B_{τ_2} follow those described in [27]. Each of the two sub communicators is further recursively split unless the communicator has only one process that stores and computes D_{τ} . Once constructed following this layout, the HODBF representation can also be inverted following this layout. See more detail on parallelization in [41].

4 RANK STRUCTURED MULTIFRONTAL FACTORIZATION

Algorithm 5 outlines the rank-structured multifrontal factorization using BLR compression for medium sized fronts and HODBF compression for large sized fronts. Since the more complicated HODBF matrix format has larger overhead for smaller matrices compared to the BLR compression, HODBF compression is only used for fronts corresponding to a separator size larger than a certain threshold n_{Hmin}. In addition, BLR compression is only used for fronts corresponding to a separator size smaller than n_{Hmin} and larger than a threshold n_{Bmin} . We have the option to add zfp compression [24] for small fronts, below threshold n_{Bmin} and larger than or equal to a separator size of 8. All fronts corresponding to a separator size smaller than 8 will not be compressed. The advantage of adding zfp compression is an additional Manuscript submitted to ACM

731

732

733 734

735

736

739

741

749

750

751

752

753 754

755

756 757

758

759

760

761 762

763

764

765

766 767

768

769

770

777

reduction in memory consumption, however the solve time increases due to the decompression of zfp fronts during the solve step, see for example Fig. 8c and Fig. 8e.

Typically, the larger fronts are found closer to the root of the multifrontal assembly tree. This is illustrated in Fig. 5 for a small regular $5 \times 5 \times 4$ mesh (Fig. 5a), and Fig. 5c shows the corresponding multifrontal assembly tree, where only the top three fronts are compressed using HODBF, the next two levels down in the assembly tree consist of BLR compressed fronts and the remaining fronts are compressed with zfp.



Fig. 5. (a) The top two levels of nested dissection for a $5 \times 5 \times 4$ mesh. The top three separators are S^0 , S_0^1 and S_1^1 . (b) The root separator S^0 is a vertical plane of 20 points, which is recursively bisected to define level 1 (T_0^1 to T_1^1) and level 2 (T_0^2 to T_3^2) of the hierarchical matrix partitioning. (c) The root separator corresponds to the top level front, and its HODBF partitioning is defined by the recursive bisection of the root separator, as shown in (b), similar for the next level down; the following two levels in the assembly/frontal tree consists of BLR fronts and the remaining fronts are compressed with zfp.

We now discuss the construction as well as the partial factorization of frontal matrices within a multifrontal solver with a focus on BLR and HODBF compressed fronts. Recall from Section 2 that a front F_{τ} is built up from elements of the reordered sparse input matrix A and the contribution blocks of the children of the front in the assembly tree: $F_{22;\nu_1}$ and $F_{22;\nu_2}$, where ν_1 and ν_2 are the two children of τ . Since the multifrontal factorization traverses the assembly tree from the leaves to the root, these children contribution blocks might already be compressed using the BLR or the HODBF format. Hence, extracting frontal matrix elements requires getting them from fronts previously compressed. There are four different options that we describe in detail in the following subsections, Section 4.1 - Section 4.4: extracting from HODBF to construct HODBF, extracting from BLR to construct HODBF, extracting from LL/RL BLR to construct LL/RL BLR, extracting from hybrid BLR to construct hybrid BLR. In addition to the above mentioned extracting operations, also called extend-add operations, there are extracting operations that we don't discuss here either because they are straightforward operations, like extracting from dense matrix to construct a matrix in BLR form, or because they are similar to the other operations that we explain in detail in the following sections, like extracting from dense or zfp matrix to construct HODBF.

771 The block clustering for the BLR representation of F_{11} , and the HODBF cluster tree for the F_{11} part of a front, are 772 defined by performing a recursive bisection (not to be confused with nested dissection), using METIS, of the graph 773 corresponding to $A(I_{\tau}^s, I_{\tau}^s)$, where I_{τ}^s denotes the index sets of F_{11} . Recursive bisection leads to a tree structure that can 774 be used to define the HODBF cluster tree and a corresponding permutation of the rows/columns of F_{11} . For BLR, only 775 the leaves of this tree are considered for the definition of the blocks. 776

4.0.1 BLR admissibility condition. The admissibility condition determines which blocks in the BLR matrix should 778 779 be considered as compressible. We implement two different admissibility conditions. As the default strategy, each 780 Manuscript submitted to ACM

off-diagonal block is compressed, but if the rank of an off-diagonal block is too large, such that the compression does not decrease memory consumption, then that block is stored in its dense format. The diagonal blocks B_{ii} are always stored as full-rank matrices ($\tilde{B}_{ii} = B_{ii}$). We also provide an alternative strategy, called strong admissibility, where we use the graph of $A(I_{\tau}^{s}, I_{\tau}^{s})$ to determine whether a block is admissible. For BLR with strong admissibility, we say that an interaction $\sigma \times \tau$ is inadmissible if $\sigma \equiv \tau$, or if the matrix block $B_{\sigma\tau}$ contains a nonzero entry coming from sparse matrix *A*. We found that trying to compress each block performed slightly better in terms of both compression ratio and runtime.

4.1 HODBF with HODBF Children Nodes

Extracting an HODBF frontal matrix to construct an HODBF matrix looks like:



with F_{11} and F_{22} compressed as HODBF, and F_{12} and F_{21} compressed as butterfly. For the construction of an HODBF front with HODBF children, the following tasks need to be executed:

- (1) Since fronts are constructed as a combination (extend-add) of other smaller fronts, a list of submatrices needs to be extracted from other fronts which are already compressed using HODBF. Therefore it is critical for performance to use an efficient algorithm to extract a list of submatrices from a butterfly matrix. This is presented as extract_BF in [27]. The *F*₁₁ block of *F* ≡ *F*_τ is compressed as an HODBF matrix, see [27] for details, which uses the extract_BF routine to extract elements from *F*₁₁ = *A*(*I*^s_τ, *I*^s_τ) ← *F*_{22;ν1} ← *F*_{22;ν2}, see line 32 in Algorithm 5. Note that in this case, the extend-add operation just requires checking whether the required matrix entries appear in the sparse matrix, or in the child contribution blocks, and then adding those different contributions together.
- (2) Line 33 approximates F_{11}^{-1} from the butterfly representation of F_{11} , see [27] for details.
- (3) Lines 34 and 35, the off-diagonal blocks/fronts F₁₂ and F₂₁ are each approximated as a single butterfly matrix, using routines to extract elements from A(I^s_τ, I^u_τ) ↔ F_{22;ν1} ↔ F_{22;ν2} and A(I^u_τ, I^s_τ) ↔ F_{22;ν1} ↔ F_{22;ν2} respectively. For F₁₂, the tree T_H corresponding to F₁₁ is used as T_O, and the tree corresponding to F₂₂ is used for T_S, and vice versa for F₂₁.
- (4) The final step for this front is to construct the contribution block of τ. F₂₂ as an HODBF matrix, again using element extraction, now from F_{22;ν1} ↔ F_{22;ν2} S, where F_{22;ν1} and F_{22;ν2} are in HODBF form and S = F₂₁F₁₁⁻¹F₁₂ is a single butterfly matrix compressed via the randomized algorithm in [27]. S can be released as soon as the contribution block has been assembled, and the contribution block is kept in memory until it has been used to assemble the parent front.

- 832 Manuscript submitted to ACM

Sparse approximate multifrontal factorization with composite compression methods

Inp	put: $A \in \mathbb{R}^{N \times N}$, $b \in \mathbb{R}^N$	
Ou	tput: $x \approx A^{-1}b$	
1:	$A \leftarrow P(D_r A D_c Q_c) P^\top \qquad \qquad \triangleright \text{ scaling, and permut}$	tation for stability and fill reduction
2:	$A \leftarrow \hat{P}A\hat{P}^{\top} \qquad \qquad \triangleright$	rank-reducing separator reordering
3:	Build assembly tree: define I_{τ}^{s} and I_{τ}^{u} for every frontal matrix F_{τ}	
4:	for nodes τ in assembly tree in topological order do	
5:	if dimension $(I_{\tau}^{s}) < 8$ then	
6:	construct F_{τ} as a dense matrix	▶ Algorithm 1
7:	else if dimension $(l_{\tau}^{s}) < n_{\text{Bmin}}$ then	
8:	construct F_{τ} as a dense matrix	► Algorithm 1
9:	if ztp enabled then	
10:	compress as zfp matrix	
11:		
12:	else if dimension $(I_{\tau}^{r}) < n_{\text{Hmin}} \&\& \text{ RL- or LL-BLR then}$	
13:	$F_{\tau} \leftarrow \begin{bmatrix} A(I_{\tau}, I_{\tau}) & A(I_{\tau}, I_{\tau}) \\ A(I_{\tau}^{u}, I_{\tau}^{z}) & 0 \end{bmatrix} \Leftrightarrow F_{22;\nu_{1}} \Leftrightarrow F_{22;\nu_{2}}$	
14:	$P_{\tau}L_{\tau}U_{\tau} \leftarrow F_{11}$	LU with partial pivoting
15:	$F_{11}, F_{12}, F_{21} \leftarrow \text{BLR compress}(F_{11}, F_{12}, F_{21})$	
16:	$F_{12} \leftarrow L_{\tau}^{-1} P_{\tau}^{\top} F_{12}$	
17:	$F_{21} \leftarrow F_{21} U_{\tau}^{-1}$	
18:	$F_{22} \leftarrow F_{22} - F_{21}F_{12}$	⊳ Schur update
19:	else if dimension $(I_{\tau}^{s}) < n_{\text{Hmin}}$ & Hybrid-BLR then	
20:	for columns <i>C</i> in frontal matrix F_{τ} corresponding to I_{τ}^{s} do	,
21:	$F_{11;C} \leftarrow \text{BLR compress}\left(A(I_{\tau}^{\text{Spartial}}, I_{\tau}^{\text{Spartial}}) \Leftrightarrow F_{22;\nu_{j};\text{partial}} \leftrightarrow F_{22;\nu_{j}}\right)$;partial)
22:	$P_{\tau}L_{\tau}U_{\tau} \leftarrow F_{11:C}$	► LU with partial pivoting
23:	$F_{11:C}, F_{21:C} \leftarrow \text{BLR compress}(F_{11:C}, F_{21:C})$	1 1 0
24:	$F_{21\cdot C} \leftarrow F_{21\cdot C} U_{\tau}^{-1}$	
25:	end for	
26:	for columns <i>C</i> in frontal matrix F_{τ} corresponding to I_{τ}^{u} do	
27:	$F_{12:C} \leftarrow \text{BLR compress}(F_{12:C})$	
28:	$F_{12:C} \leftarrow L_{\tau}^{-1} P_{\tau}^{\top} F_{12:C}$	
29:	$F_{22:C} \leftarrow F_{22:C} - F_{21}F_{12}$	⊳ Schur update
30:	end for	-
31:	else	
32:	$F_{11} \leftarrow \text{HODBF compress} \left(A(I_{\tau}^{s}, I_{\tau}^{s}) \Leftrightarrow F_{22;\nu_{1}} \Leftrightarrow F_{22;\nu_{2}} \right)$	
33:	$F_{11}^{-1} \leftarrow \text{HODBF invert}(F_{11})$	
34:	$F_{12} \leftarrow \text{butterfly compress} \left(A(I_{\tau}^{s}, I_{\tau}^{u}) \Leftrightarrow F_{22;\nu_{1}} \leftrightarrow F_{22;\nu_{2}} \right)$	
35:	$F_{21} \leftarrow \text{butterfly compress} \left(A(I_{\tau}^{u}, I_{\tau}^{s}) \Leftrightarrow F_{22;\nu_{1}} \leftrightarrow F_{22;\nu_{2}} \right)$	
36:	$S \leftarrow \text{butterfly compress} (F_{21}F_{11}^{-1}F_{12})$	
37:	$F_{22} \leftarrow \text{HODBF compress} \left(F_{22:1h} \leftarrow F_{22:1h} - S \right)$	⊳ Schur update
38:	end if	
39:	end for	
40	$r \leftarrow \text{GMRFS}(A \mid h \mid M : u \leftarrow D_{-}O_{-}P^{\top}\hat{P}^{\top} \text{ bwd-solve}(\text{fwd-solve}(\hat{P}PD_{-}u)))$	

4.2 HODBF with BLR Children Nodes

Extracting a BLR(Hybrid) frontal matrix to construct an HODBF matrix is similar to an extraction from HODBF and looks like:



For the construction of an HODBF front with BLR children, the following tasks need to be changed compared to Section 4.1:

For line 32 in Algorithm 5 a different routine to extract elements from F₁₁ = A(I^s₇, I^s₇) ↔ F_{22;ν1} ↔ F_{22;ν2} needs to be executed. For RL- and LL-BLR the extract is straightforward since F_{22;ν1} and F_{22;ν2} remain dense at the time of extraction. Hence, we simply extract the elements from a dense submatrix using a scatter operation. In case of the Hybrid-BLR children F_{22;ν1} and F_{22;ν2} are compressed. Hence, more computational effort is needed to extract from compressed BLR tiles.

- Lines 34 and 35, the F₁₂ and F₂₁ front off-diagonal blocks are each approximated as a single butterfly matrix, using routines to extract elements from A(I^s_τ, I^u_τ) ↔ F_{22;ν1} ↔ F_{22;ν2} and A(I^u_τ, I^s_τ) ↔ F_{22;ν1} ↔ F_{22;ν2} respectively, where F_{22;ν1} and F_{22;ν2} are in compressed form as in Hybrid-BLR or dense as in LL/RL-BLR.
- The final step to construct the contribution block of *τ*, *F*₂₂, as an HODBF matrix, uses element extraction, now from *F*_{22;ν1} ↔ *F*_{22;ν2} − *S*, where *F*_{22;ν1} and *F*_{22;ν2} are in BLR form and *S* is a single butterfly matrix.

Note that for BLR fronts, the extend-add operation just requires checking whether the required matrix entries appear in the sparse matrix, or in the child contribution blocks, and then adding those different contributions together.

4.3 BLR with BLR Children Nodes with Immediate Construction of All Tiles (LL/RL-BLR)

Extracting a BLR frontal matrix to construct a BLR matrix looks like:



and the following tasks need to be executed for the RL and LL BLR versions:

- (1) The $F \equiv F_{\tau}$ frontal matrix is constructed as a BLR matrix, which implies that it is separated into smaller dense tiles. First, we apply extend-add operations to update $F_{\tau} \leftarrow \begin{bmatrix} A(I_{\tau}^{s}, I_{\tau}^{s}) & A(I_{\tau}^{s}, I_{\tau}^{u}) \\ A(I_{\tau}^{u}, I_{\tau}^{s}) & 0 \end{bmatrix} \stackrel{\uparrow}{\leftrightarrow} F_{22;\nu_{1}} \stackrel{\uparrow}{\leftrightarrow} F_{22;\nu_{2}}$, see line 13 in Algorithm 5. Since $F_{22;\nu_{1}}$ and $F_{22;\nu_{2}}$ are not compressed the extract is straightforward, we simply extract the elements from a dense submatrix using a scatter operation.
- ⁹³² (2) Line 14 computes the *LU* decomposition of F_{11} tile by tile.
- (3) Line 15 compresses the off-diagonal tiles of F_{11} and the tiles of F_{12} , F_{21} via QR factorization with column pivoting.
- 935 (4) Lines 16 and 17, the off-diagonal blocks F_{12} and F_{21} are updated.
- 936 Manuscript submitted to ACM

 (5) The final step for this front is to construct the contribution block F_{22} as a BLR matrix with dense tiles, see line 18. The contribution block is kept in memory until it has been used to assemble the parent front.

4.3.1 Construction of a parallel BLR parent front with immediate construction of all tiles. In our implementation, the BLR matrix is distributed among MPI communicators, i.e., a certain amount of BLR tiles is assigned to each MPI communicator. The parallel construction is executed for each MPI communicator. First, each MPI communicator collects their contribution from children fronts $F_{22;v_1}$ and $F_{22;v_2}$. This is followed by an all-to-all exchange of BLR tiles, such that each MPI process has access to the necessary update tiles. Afterward, each MPI communicator updates their tiles within the parent front. A more detailed description and a visualization can be found for the similar hybrid BLR algorithm with fragmentary construction of the BLR matrix, see Section 4.4.

4.4 BLR with BLR Children Nodes with Fragmentary Construction of the BLR Parent Matrix (hybrid-BLR)

Extracting a BLR frontal matrix to construct a parent BLR matrix in fragments (few columns at a time) is similar to a BLR construction where we construct all tiles immediately, see Section 4.3. However it requires some algorithmic updates as represented in the following:



The tasks to be executed need to be updated as described below:

(1) The extend-add operations to update $F_{\tau} \leftarrow \begin{bmatrix} A(I_{\tau}^{s}, I_{\tau}^{s}) & A(I_{\tau}^{s}, I_{\tau}^{u}) \\ A(I_{\tau}^{u}, I_{\tau}^{s}) & 0 \end{bmatrix} \Leftrightarrow F_{22;\nu_{1}} \Leftrightarrow F_{22;\nu_{2}}$ requires a different routine to only update the columns that have been constructed. Additional extend-add steps need to be executed every time block columns are added to the front. Eq. (6) visualizes the extend-add operation involving three newly constructed columns of the parent front F_{τ} and columns of the child fronts $F_{22;\nu_{1}}$ and $F_{22;\nu_{2}}$, highlighted in yellow. $F_{22;\nu_{1}}$ and $F_{22;\nu_{2}}$ are in BLR matrix form, fully constructed and compressed. Note, only the columns of the contribution blocks corresponding to the data points represented in the columns of the parent front need to be extracted for the extend add operation. The extract operation for the required block columns can be executed as follows:

- If the tiles of $F_{22;\nu}$ needed for the extend-add operation are compressed, we first decompress the block columns and then apply a scatter operation to construct F_{τ} .
- If the tiles of *F*_{22;ν} needed for the extend-add operation have been decompressed already, we simply apply a scatter operation to construct *F*_τ.

These steps need to be repeated for each column that is constructed.

(2) The remaining steps are adjusted such that only the constructed block columns are considered. A loop is added for partial construction and update of block columns, see lines 20 and 26 in Algorithm 5.

4.4.1 F_{22} compression. For the final step, the so-called update step of the contribution block F_{22} of a frontal matrix, see line 29 in Algorithm 5, we include an additional compression step for the hybrid BLR algorithm. In contrast to the BLR matrices with immediate construction of all tiles, compare Section 4.3, we do a simple compression step after the General Matrix Multiply (GeMM) operation is executed.

4.4.2 Construction of a parallel hybrid BLR parent front. In our implementation, the BLR matrix is distributed among MPI communicators, i.e., a certain amount of BLR tiles is assigned to each MPI process. The parallel construction is executed for each MPI communicator. First, each MPI process collects their contribution from children fronts $F_{22;\nu_1}$ and $F_{22;\nu_2}$ needed for the update of the block columns in the parent front. This is followed by an all-to-all exchange of BLR tiles, such that each MPI process has access to the necessary update tiles. Afterward, each MPI communicator updates their tiles within the columns of the parent front.

As discussed in Section 3.2.2 the communication cost of the BLR matrix incorporated in the sparse multifrontal solver vary based on the BLR variant. We concluded higher communication cost for a BLR(Hybrid) matrix compared to BLR(RL) and BLR(LL) in Section 3.2.2. In addition, when BLR is used within a multifrontal solver, BLR(Hybrid) requires additional communication steps for the extend-add operation sending updates from children fronts to parent frontal matrices. These observations lead to an increased factorization time as can be seen in Section 5.

997

998 999

1000

5 EXPERIMENTAL RESULTS

1005 Experiments reported here are all performed on the Haswell nodes of the Cori machine, a Cray XC40, at the National 1006 Energy Research Scientific Computing Center in Berkeley. Each of the 2, 388 Haswell nodes has two 16-core Intel 1007 Xeon E5-2698v3 processors and 128GB of 2133MHz DDR4 memory. The approximate multifrontal solver is used 1008 1009 as a preconditioner for restarted GMRES(30) with modified Gram-Schmidt and a zero initial guess. Unless noted 1010 otherwise, all experiments are performed in double precision with absolute or relative stopping criteria $||u_i|| \leq 10^{-10}$ or 1011 $||u_i||/||u_0|| \le 10^{-6}$, where $u_i = M^{-1}(Ax_i - b)$ is the residual at Krylov iteration *i*, with M the approximate multifrontal 1012 factorization of A. We use iterative refinement instead of GMRES for the exact multifrontal solver, which is also called 1013 1014 multifrontal solver with no compression. For simplicity all experiments are in double precision. For a discussion on 1015 mixed precision iterative refinement for approximate sparse solvers see [3]. 1016

1018 5.1 Visco-Acoustic Wave Propagation

We first consider the 3D visco-acoustic wave propagation governed by the Helmholtz equation

1030

1031

1032

1017

1019

 $\left(\sum_{i} \rho(\mathbf{x}) \frac{\partial}{\partial x_{i}} \frac{1}{\rho(\mathbf{x})} \frac{\partial}{\partial x_{i}}\right) p(\mathbf{x}) + \frac{\omega^{2}}{\kappa^{2}(\mathbf{x})} p(\mathbf{x}) = -f(\mathbf{x}).$ (7)

Here $\mathbf{x} = (x_1, x_2, x_3)$, $\rho(\mathbf{x})$ is the mass density, $f(\mathbf{x})$ is the acoustic excitation, $p(\mathbf{x})$ is the pressure wave field, ω is the angular frequency, $\kappa(\mathbf{x}) = v(\mathbf{x})(1 - i/(2q(\mathbf{x})))$ is the complex bulk modulus with the velocity $v(\mathbf{x})$ and quality factor $q(\mathbf{x})$. We solve Eq. (7) by a finite-difference discretization on staggered grids using a 27-point stencil and 8 PML absorbing boundary layers [36]. This requires direct solution of a sparse linear system where each matrix row contains 27 nonzeros, whose values depend on the coefficients and frequency in Eq. (7).

We consider a cubed domain with $v(\mathbf{x}) = 4000$ m/s, $\rho(\mathbf{x}) = 1$ kg/m³, $q(\mathbf{x}) = 10^4$. The frequency is set to $\omega = 8\pi$ Hz and the grid spacing is set such that there are 15 grid points per wavelength.

First, we vary *N* from 200³ to 420³ and compare four types of multifrontal solvers: "no compression"- exact solver, "BLR(RL)"- BLR compression with RL variant, "BLR(LL)"- BLR compression with LL variant and "BLR(Hybrid)"- BLR compression with hybrid variant, see Fig. 6. All fronts corresponding to separators with size $n_{\text{Bmin}} \ge 200$ are compressed with tolerance $\varepsilon = 10^{-3}$. For all experiments, the relative error was $||x - \tilde{x}||_2 / ||\tilde{x}||_2 < 10^{-5}$. We observed that all variants of the BLR multifrontal solver outperform the exact solver in terms of factor flops, factor time, factor nonzeros and peak memory, see Fig. 6. The iteration counts are shown in Fig. 6e for BLR(Hybrid), see Table 1 for iteration counts Manuscript submitted to ACM



Fig. 6. Results for high frequency 3D Helmholtz using the $BLR(10^{-3})$ multifrontal solver with GMRES. (a) Flop counts for factorization. (b) CPU time for factorization. (c) Memory usage for factorization, i.e. factor nonzeros/factor memory. (d) Flop counts for solve per iteration in GMRES. (e) CPU time for solve. The number of GMRES iterations are shown at every data point (additional data in Table 1). (f) Peak working memory. We use 64 compute nodes, with 4 MPI ranks per node and 8 OpenMP threads per MPI process.

of other variants. The BLR(RL) version outperforms the other two BLR variants in terms of factor time. This can be explained with additional communication cost for the other two BLR variants. As explained in Section 3.2.2, BLR(RL) and BLR(Hybrid) consist of broadcast operations only, while BLR(LL) needs additional send operations. The formula in Section 3.2.2 indicate that BLR(RL) uses the least amount of broadcast operations and BLR(Hybrid) uses significantly more than the other two variants, see Fig. 7a, which is due to additional broadcast operations needed for each newly constructed set of columns. Fig. 7b shows that the BLR(LL) variant dominates the amount of send operations since they are not used within fronts in the other two variants. The BLR(Hybrid) algorithm has lower peak memory consumption which allows to solve larger problem sizes. BLR(RL) and BLR(LL) run out of memory when solving Eq. (7) with sizes larger than 320³. Among all three variants, BLR(RL) requires the least amount of communication and is the fastest, but it requires the largest amount of memory.

problem size	200 ³	220 ³	250 ³	280 ³	300 ³	320 ³	350 ³	380 ³	400^{3}	420 ³	
BLR(Hybrid)	5	4	5	5	4	6	4	5	5	8	
BLR(RL)	4	4	5	5	4	6					
BLR(LL)	4	4	5	5	4	6					
Table 1. Iteration counts corresponding to results in Fig. 6											

Next, we consider a problem with size $N = k^3$, with k ranging from 200 to 420 and compare the performance of four multifrontal solver, with BLR, HODBF, HODBF_BLR and HODBF_BLR_ZFP, see Fig. 8. For the BLR multifrontal solver, Manuscript submitted to ACM



Fig. 7. Amount of communication for the BLR(LL), BLR(RL) and BLR(Hybrid). (a) Number of broadcast operations. (b) Number of send operations. BLR(RL) and BLR(Hybrid) require zero send operations, see Section 3.2.2.

we use the BLR(Hybrid) variant which allows to solve problems of size up to 420^3 . We set the tolerance $\varepsilon = 10^{-2}$ for HODBF and $\varepsilon = 10^{-3}$ for BLR, respectively. The zfp compressed fronts use 16 bitplanes in HODBF BLR ZFP.

For HODBF compression, all fronts corresponding to separators with sizes $n_{\text{Hmin}} \geq 7$ K are compressed. For HODBF_BLR_ZFP and HODBF_BLR compression, all fronts corresponding to separators with sizes $n_{\text{Hmin}} \ge 15K$ are compressed with HODBF, all fronts corresponding to separators $200 \le n_{Bmin} \le 15$ K are compressed with BLR, and all fronts corresponding to separators smaller than 200 are either not compressed or compressed with zfp compression. For all experiments, the relative error was $\|x - \tilde{x}\|_2 / \|\tilde{x}\|_2 < 10^{-5}$. Compared to the $O(N^2)$ computation and $O(N^{4/3})$ memory complexities using the exact multifrontal solver and the BLR multifrontal solver, we observe the predicted $O(N \log^2 N)$ computation and O(N) memory complexities (see Fig. 8a-Fig. 8e) for the HODBF multifrontal solver variants. The iteration counts are shown in Fig. 8e for BLR(Hybrid) and HODBF BLR ZFP, see Table 2 for iteration counts of the other multifrontal solvers.

Note that HODBF BLR ZFP and HODBF BLR outperform the other solvers in terms of factor time, factor flops and solve flops. HODBF_BLR_ZFP outperforms all solvers in terms of factor nonzeros. Due to its lower peak memory consumption the BLR(Hybrid) multifrontal solver allows to solve larger problem sizes up to 420³ with a low solve time.

problem size	200 ³	220 ³	250 ³	280 ³	300 ³	320 ³	350 ³	380 ³	400 ³	420^{3}
BLR(Hybrid)	5	4	5	5	4	6	4	5	5	8
BLR(RL)	4	4	5	5	4	6				
HODBF	6	6	9	8	15					
HODBF_BLR(Hybrid)	8	8	29	41	20	43	73	196	317	
HODBF_BLR(Hybrid)_ZFP	8	8	29	41	20	43	72	179	354	
							1		_	

Table 2	Iteration	counts	corresponding	to	results	in	Fig	8
1abic 2	nucration	counts	CONCODONUNE	ιu	results		116.	L.

5.2 Singularly Perturbed PDE

Next, we consider the following three-dimensional singularly perturbed reaction-diffusion differential equation (SPDE) that arises in fluid dynamics, computational chemistry, and biological applications (see, for instance, [40]):

> $-\delta^2 \Delta u + u = f$, on $\Omega = (0, 1)^3$, and $u(\partial \Omega) = g$, (8)

where the *perturbation parameter* δ is small and positive, *q* and *f* are some given functions. Solving large sparse systems arising from finite difference discretizations, even for the two-dimensional analogue of (8), is a challenging task. For example, [30] showed that standard Cholesky-based solvers exhibit poor performance when δ is small. The underlying Manuscript submitted to ACM





Fig. 8. Results for high frequency 3D Helmholtz using the four different multifrontal solver with GMRES. (a) Flop counts for factorization. (b) CPU time for factorization. (c) Memory usage for the factorization, i.e. factor nonzeros/factor memory. (d) Flop counts for solve per iteration in GMRES. (e) CPU time for solve. The number of GMRES iterations are shown at every datapoint (additional data in Table 2). (f) Peak working memory. We use 64 compute nodes, with 4 MPI ranks per node and 8 OpenMP threads per MPI process.

reason for this is that the fill-in entries in the Cholesky factors are so small as to fall into the range of subnormal numbers, which are expensive to compute [37]. A thorough investigation of how the subnormal numbers propagate in the Cholesky factors for the 2D problems can be found in [34].

1182 Here, we set $\delta = 10^{-4}$ and solve Eq. (8) by a 7-point finite difference discretization. We vary N from 64^3 to 512^3 and 1183 compare the performance of four multifrontal solvers, with BLR, BLR_ZFP, HODBF_BLR and HODBF_BLR_ZFP, see 1184 Fig. 9. We set the tolerance $\varepsilon = 10^{-2}$ for both the BLR and the HODBF compression tolerance and use 16 bitplanes 1185 for zfp compression. For the multifrontal solver with BLR or BLR_ZFP compression, all fronts with separator sizes 1186 1187 $n_{\text{Bmin}} \ge 200$ are compressed with BLR. For the multifrontal solver with HODBF_BLR_ZFP or HODBF_BLR compression, 1188 all fronts with separator sizes $n_{\text{Hmin}} \ge 5$ K are compressed with HODBF, all fronts corresponding to separator sizes 1189 $200 \le n_{\text{Bmin}} \le 5$ K are compressed with BLR, and all fronts corresponding to separator sizes below 200 are either not 1190 compressed or compressed with zfp compression. For all experiments, the relative error was $||x - \tilde{x}||_2 / ||\tilde{x}||_2 < 10^{-5}$. We observed that when adding the HODBF compression one can attain the $O(N \log^2 N)$ computation and O(N) memory complexities. Adding zfp on top of BLR and HODBF can slightly improve the total factor memory, but can increase the solve time per iteration. 1195

Manuscript submitted to ACM

1196

1169

1170

1171

1172

1173

1179

1180



Fig. 9. Results for 3D singularly perturbed differential equations (SPDE) using the BLR, BLR-ZFP, BLR-HODBF and BLR-HODBF-ZFP (10^{-2}) multifrontal solver with GMRES. (a) Flop counts for factorization. (b) CPU time for factorization. (c) Memory usage for factorization, i.e. factor nonzeros/factor memory. (d) Flop counts for solve per iteration in GMRES. (e) CPU time for solve. (f) Peak working memory per MPI. We use 64 compute nodes, with 2 MPI ranks per node and 16 OpenMP threads per MPI process.

5.3 Incompressible Navier-Stokes Flow

We solve linear systems with the Stokes operator, modeling incompressible flow described by the Navier-Stokes equations. The system is discretized on a regular 3D mesh using a staggered grid with the velocity components approximated at the cell faces and the pressure at the cell centers. Velocity boundary conditions are used on all (six) faces of the cube. For a 3D cube with *k* mesh points in each direction, there are $3(k + 1)k^2$ velocity degrees of freedom and k^3 pressure degrees of freedom. The discretization is performed using IBAMR [13, 14, 33] (Immersed Boundary Method Adaptive Mesh Refinement Software Infrastructure), which is built on top of SAMRAI [5] (Structured Adaptive Mesh Refinement Application Infrastructure). Matrix assembly is done through PETSc [7].

Discretization of the governing equations leads to a linear system Mx = b, or

$$\begin{array}{c} A & G \\ -D & 0 \end{array} \begin{vmatrix} x_u \\ x_p \end{vmatrix} = \begin{bmatrix} b_u \\ b_p \end{vmatrix}$$

$$(9)$$

where *A* corresponds to the temporal and viscous terms, *G* to the (pressure) gradient and *D* to the diverge (of the velocity). x_u and x_p are the velocity (3 spatial components) and pressure respectively.

¹²⁴⁵ Since the pressure is only defined up to a constant, the matrix *M* is singular. The nullspace is $Z = k^{-3/2} [0 \dots 01 \dots 1]$, ¹²⁴⁷ with the last k^3 elements of *Z* corresponding to the pressure degrees of freedom. We construct an exact solution ¹²⁴⁸ Manuscript submitted to ACM

24

1232

1233

1234

1235 1236

1237

1238

1239 1240

1241 1242 1243

1244



Fig. 10. Results for linear systems modeling incompressible Navier-Stokes flow using the BLR and BLR-ZFP multifrontal solver with BiCGStab. (a) Flop counts for factorization. (b) CPU time for factorization. (c) Memory usage for factorization, i.e. factor nonzeros/factor memory. (d) Total solve flop counts in BiCGStab and the multifrontal preconditioner. (e) CPU time for solve. (f) Peak working memory per MPI rank. We use 64 compute nodes, with 4 MPI ranks per node and 8 OpenMP threads per MPI process.

1276

1277

1278 1279

1280 $\tilde{x} = x_r - ZZ^T x_r$ with x_r a vector with elements in $\mathcal{N}(0, 1)$. From the solution vector x obtained using BiCGStab with 1281 the approximate multifrontal preconditioner, we compute the final solution $x \leftarrow x - ZZ^T x$ and compare that to \tilde{x} . The 1282 BiCGStab stopping criterion is a 10^{-10} relative residual decrease. Results of the BLR ZFP multifrontal solver and the 1283 BLR multifrontal solver are shown in Fig. 10 with k varying from 100 to 250. For BLR, fronts corresponding to separator 1284 1285 sizes n_{Bmin} above 1000 are compressed with a relative compression tolerance 10^{-6} and all fronts corresponding to 1286 separators sizes below 1000 are either not compressed or compressed with zfp compression using 32 bitplanes. M 1287 has a large zero block on the diagonal. Our solver can apply a static permutation, before the start of the numerical 1288 factorization, to make the main diagonal of the matrix nonzero. This permutation is implemented using the MC64 1289 1290 matching code. However, doing so completely destroys the symmetry of the pattern of M. Note that our solver computes 1291 a (non-symmetric) LU factorization, but using a symmetric nonzero pattern. Furthermore, we observed numerical 1292 difficulties when trying to solve the linear system with the matrix M permuted with the MC64 matching. Our solver 1293 also implements a small pivot replacement option, in which, during numerical factorization, small pivots, which 1294 1295 would cause overflow during triangular solution, are replaced with a slightly larger value of $\sqrt{\epsilon_{\text{mach}}} \|M\|_1$. However, 1296 since the operator M is highly ill-conditioned, instead of relying on this small pivot replacement, we replace the zero 1297 diagonal elements of M with $\tau \sqrt{\epsilon_{\text{mach}}} \|M\|_1$ before starting the numerical factorization. Since we also apply a matrix 1298 equilibration similar to LAPACK's dgeequ/dlaqge, $||M||_1 = 1$. The factor τ needs to be chosen carefully. A larger τ 1299 1300 Manuscript submitted to ACM

1301 reduces the condition number of the preconditioner and allows for better compression. However, a larger τ leads to a 1302 worse preconditioner, resulting in more BiCGStab iterations. We pick $\tau = 10^4$. The diagonal shift is also applied for the 1303 multifrontal solver without compression, which is then also used with BiCGStab (instead of iterative refinement). For 1304 all experiments, the relative error was $\|x - \tilde{x}\|_2 / \|\tilde{x}\|_2 < 10^{-5}$. Fig. 10 shows a lower peak memory consumption when 1305 1306 using BLR(Hybrid)_ZFP which allows to solve larger problem sizes. However, BLR(RL) and BLR(RL)_ZFP outperform 1307 BLR(Hybrid)_ZFP in terms of factor and solve time. 1308

5.4 SuiteSparse	Matrix	Col	lection
-----------------	--------	-----	---------

																		0.					
	l r					no compr. BLR(RL,10 ⁻²) compression						BLR(RL,10 ⁻²)_ZFP(32) compression											
		Ν	nnz	fact	sol		#fro	nts	fact	sol		comp	peak	rel.	#:	fronts		fact	sol		comp	peak	rel.
matr	rix	$\times 10^3$	$\times 10^3$	(s)	(s)	its	dense	BLR	(s)	(s)	its	(%)	(%)	err.	dense	ZFP	BLR	(s)	(s)	its	(%)	(%)	err.
bone	eS01	127	6,715	0.45	0.02	1	15K	4	0.38	0.20	14	87.3	101.0	8e-06	11K	4K	4	0.43	4.92	14	46.7	74.0	8e-06
xeno	m2	157	3,866	0.49	0.02	1	19K	8	0.47	0.23	12	87.4	102.0	9e-07	13K	5K	8	0.56	4.49	12	50.2	76.0	9e-07
scirc	uit	170	959	0.12	0.01	1	21K	0	0.12	0.01	1	100.0	102.0	1e-10	12K	8K	0	0.12	0.11	4	53.7	47.0	2e-08
pwtk	c	217	11,634	0.27	0.02	1	27K	1	0.27	0.06	3	99.7	99.0	6e-08	20K	6K	1	0.33	1.06	3	50.8	70.0	1e-06
torse	53	259	4,632	0.81	0.03	1	32K	14	0.59	0.10	2	72.1	100.0	2e-07	25K	7K	14	0.70	1.66	2	37.0	70.0	2e-07
cage	13	445	7,479	75.11	0.25	1	55K	56	28.74	0.53	2	24.1	54.0	6e-07	46K	8K	56	31.69	8.50	2	17.8	50.0	6e-07
audil	kw_1	943	77,651	12.54	0.42	1	117K	124	6.60	4.90	48	47.7	47.0	5e-04	94K	22K	124	7.04	141.04	58	29.2	37.0	1e-03
atmo	osmodd	1,270	8,814	9.82	0.12	1	158K	63	5.60	1.03	8	44.5	64.0	5e-07	106K	52K	63	5.94	15.32	8	29.3	51.0	5e-07
Serei	na	1,391	64,531	42.39	0.36	1	173K	134	18.66	1.87	10	34.8	52.0	2e-05	137K	36K	134	19.01	55.42	10	22.3	39.0	2e-02
Geo	1438	1,437	63,156	26.76	0.23	1	179K	177	12.76	2.01	13	45.4	47.0	8e-05	142K	36K	177	13.39	41.93	12	30.1	36.0	1e-03
atmo	osmodl	1,489	10,319	9.18	0.13	1	186K	64	5.40	0.60	4	45.8	64.0	9e-07	126K	59K	64	5.85	10.36	4	27.4	52.0	9e-07
Hool	k_1498	1,498	60,917	15.00	0.22	1	187K	135	7.81	4.81	34	46.9	48.0	4e-05	137K	49K	135	8.37	139.68	34	27.3	31.0	4e-05
ML_	Geer	1,504	110,879	4.79	0.14	1	187K	115	2.97	2.32	24	65.4	40.0	2e-05	144K	43K	115	3.79	84.81	23	33.5	24.0	2e-05
Tran	sport	1,602	23,500	8.76	0.18	1	200K	128	4.69	3.01	26	51.7	60.0	1e-05	152K	47K	128	5.41	81.27	26	30.5	42.0	7e-06
mem	nchip	2,707	15,950	0.35	0.13	1	338K	0	0.35	0.17	1	100.0	100.0	7e-15	204K	134K	0	0.42	0.68	1	48.3	51.0	2e-06

Table 3. Results for the numerical factorization and solve for a number of matrices from the SuiteSparse matrix collection. Compression 1327 ratio (comp %) refers to the size of the final LU factors relative to the exact solver without compression, while peak (%) refers to 1328 peak memory usage during factorization, also relative to the exact solver. The factorization is always fastest for the solver with BLR 1329 compression. The hybrid BLR_ZFP solver uses the least memory. 1330

1331

1332 1333 1334

1335

1337

1338

1339

1340 1341

The problems shown so far where all defined on a regular three-dimensional domain. Table 3 shows results for a number of matrices from the SuiteSparse matrix collection. These are some of the larger problems in this collection, and they correspond to a range of different applications, with varying numerical and structural properties. Table 3 shows 1336 a comparison between the exact solver (no compression), and the approximate multifrontal solver with either BLR compression (RL variant with relative compression tolerance 10^{-2}), or the BLR_ZFP compression (lossy compression with 32 bitplanes). For all other parameters, and for each problem, the default values are used, e.g., $n_{\text{Bmin}} = 512$, $n_{\text{Lmin}} = 8.$

1342 In Table 3, we notice the speedup obtained for the factorization when using the BLR compression. When using 1343 BLR_ZFP compression, there is a small overhead in factorization time, compared to BLR compression only. Likewise, 1344 when enabling zfp compression, the solve becomes significantly slower. However, enabling zfp leads to better compres-1345 sion ratios. For instance for the scircuit and memchip systems, using BLR only does not give any compression, whereas 1346 1347 zfp compresses the factors by at least 2×. Unlike most of the other problems, the scircuit and memchip matrices are not 1348 derived from PDE discretizations, which explains why their graphs do not have separators larger than $n_{Bmin} = 512$. The 1349 fact that there are no large separators means that the amount of fill-in will be relatively small, and hence the exact 1350 sparse direct solver should be an efficient solver. 1351

1352 Manuscript submitted to ACM

¹³⁵³ 6 CONCLUSION

1354

1367

1368

1369

1370 1371

1372

1382 1383

1384

1385

1386

1387

1388

1389

This paper presents a fast and approximate multifrontal solver for large sparse linear systems. The solver leverages 1355 the Hierarchically Off-Diagonal Butterfly compression, HODBF, a reduced-memory version of the non-hierarchical 1356 1357 Block Low-Rank format, BLR, and lossy compression. Depending on the application as well as problem sizes, we make 1358 use of different combinations of the three compression methods. In general, HODBF is used to compress large frontal 1359 matrices, BLR for medium sized frontal matrices and lossy compression for small frontal matrices. The reduced-memory 1360 version of the BLR format, BLR(Hybrid), leads to a reduction in peak memory consumption which allows to solve 1361 larger problem sizes. The resulting solver can attain the $O(N \log^2 N)$ computation and O(N) memory complexities 1362 1363 when adding the HODBF compression. Some of the presented results for smaller problem sizes do not make use of the 1364 HODBF compression since HODBF is beneficial for really large fronts only. Adding zfp on top of BLR and/or HODBF 1365 can improve the compression ratios and the total factor memory, but can increase the solve time per iteration. 1366

The code is made publicly available through the sparse solver package STRUMPACK 2 . The HODBF implementation is integrated using the dense solver package ButterflyPACK 3 , lossy compression is provided with the software package zfp 4 .

ACKNOWLEDGMENTS

1373 This research was supported in part by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the 1374 U.S. Department of Energy Office of Science and the National Nuclear Security Administration, and in part by the U.S. 1375 Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through 1376 1377 Advanced Computing (SciDAC) program through the FASTMath Institute under Contract No. DE-AC02-05CH11231 1378 at Lawrence Berkeley National Laboratory. This research used resources of the National Energy Research Scientific 1379 Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. 1380 1381 DE-AC02-05CH11231.

REFERENCES

- Sivaram Ambikasaran and Eric Darve. 2013. An O(N log N) fast direct solver for partial hierarchically semi-separable matrices. SIAM J. Sci. Comput 57, 3 (Dec. 2013), 477–501.
- [2] Patrick Amestoy, Cleve Ashcraft, Olivier Boiteau, Alfredo Buttari, Jean-Yves L'Excellent, and Clément Weisbecker. 2015. Improving multifrontal methods by means of block low-rank representations. SIAM J. Sci. Comput. 37, 3 (2015), A1451–A1474.
- [3] Patrick Amestoy, Alfredo Buttari, Nicholas J Higham, Jean-Yves L'Excellent, Théo Mary, and Bastien Vieuble. 2023. Combining sparse approximate factorizations with mixed-precision iterative refinement. ACM Trans. Math. Software 49, 1 (2023), 1–29.
- Patrick R. Amestoy, Alfredo Buttari, Jean-Yves L'Excellent, and Theo Mary. 2019. Performance and Scalability of the Block Low-Rank Multifrontal Factorization on Multicore Architectures. ACM Trans. Math. Softw. 45, 1, Article 2 (Feb. 2019), 26 pages. https://doi.org/10.1145/3242094
- 1392 [5] R Anderson, W Arrighi, N Elliott, B Gunney, and R Hornung. 2013. SAMRAI concepts and software design. Technical Report. LLNL.
- 1393 [6] Ariful Azad, Aydin Buluç, Xiaoye S. Li, Xinliang Wang, and Johannes Langguth. 2020. A Distributed-Memory Algorithm for Computing a Heavy-
- Weight Perfect Matching on Bipartite Graphs. *SIAM Journal on Scientific Computing* 42, 4 (2020), C143–C168. https://doi.org/10.1137/18M1189348
 Satish Balay, Shrirang Abhyankar, Mark Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Alp Dener, Victor Eijkhout, W Gropp, et al. 2018. *PETSc users manual*. Argonne National Laboratory.
- [8] Jack J. Dongarra, Iain S. Duff, Danny C. Sorensen, and Henk A. van der Vorst. 1998. Numerical Linear Algebra for High-Performance Computers. Society for Industrial and Applied Mathematics, Philadelphia, United States. https://doi.org/10.1137/1.9780898719611
 [398] arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9780898719611
- [9] Pieter Ghysels. 2014. STRUMPACK: STRUctured Matrix PACKage. http://portal.nersc.gov/project/sparse/strumpack/.
- 1400

1404

^{1401 &}lt;sup>2</sup>https://github.com/pghysels/STRUMPACK

 ^{1402 &}lt;sup>3</sup>https://github.com/liuyangzhuan/ButterflyPACK

^{1403 &}lt;sup>4</sup>https://github.com/LLNL/zfp

- [10] Pieter Ghysels, Xiaoye Sherry Li, Christopher Gorman, and François-Henry Rouet. 2017. A robust parallel preconditioner for indefinite systems
 using hierarchical matrices and randomized sampling. In 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS). IEEE,
 Orlando, FL, USA, 897–906.
- [11] G.Karypis and V.Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing 20, 1 (1998), 359–392.
- [12] Lars Grasedyck and Wolfgang Hackbusch. 2003. Construction and arithmetics of H-matrices. Computing 70, 4 (2003), 295–334.
- [13] Boyce Eugene Griffith. 2005. Simulating the blood-muscle-valve mechanics of the heart by an adaptive and parallel version of the immersed boundary method. Ph. D. Dissertation. New York University.
- [14] Boyce E Griffith, Richard D Hornung, David M McQueen, and Charles S Peskin. 2007. An adaptive, formally second order accurate version of the immersed boundary method. *Journal of computational physics* 223, 1 (2007), 10–49.
- [11] Wolfgang Hackbusch. 1999. A sparse matrix arithmetic based on *H*-matrices. Part I: Introduction to *H*-matrices. Computing 62, 2 (April 1999),
 [15] 89–108.
- [16] Wolfgang Hackbusch, Boris N Khoromskij, and Ronald Kriemann. 2004. Hierarchical matrices based on a weak admissibility criterion. Computing
 73, 3 (2004), 207–243.
- 1418[17]Pascal Hénon, Pierre Ramet, and Jean Roman. 2002. PaStiX: a High-Performance Parallel Direct Solver for Sparse Symmetric Positive Definite1419Systems. Parallel Comput. 28, 2 (2002), 301–321.
- [18] Nicholas J Higham and Theo Mary. 2022. Solving block low-rank linear systems by LU factorization is numerically stable. *IMA J. Numer. Anal.* 42, 2 (2022), 951–980.
- [19] I.S.Duff and J.Koster. 1999. The design and use of algorithms for permuting large entries to the diagonal of sparse matrices. *SIAM J MATRIX ANAL A.* 20, 4 (1999), 889–901.
- [20] I.S.Duff and J.K.Reid. 1983. The multifrontal solution of indefinite sparse symmetric linear. ACM Trans. Math. Softw. 9, 3 (1983), 302–325
- [21] Yingzhou Li and Haizhao Yang. 2017. Interpolative butterfly factorization. SIAM J. Sci. Comput. 39, 2 (2017), A503–A531. https://doi.org/10.1137/
 16M1074941
- [22] Yingzhou Li, Haizhao Yang, Eileen R Martin, Kenneth L Ho, and Lexing Ying. 2015. Butterfly factorization. *Multiscale Model. Sim.* 13, 2 (2015),
 714–732.
- 1428
 [23] Peter Lindstrom. 2014. Fixed-Rate Compressed Floating-Point Arrays. IEEE Transactions on Visualization and Computer Graphics 20 (08 2014).

 1429
 https://doi.org/10.1109/TVCG.2014.2346458
- [24] Peter Lindstrom. 2014. zfp: Compressed Floating-Point and Integer Arrays. https://computing.llnl.gov/projects/zfp.
- [25] J. W. H. Liu. 1992. The multifrontal method for sparse matrix solution: Theory and practice. *SIAM Rev.* 34 (1992), 82–109. https://doi.org/10.1137/ 1034004
- [26] Yang Liu. 2018. ButterflyPACK. https://portal.nersc.gov/project/sparse/butterflypack/.
- [433 [27] Yang Liu, Pieter Ghysels, Lisa Claus, and Xiaoye Sherry Li. 2021. Sparse Approximate Multifrontal Factorization with Butterfly Compression for
 ¹⁴³⁴ High-Frequency Wave Equations. *SIAM Journal on Scientific Computing* 43, 5 (2021), S367–S391. https://doi.org/10.1137/20M1349667
- [435 [28] Yang Liu, Han Guo, and Eric Michielssen. 2017. An HSS matrix-inspired butterfly-based direct solver for analyzing scattering from two-dimensional
 biperts. *IEEE Antennas Wirel. Propag. Lett.* 16 (2017), 1179–1183.
- [29] Yang Liu, Xin Xing, Han Guo, Eric Michielssen, Pieter Ghysels, and Xiaoye Sherry Li. 2021. Butterfly Factorization Via Randomized Matrix-Vector
 Multiplications. *SIAM Journal on Scientific Computing* 43, 2 (2021), A883–A907. https://doi.org/10.1137/20M1315853
- 1439[30]S. MacLachlan and N. Madden. 2013. Robust Solution of Singularly Perturbed Problems Using Multigrid Methods. SIAM J. Sci. Comput. 35, 5 (2013),1440A2225-A2254. https://doi.org/10.1137/120889770
- [31] Théo Mary. 2017. Block Low-Rank multifrontal solvers: complexity, performance, and scalability. Ph. D. Dissertation. l'Université de Toulouse.
- [32] Eric Michielssen and Amir Boag. 1994. Multilevel evaluation of electromagnetic fields for the rapid solution of scattering problems. *Microw Opt Technol Lett.* 7, 17 (1994), 790–795.
- [33] Nishant Nangia, Boyce E Griffith, Neelesh A Patankar, and Amneet Pal Singh Bhalla. 2019. A robust incompressible Navier-Stokes solver for high density ratio multiphase flows. *J. Comput. Phys.* 390 (2019), 548–594.
- [34] T. A. Nhan and N. Madden. 2015. Cholesky Factorisation of Linear Systems Coming from Finite Difference Approximations of Singularly Perturbed
 Problems. In Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2014, Petr Knobloch (Ed.). Springer International Publishing,
 Cham, 209–220.
- 1448[35] Richard Nies and Matthias Hoelzl. 2019. Testing performance with and without block low rank compression in MUMPS and the new PaStiX 6.0 for1449JOREK nonlinear MHD simulations. arXiv e-prints (2019), arXiv:1907.13442.
- [145] [36] Stéphane Operto, Jean Virieux, Patrick Amestoy, Jean-Yves L'Excellent, Luc Giraud, and Hafedh Ben Hadj Ali. 2007. 3D finite-difference frequencydomain modeling of visco-acoustic wave propagation using a massively parallel direct solver: A feasibility study. *Geophysics* 72, 5 (2007), SM195–SM211.
- [37] Michael L. Overton. 2001. Numerical Computing with IEEE Floating Point Arithmetic. Society for Industrial and Applied Mathematics, Philadelphia,
 [453 United States. https://doi.org/10.1137/1.9780898718072 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9780898718072
- [454 [38] Qiyuan Pang, Kenneth L. Ho, and Haizhao Yang. 2020. Interpolative decomposition butterfly factorization. SIAM J. Sci. Comput. 42, 2 (2020),
 [455 A1097–A1115. https://doi.org/10.1137/19M1294873
- 1456 Manuscript submitted to ACM

Sparse approximate multifrontal factorization with composite compression methods

- [39] François Pellegrini and Jean Roman. 1997. Sparse matrix ordering with Scotch. In High-Performance Computing and Networking, Bob Hertzberger and Peter Sloot (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 370-378. https://doi.org/10.1007/BFb0031609
- [40] H.-G. Roos, M. Stynes, and L. Tobiska. 2008. Robust numerical methods for singularly perturbed differential equations (second ed.). Springer Series in Computational Mathematics, Vol. 24. Springer-Verlag, Berlin. xiv+604 pages.
- [41] Sadeed Bin Sayed, Yang Liu, Luis J. Gomez, and Abdulkadir C. Yucel. 2022. A Butterfly-Accelerated Volume Integral Equation Solver for Broad Permittivity and Large-Scale Electromagnetic Analysis. IEEE Transactions on Antennas and Propagation 70, 5 (2022), 3549-3559. https://doi.org/10. 1109/TAP.2021.3137193
- [42] John Shaeffer. 2008. Direct Solve of Electrically Large Integral Equations for Problem Sizes to 1 M Unknowns. IEEE Trans. Antennas Propag. 56, 8 (2008), 2306-2313.
- [43] Raf Vandebril, Marc Van Barel, Gene Golub, and Nicola Mastronardi. 2005. A bibliography on semiseparable matrices. Calcolo 42, 3-4 (2005), 249 - 270