

Restructuring the Symmetric QR Algorithm for Performance



Field Van Zee

Gregorio Quintana-Orti

Robert van de Geijn



For details:

Field Van Zee, Robert van de Geijn, and Gregorio Quintana-Orti. “Restructuring the QR algorithm for Performance.” *ACM TOMS*. Accepted (pending minor modifications)

This work was supported by

- UTAustin-Portugal Colab program
- Microsoft
- NSF under grants OCI-0850750, CCF-0917167, and OCI-1148125

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).



Overview

- 50+ years of progress
- The hidden costs of MRRR and D&C
- QR algorithm basics
- Accumulating and applying rotations
- Performance
- Conclusion



Overview

- 50+ years of progress
- The hidden costs of MRRR and D&C
- QR algorithm basics
- Accumulating and applying rotations
- Performance
- Conclusion



Symmetric EVD/SVD: 50+ Years of Progress

- Recent progress focuses a lot on the mathematics side
 - Divide & Conquer (Cuppen's) algorithm (D&C)
 - Method of Relatively Robust Representations (MRRR)
- Occasional revisit of Jacobi's method
- Progress on QR has been for non-symmetric problem.
 - Aggressive Early Deflation
 - Multishift



Two Insights

- **WHEN COMPUTING THE DENSE EVD (all eigenvalues and vectors), D&C and MRRR have hidden $O(n^3)$ cost**
- QR becomes competitive if rotations are applied in batches
 - Classical QR: cast in terms of vector-vector operations
 - Batched application: cast in terms of computation that reuses data in cache, like matrix-matrix operations.



Overview

- 50+ years of progress
- **The hidden costs of MRRR and D&C**
- QR algorithm basics
- Accumulating and applying rotations
- Performance
- Conclusion



The Hidden Cost of D&C and MRRR

- Start with symmetric, dense A
- Reduce to tridiagonal form:

$$A \longrightarrow Q_A T Q_A^T$$

- Compute Spectral Decomposition of T :

$$T \longrightarrow Q_T D Q_T^T$$

- Backtransform:

$$A = \underbrace{Q_A Q_T}_Q D \underbrace{(Q_A Q_T)^T}_{Q^T} = Q D Q^T$$



Reduction to Tridiagonal Form

$$A \longrightarrow Q_A T Q_A^T$$

$$\begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}$$



Reduction to Tridiagonal Form

$$A \longrightarrow Q_A T Q_A^T$$

$$H_{2:5} \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix} H_{2:5}$$

=

$$\begin{pmatrix} \times & \times & & & \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}$$



Reduction to Tridiagonal Form

$$A \longrightarrow Q_A T Q_A^T$$

$$H_{3:5} H_{2:5} \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix} H_{2:5} H_{3:5}$$

=

$$\begin{pmatrix} \times & \times & & & \\ \times & \times & \times & & \\ & & & & \\ \times & & & \times & \times \\ & & & \times & \times \\ & & & \times & \times \\ & & & \times & \times \end{pmatrix}$$



Reduction to Tridiagonal Form

$$A \longrightarrow Q_A T Q_A^T$$

$$H_{4:5} H_{3:5} H_{2:5} \left(\begin{array}{ccccc} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right) H_{2:5} H_{3:5} H_{4:5} =$$
$$\left(\begin{array}{ccccc} \times & \times & & & \\ \times & \times & \times & & \\ \times & \times & \times & \times & \\ \boxed{\begin{array}{c|ccc} & & & \\ \times & \times & \times & \\ \hline \times & & \times & \\ \times & & \times & \end{array}} & & & & \end{array} \right)$$



Backtransformation

$$\begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}$$

Backtransformation

$$H_{4:5} \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}$$

=

$$\begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \boxed{\times & \times & \times & \times & \times} \\ \boxed{\times & \times & \times & \times & \times} \end{pmatrix}$$

Backtransformation

$$H_{3:5} H_{4:5} \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}$$

=

$$\begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \boxed{\times & \times & \times & \times & \times} \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}$$

Backtransformation

$$H_{2:5} H_{3:5} H_{4:5} \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}$$

=

$$\begin{pmatrix} \times & \times & \times & \times & \times \\ \boxed{\times & \times & \times & \times & \times} \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix}$$



Cost of QR algorithm

- Start with symmetric, dense A
- Reduce to tridiagonal form:

$$A \longrightarrow Q_A T Q_A^T$$

- Form Q_A
- Compute Spectral Decomposition of T while updating Q_A

$$Q_A T Q_A^T \longrightarrow (Q_A Q_T) D (Q_A Q_T)^T$$



Form Q_A

$$\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$$



Form Q_A

Form Q_A

$$H_{3:5} H_{4:5} \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{pmatrix} = \begin{pmatrix} & & \\ & & \\ & \boxed{\begin{matrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{matrix}} & \\ & & \end{pmatrix}$$



Form Q_A

$$H_{2:5} H_{3:5} H_{4:5} \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \\ & & & & 1 \end{pmatrix} = \begin{pmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{pmatrix}$$



Form Q_A

$$H_{2:5} H_{3:5} H_{4:5} \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \\ & & & & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & \\ \times & \times & \times & \times \end{pmatrix}$$



Cost

- Backtransformation: $2 n^3$ flops
- Form Q_A : $\frac{4}{3} n^3$ flops
- Hidden cost of MRRII and D&C:
 $2/3 n^3$ flops

EVD OF A DENSE MATRIX!!!
(All eigenvalues and eigenvectors)



Overview

- 50+ years of progress
- The hidden costs of MRRR and D&C
- **QR algorithm basics**
- Accumulating and applying rotations
- Performance
- Conclusion



Classical QR algorithm

Q_A

T

$$\left(\begin{array}{cccc} \times & \times & \times & \times \\ \times & \times & \times & \times \end{array} \right) \quad \left(\begin{array}{ccccc} \times & \times & & & \\ \times & \times & \times & & \\ & \times & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \end{array} \right)$$

$$\left(\begin{array}{cccccc} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right) \quad \left(\begin{array}{ccccc} \times & \times \\ \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right)$$

$$\left(\begin{array}{ccccc} G & & & & \\ \swarrow & \searrow & & & \\ \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{array}} & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right)$$

$$G^T \left(\begin{array}{cc|c} G & & \\ \swarrow & \searrow & \\ \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \hline \times & \end{array}} & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \hline \times & & \times & \times & \times \\ & & & \times & \times \\ & & & & \times \end{array} \right)$$

$$\left(\begin{array}{cc|cc} \times & & & \\ \times & \times & \times & \times \end{array} \right)$$

$$G^T \left(\begin{array}{cc|cc|c} \times & \times & \times & \times & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \\ \times & & \times & \times & \\ \hline \times & \times & \times & \times & \\ \times & \times & \times & \times & \\ \times & & \times & \times & \\ \hline & & & & \end{array} \right)$$

$$\left(\begin{array}{ccccc} \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \end{array}} & \times & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \end{array} \right)$$

A 5x5 matrix with orange boxes highlighting specific submatrices. The top-left 2x2 block is highlighted, and an arrow labeled G points to it from above. The entire matrix is enclosed in large black parentheses.

$$\left(\begin{array}{ccccc} \times & \times & \boxed{\begin{array}{ccccc} \times & \times & \times & \times \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \end{array}} & \times & \\ G^T & \swarrow & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \end{array}} & \times & \\ & \boxed{\begin{array}{c} \times \\ \times \end{array}} & & & \end{array} \right)$$

A 5x5 matrix with orange boxes highlighting specific submatrices. The top-right 3x3 block is highlighted, and an arrow labeled G points to it from above. The bottom-right 2x2 block is highlighted, and an arrow labeled G^T points to it from the left. The entire matrix is enclosed in large black parentheses.

$$\left(\begin{array}{ccccc} \times & \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{array}} &) \\ \times & \times & \times & & \end{array} \right)$$

$$\left(\begin{array}{ccccc} \times & \times & & & \\ \times & \times & \times & & \\ G^T & \swarrow & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \end{array}} & \searrow & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \end{array}} \\ & & & & \end{array} \right)$$

$$\left(\begin{array}{ccccc} G & & & & \\ \swarrow & \searrow & & & \\ \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{array}} & \begin{array}{ccccc} \times & \times & \times \\ \times & \times & \times \end{array} & & & \end{array} \right)$$

$$G^T \left(\begin{array}{cc|c} G & & \\ \swarrow & \searrow & \\ \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \hline \end{array}} & \begin{array}{ccccc} \times & \times & \times \\ \times & \times & \times \\ \hline \times & & \times & \times & \times \\ & & & \times & \times \\ & & & \times & \times \end{array} & \end{array} \right)$$

$$\left(\begin{array}{cc|cc} \times & & & \\ \times & \times & \times & \times \end{array} \right)$$

$$G^T \left(\begin{array}{cc|cc|c} \times & \times & \times & \times & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \\ \times & & \times & \times & \\ \hline \times & \times & \times & \times & \\ \times & \times & \times & \times & \\ \times & & \times & \times & \\ \hline & & & & \end{array} \right)$$

$$\left(\begin{array}{ccccc} \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \end{array}} & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right)$$

$$\left(\begin{array}{ccccc} \times & \times & \boxed{\begin{array}{cccc} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{array}} & \times & \times \\ G^T & \swarrow & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \end{array}} & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \end{array}} & \times \\ \end{array} \right)$$

$$\left(\begin{array}{ccccc} \times & \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{array}} &) \\ \times & \times & \times & & \end{array} \right)$$

$$\left(\begin{array}{ccccc} \times & \times & & & \\ \times & \times & \times & & \\ G^T & \swarrow & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \end{array}} & \searrow & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \end{array}} \\ & & & & \end{array} \right)$$

$$\left(\begin{array}{ccccc} G & & & & \\ \swarrow & \searrow & & & \\ \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{array}} & \begin{array}{ccccc} \times & \times & \times \\ \times & \times & \times \end{array} & & & \end{array} \right)$$

$$G^T \left(\begin{array}{cc|c} G & & \\ \swarrow & \searrow & \\ \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \hline \end{array}} & \begin{array}{ccccc} \times & \times & \times \\ \times & \times & \times \\ \hline \times & & \times & \times & \times \\ & & & \times & \times \\ & & & \times & \times \end{array} & \end{array} \right)$$

$$\left(\begin{array}{cc|cc} \times & \times & \times & \times \\ \times & \times & \times & \times \end{array} \right)$$

A 6x4 matrix where the first two columns are highlighted by a thick orange border. A node labeled G is positioned above the first two columns, with arrows pointing from G to the first and second columns.

$$G^T \left(\begin{array}{cc|cc|c} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \hline \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right)$$

The transpose of the matrix from the left. The first two columns are highlighted by a thick orange border. A node labeled G is positioned above the first two columns, with arrows pointing from G to the first and second columns. Additionally, a node labeled G^T is positioned to the left of the matrix, with arrows pointing from G^T to the first and second columns.

$$\left(\begin{array}{ccccc} \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \end{array}} & \times & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \end{array} \right)$$

A 5x5 matrix with orange boxes highlighting a 2x2 submatrix in the third row and third column. A node labeled G is connected by arrows to the top-left cell of this submatrix.

$$\left(\begin{array}{ccccc} \times & \times & \boxed{\begin{array}{ccccc} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array}} & & \\ G^T & \rightarrow & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \end{array}} & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \end{array}} & \times \\ & \rightarrow & \boxed{\begin{array}{c} \times \\ \times \end{array}} & \boxed{\begin{array}{c} \times \\ \times \end{array}} & \times \end{array} \right)$$

A 5x5 matrix with orange boxes highlighting a 3x3 submatrix in the second row and second column, and a 2x2 submatrix in the fourth row and fourth column. A node labeled G is connected by arrows to the top-left cell of the 3x3 submatrix. A node labeled G^T is connected by arrows to the top-left cell of the 2x2 submatrix.

$$\left(\begin{array}{ccccc}
 \times & \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{array}} & \\
 \times & \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{array}} & \\
 \times & \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{array}} & \\
 \times & \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{array}} & \\
 \times & \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \end{array}} & \\
 \end{array} \right) \quad \left(\begin{array}{ccccc}
 \times & \times & & & \\
 \times & \times & \times & & \\
 \times & \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \end{array}} & \\
 \times & \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \end{array}} & \\
 \times & \times & \times & \boxed{\begin{array}{cc} \times & \times \\ \times & \times \\ \times & \times \end{array}} & \\
 \end{array} \right)$$

G

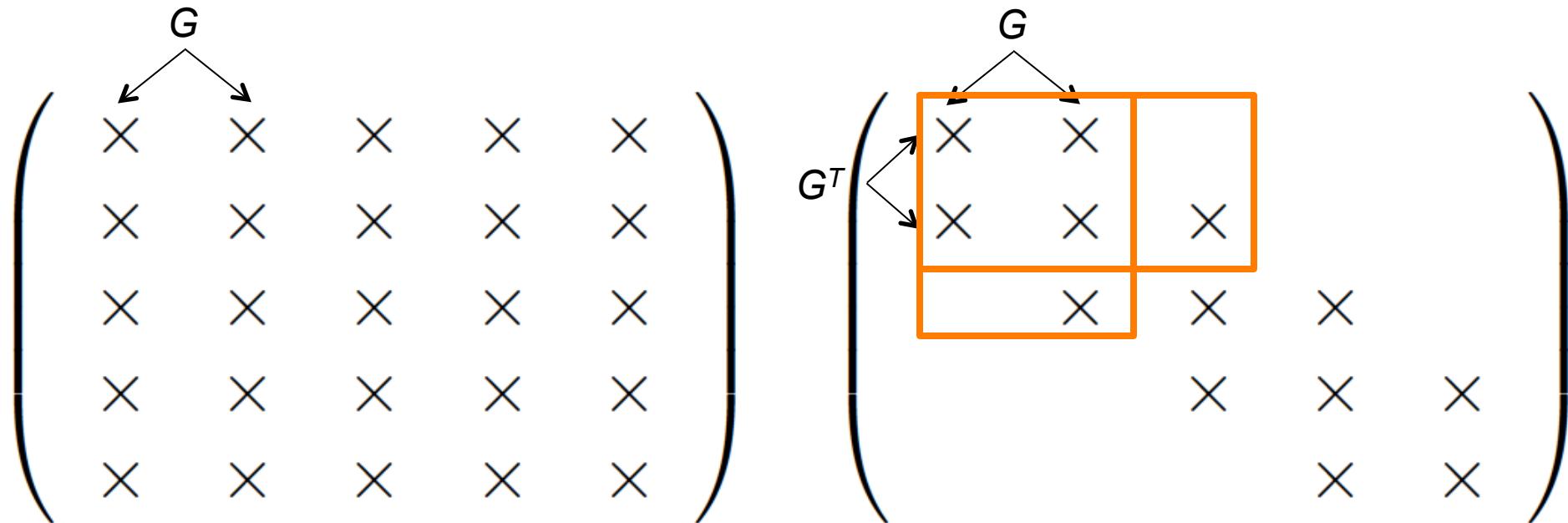
G^T



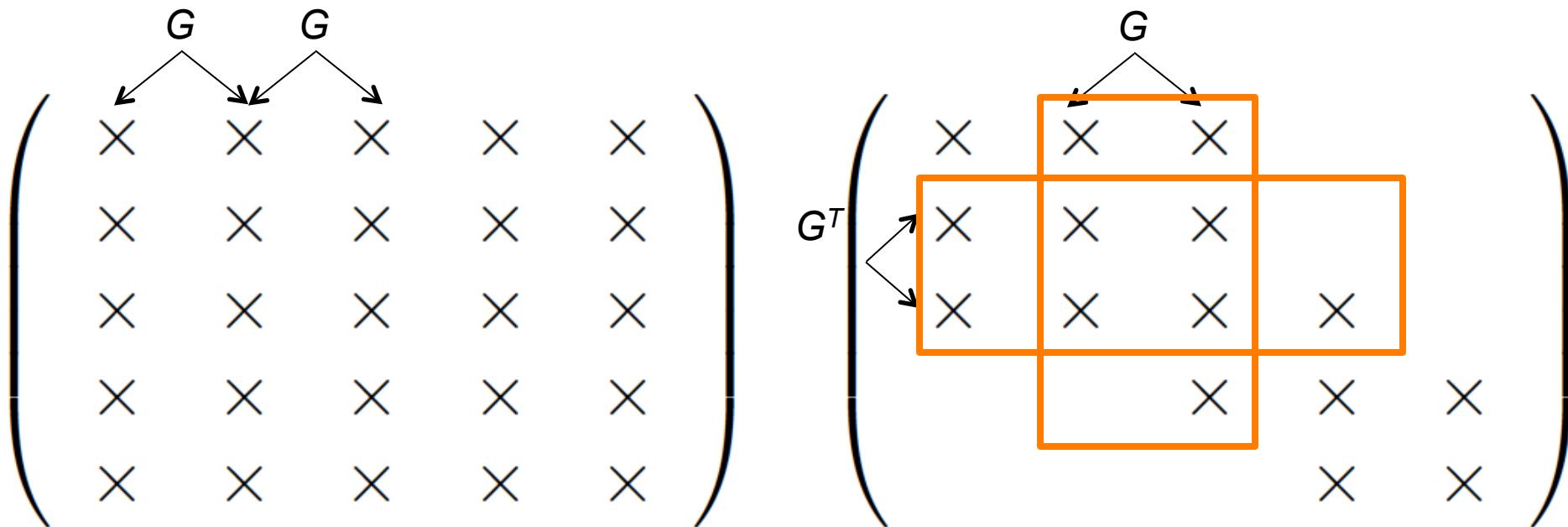
Overview

- 50+ years of progress
- The hidden costs of MRRR and D&C
- QR algorithm basics
- **Accumulating and applying rotations**
- Performance
- Conclusion

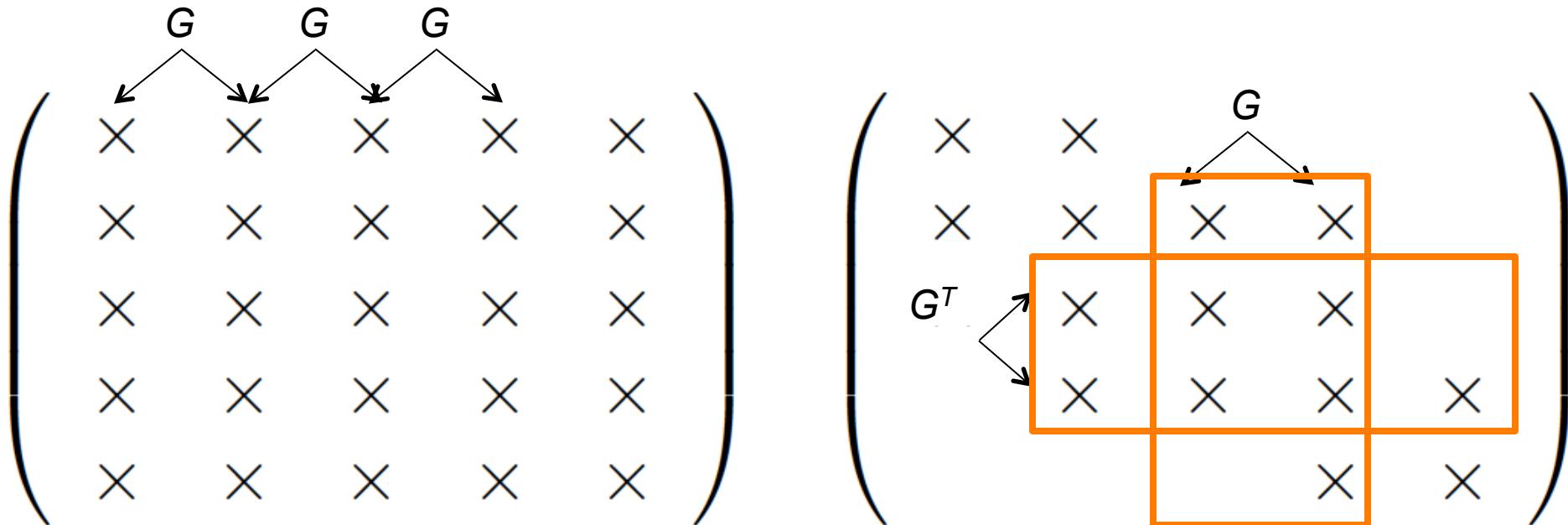
Accumulating Rotations (LAPACK)



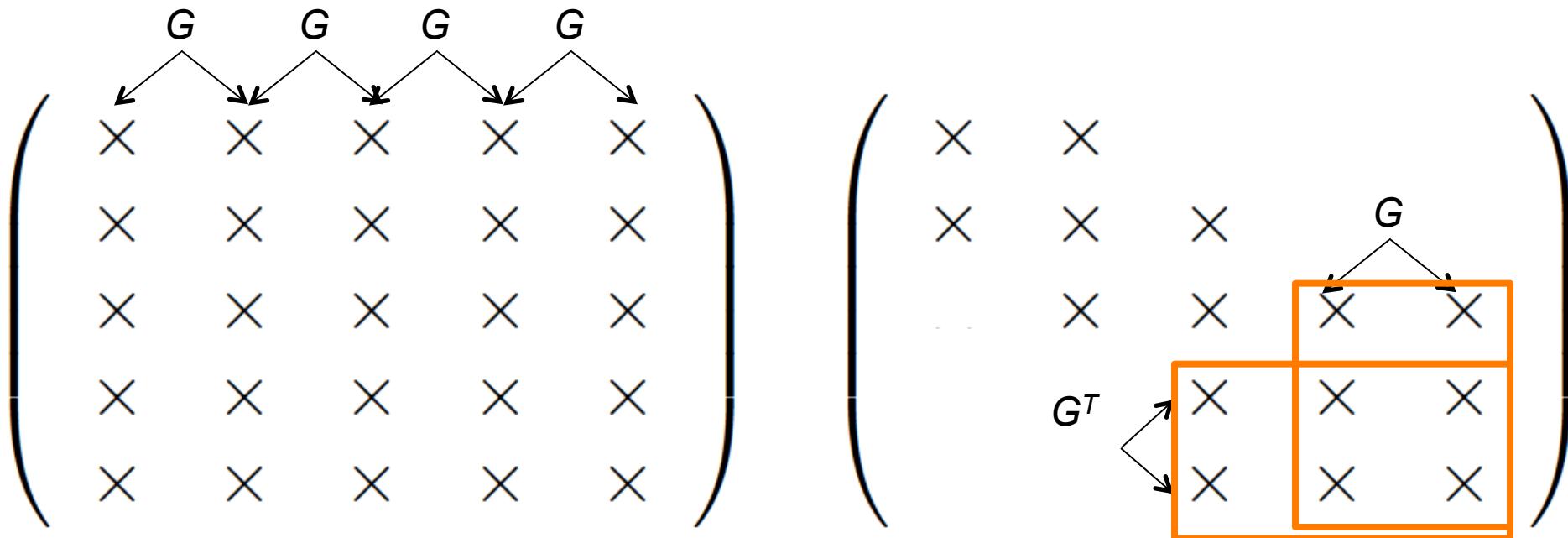
Accumulating Rotations (LAPACK)



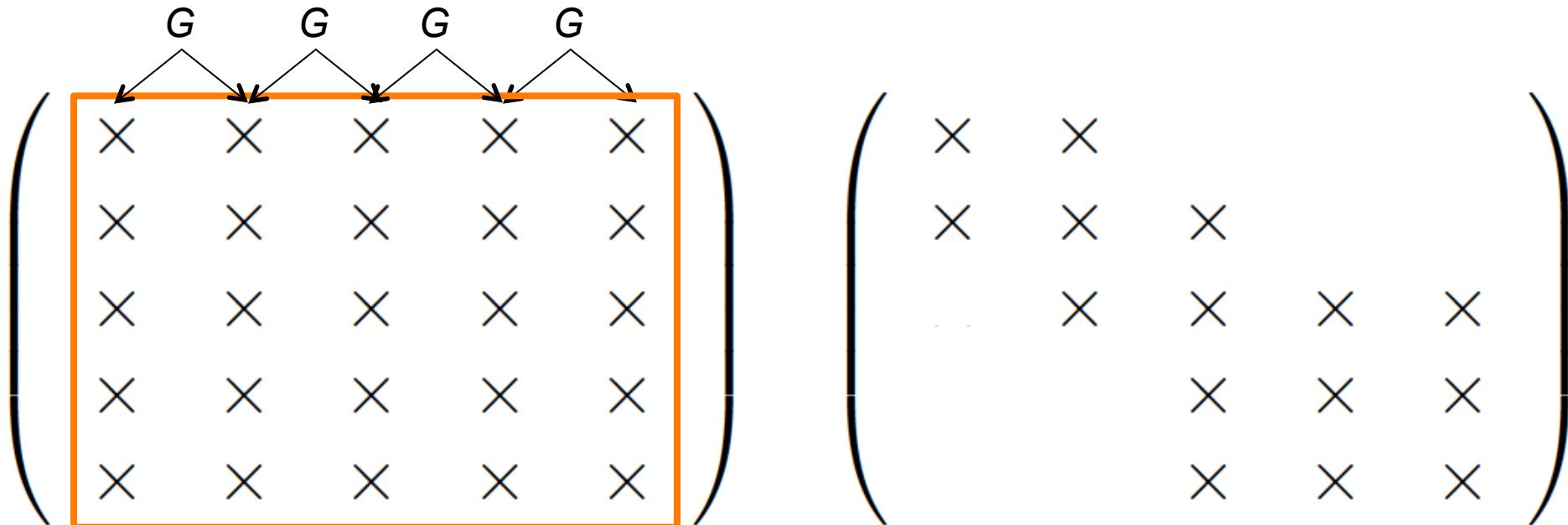
Accumulating Rotations (LAPACK)



Accumulating Rotations (LAPACK)

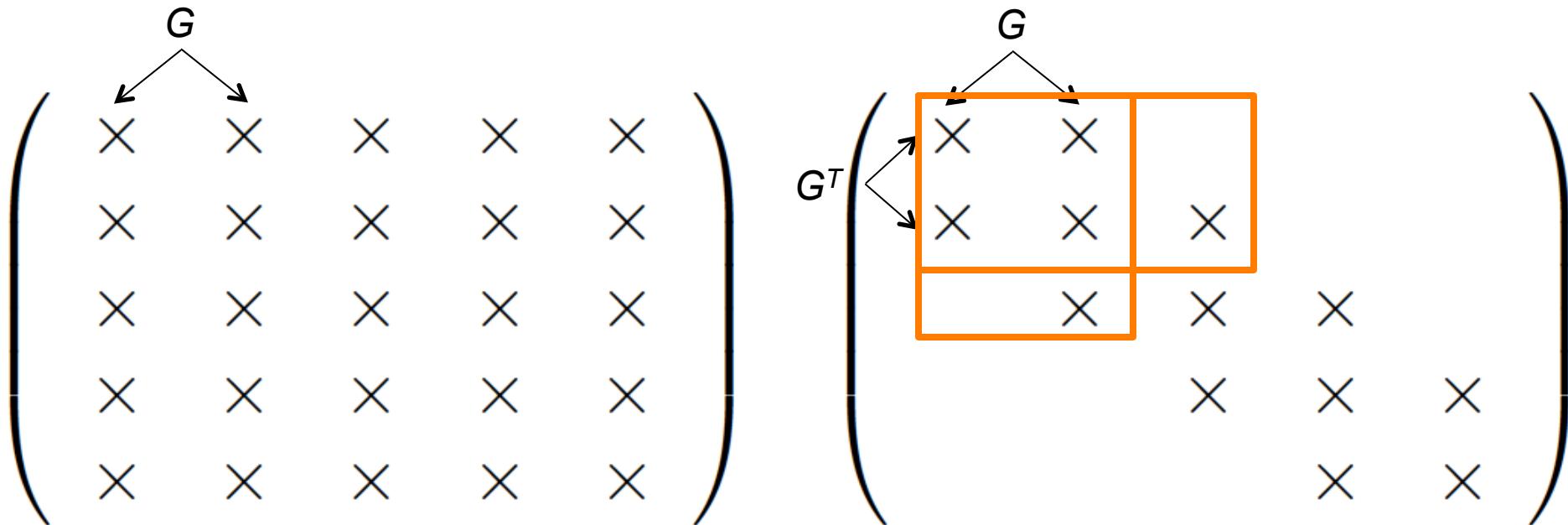


Accumulating Rotations (LAPACK)

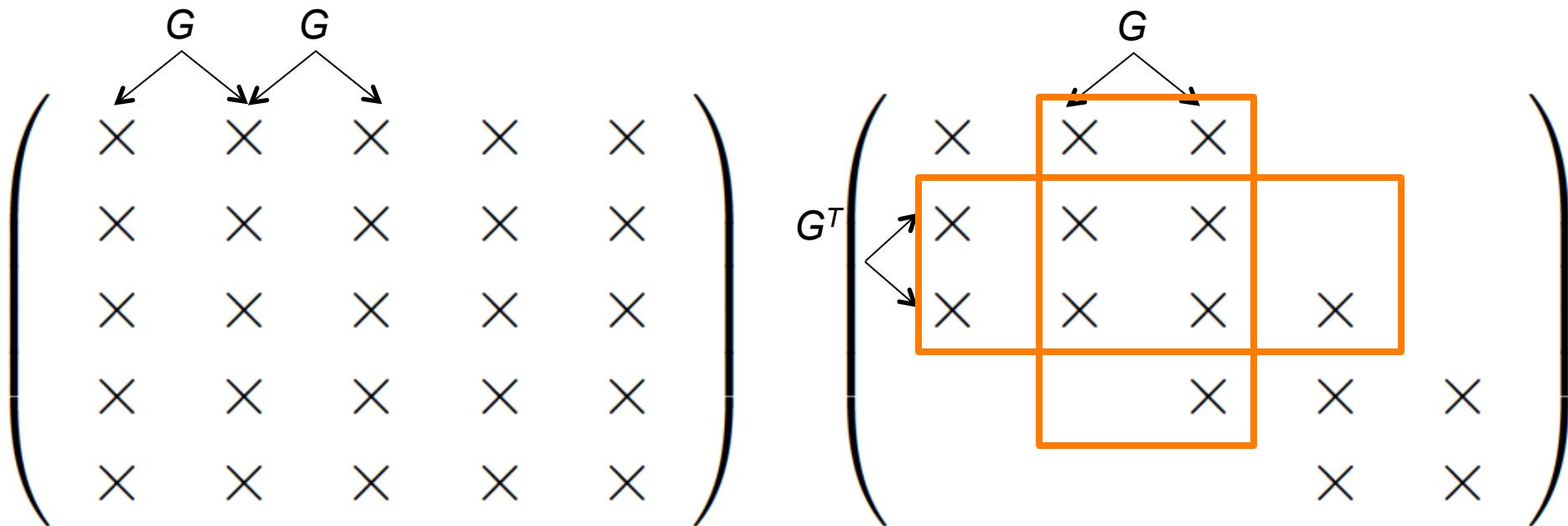


- Apply one sweep worth of rotations.
- Makes application like “level-2 BLAS”

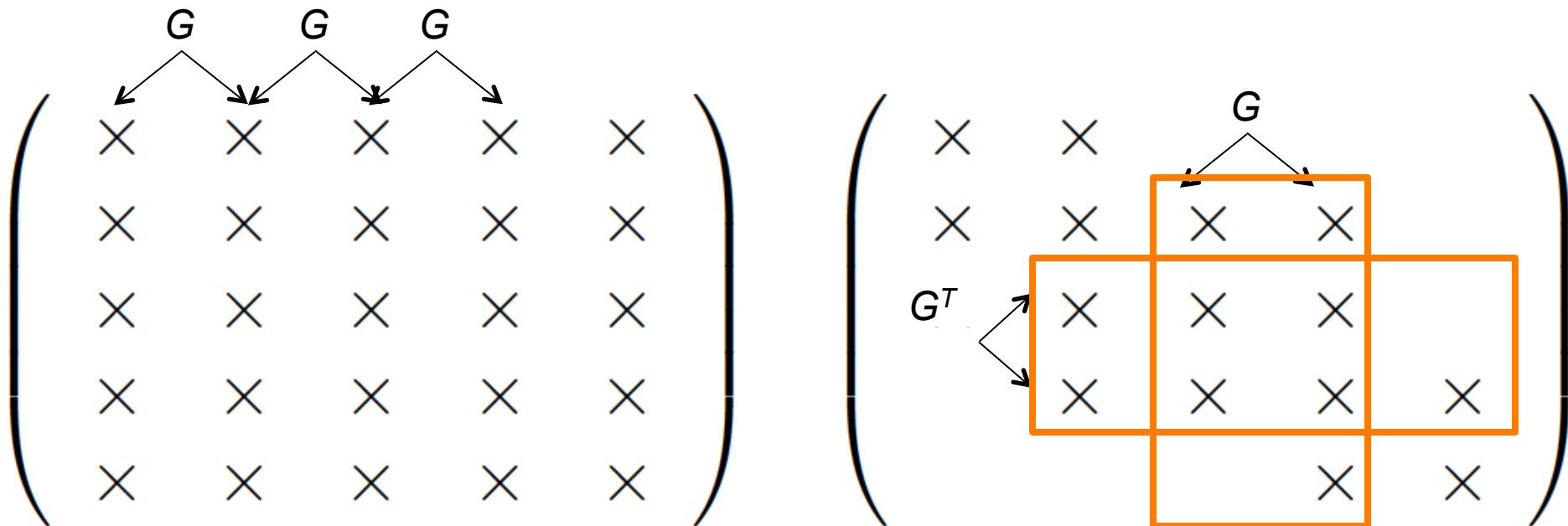
Accumulating Rotations (libflame)



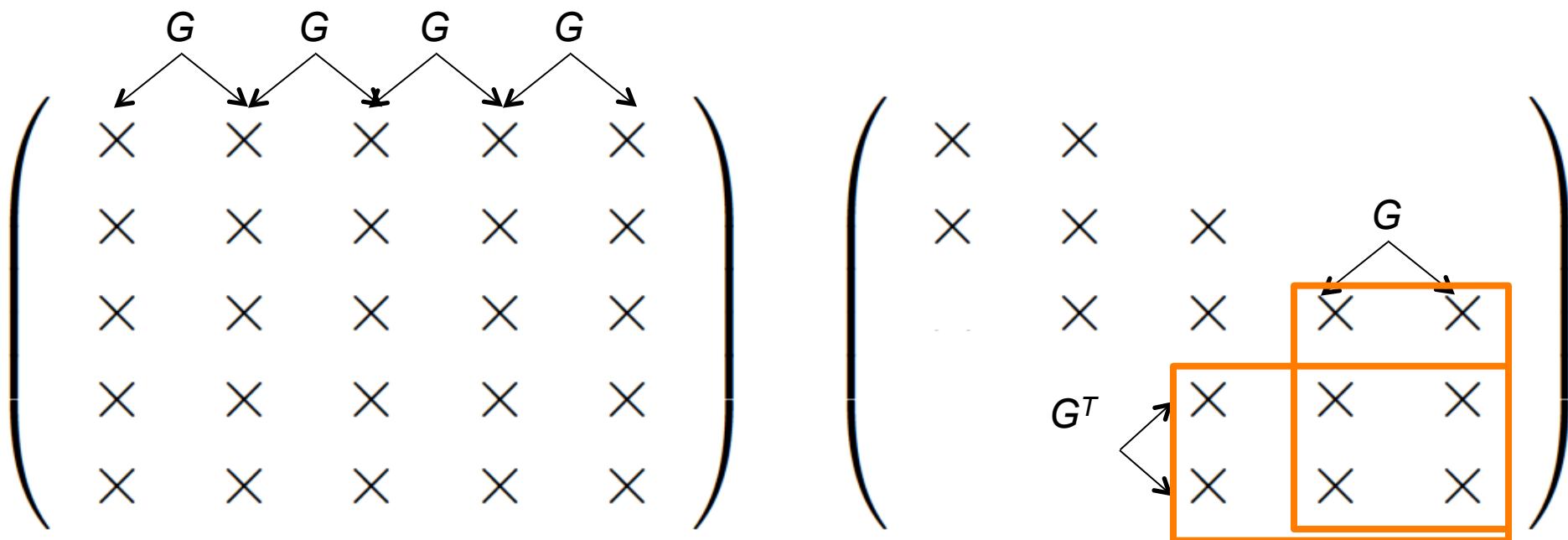
Accumulating Rotations (libflame)



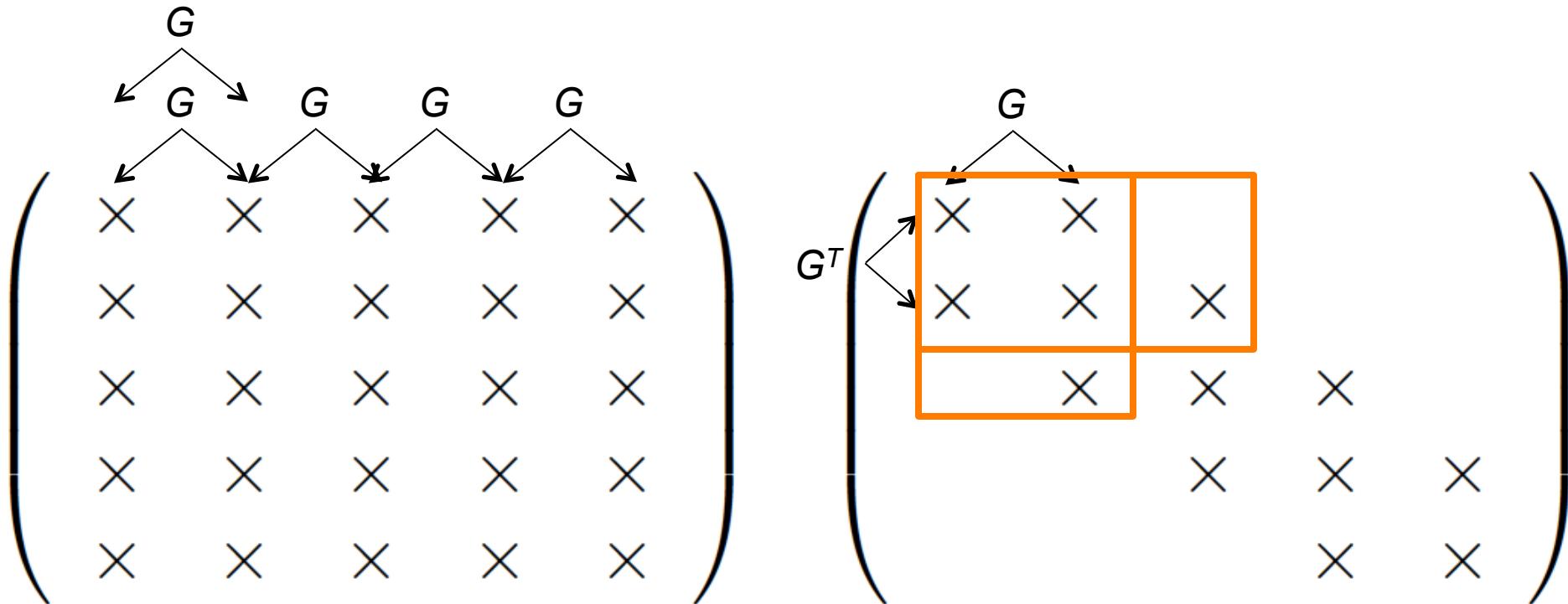
Accumulating Rotations (libflame)



Accumulating Rotations (libflame)

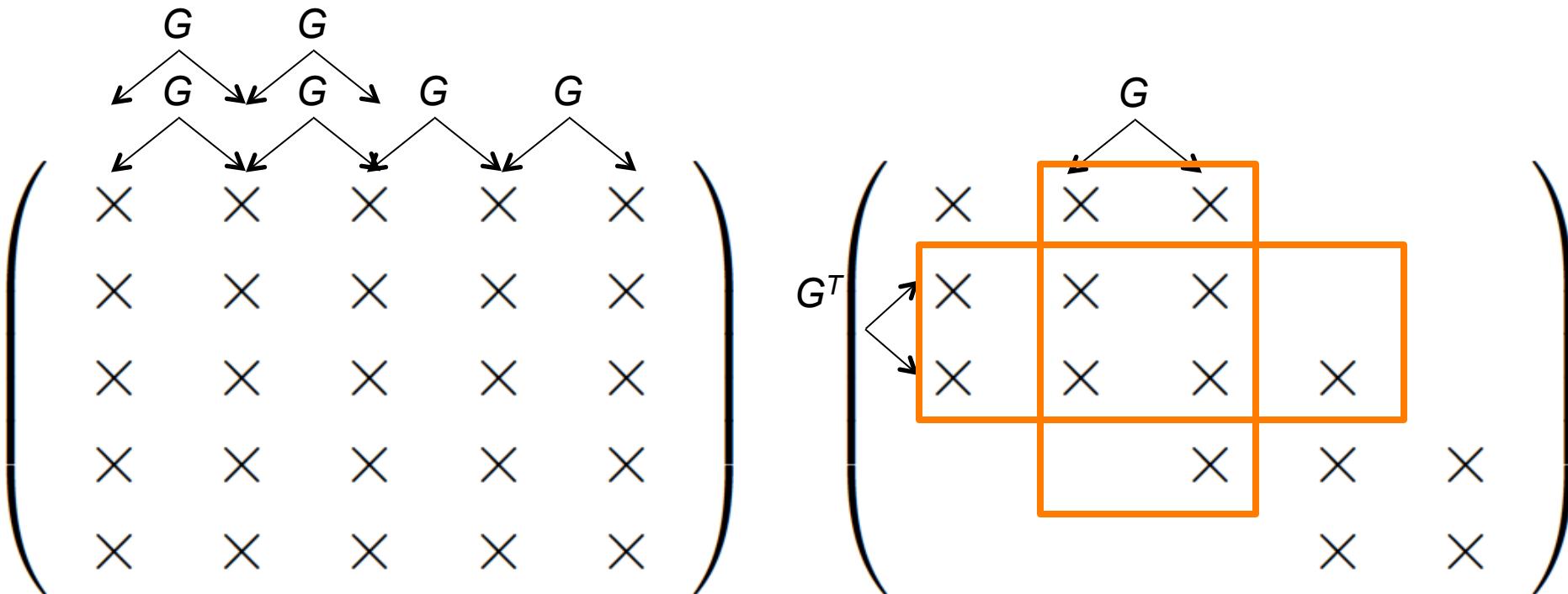


Accumulating Rotations (libflame)

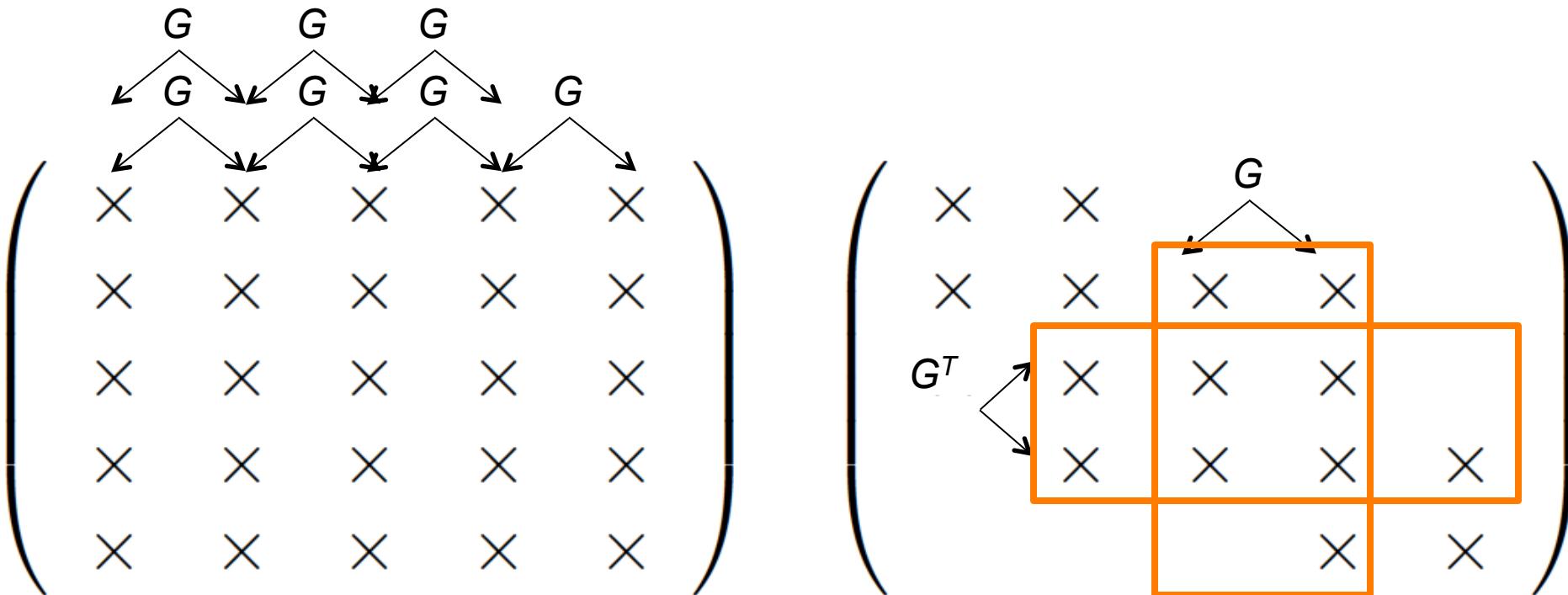




Accumulating Rotations (libflame)

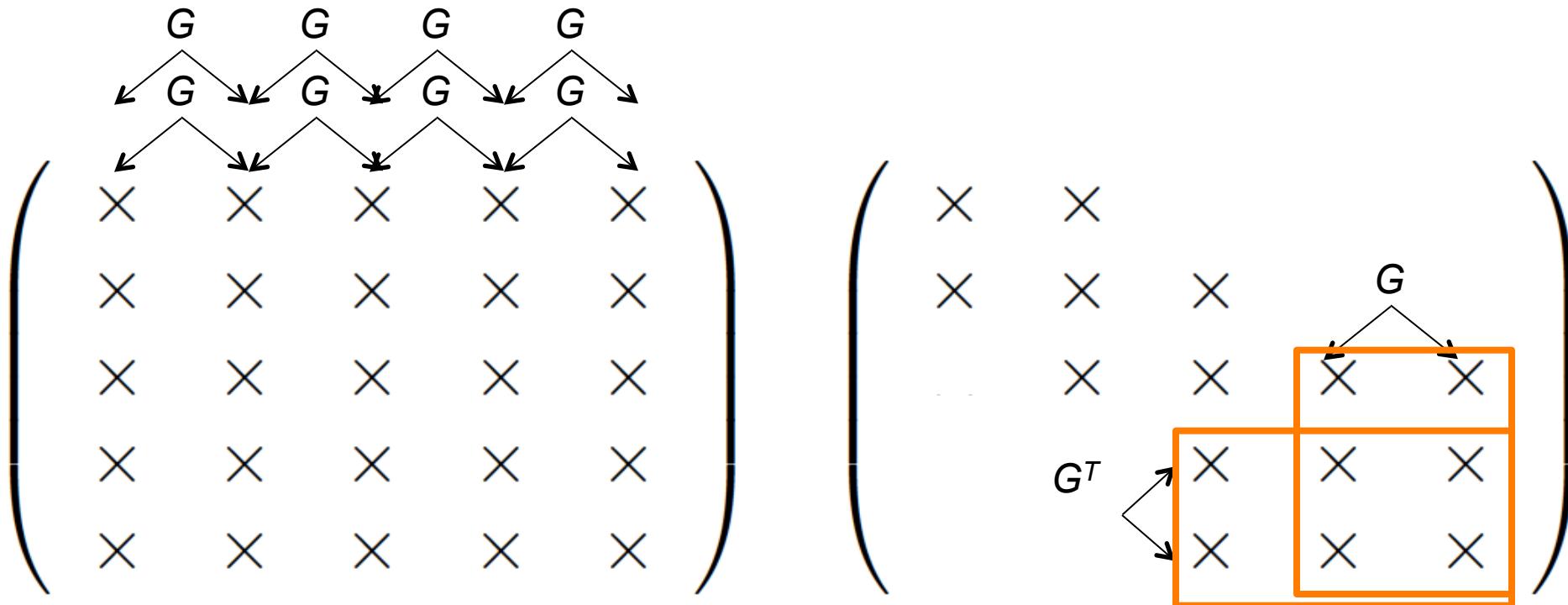


Accumulating Rotations (libflame)

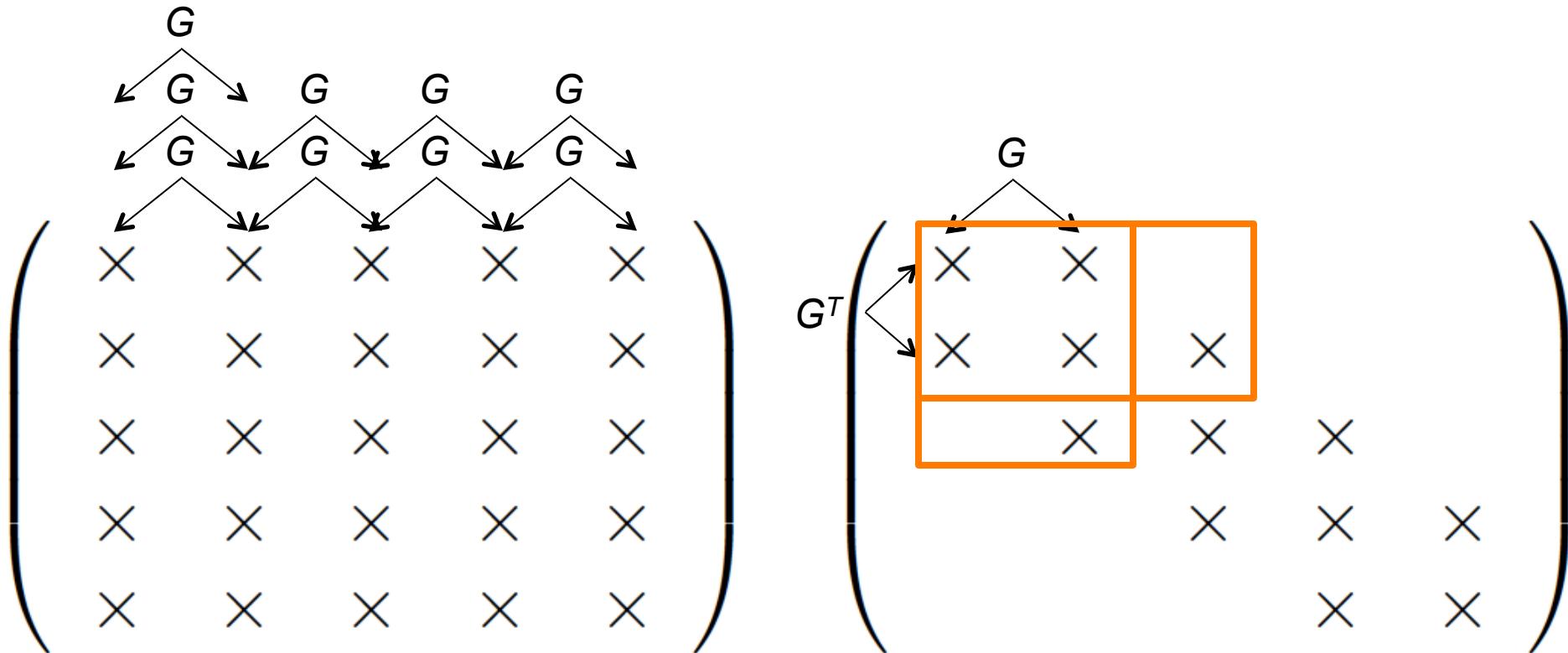




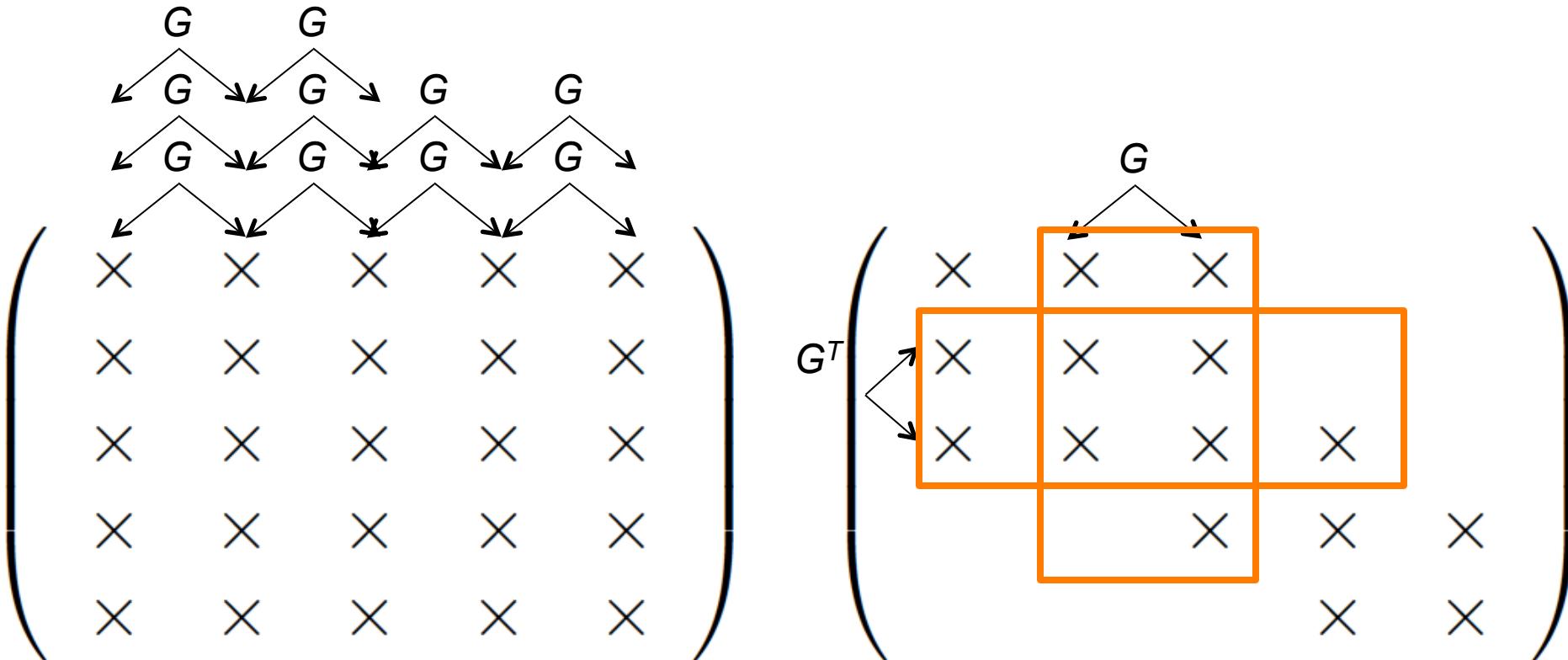
Accumulating Rotations (libflame)



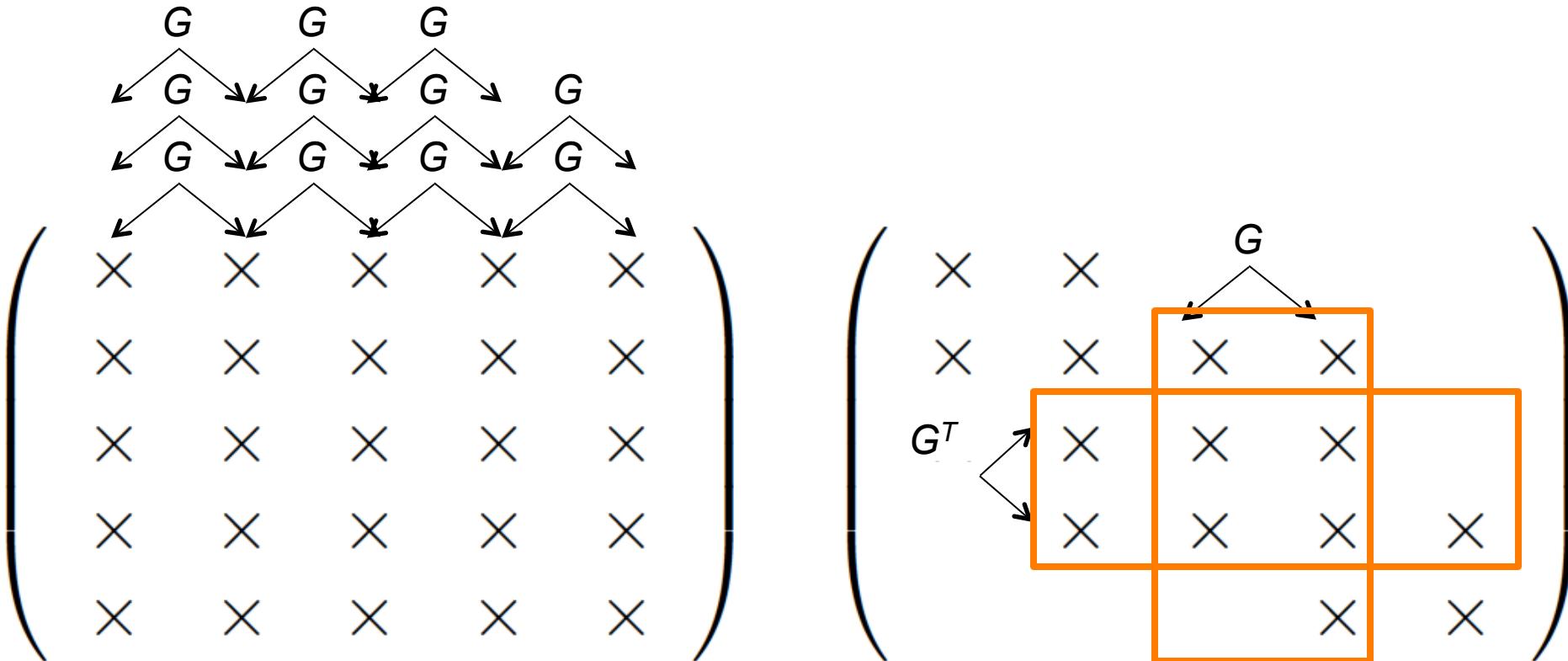
Accumulating Rotations (libflame)



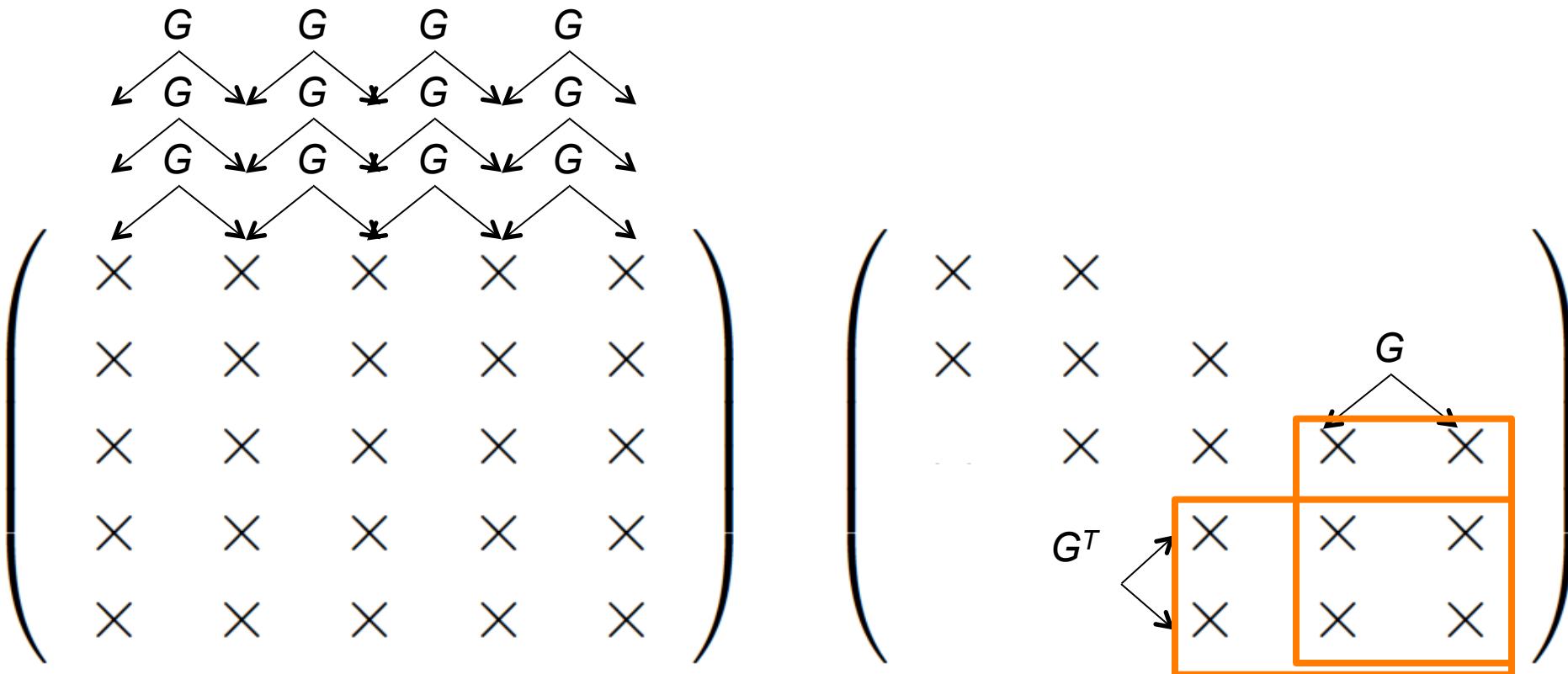
Accumulating Rotations (libflame)



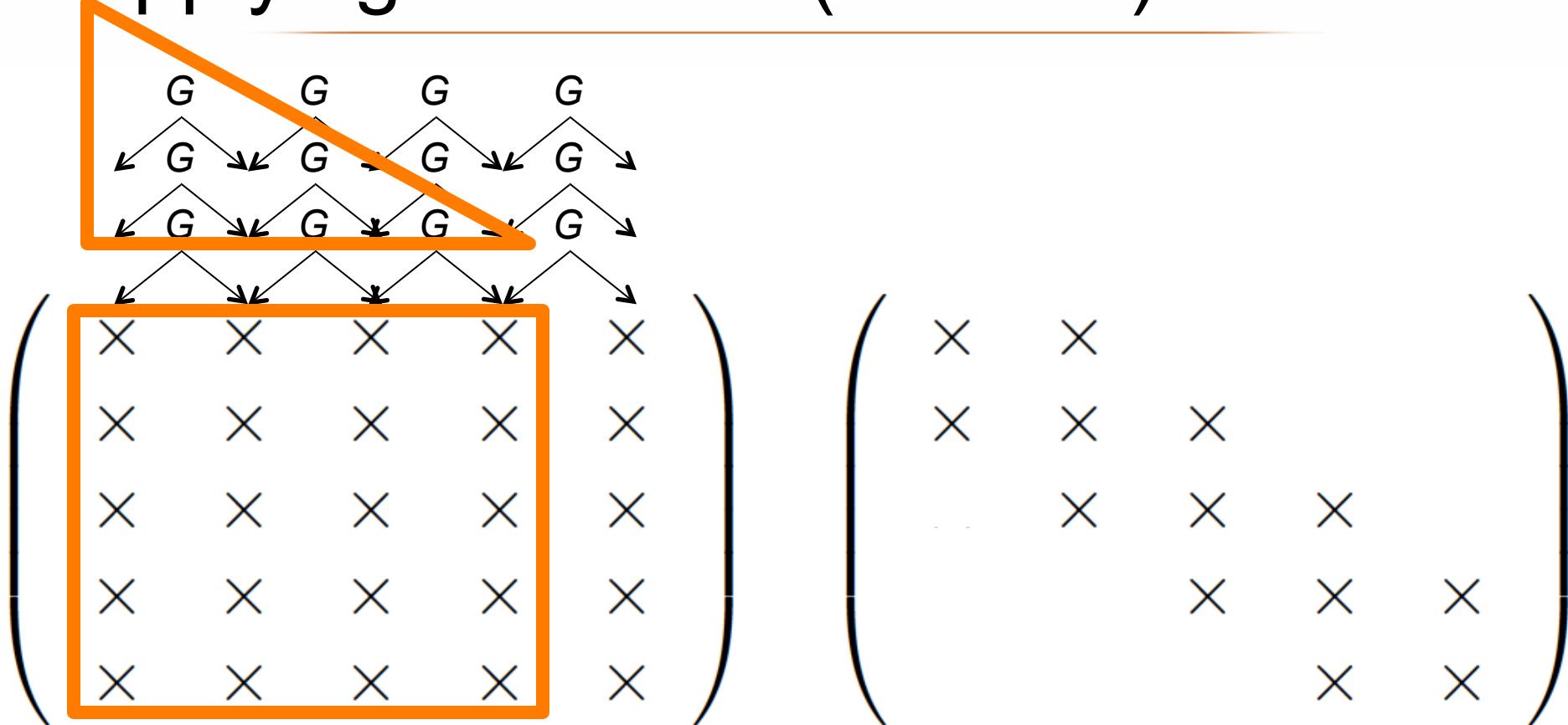
Accumulating Rotations (libflame)



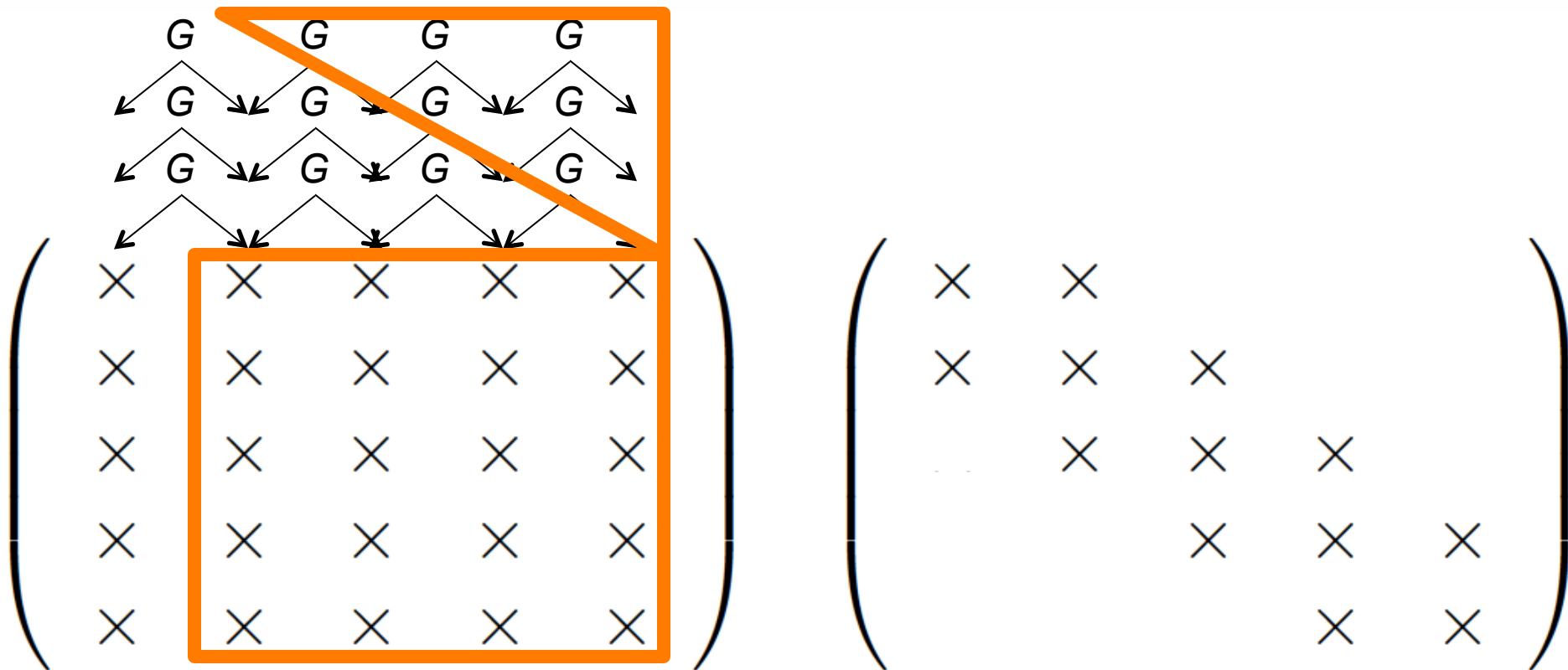
Accumulating Rotations (libflame)



Applying Rotations (libflame)



Applying Rotations (libflame)





Optimization

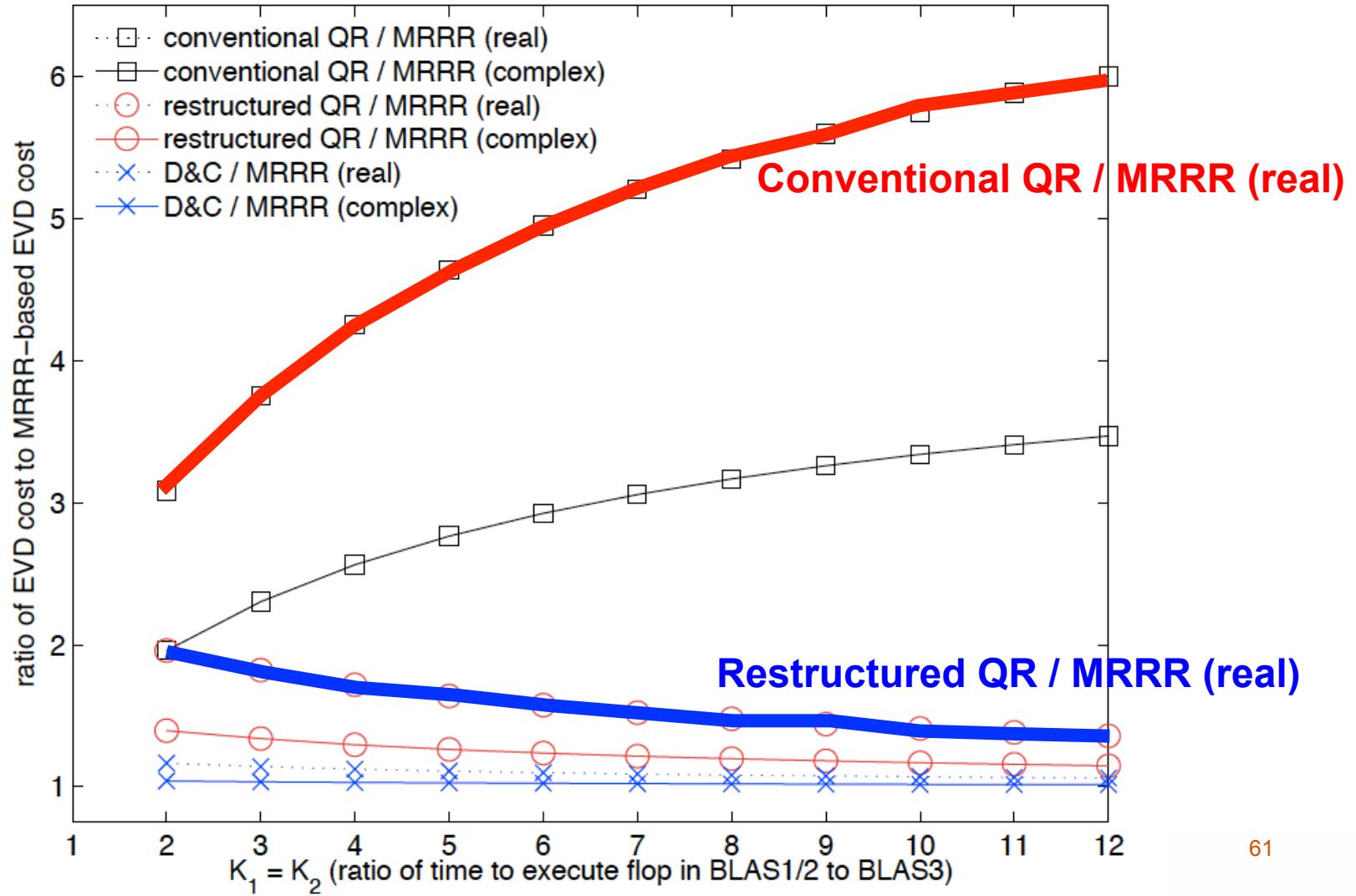
- Applying a batch of Givens' rotations:
 - $O(n^2b)$ operations on $O(n^2)$ data.
 - Can attain “level-3 BLAS” performance



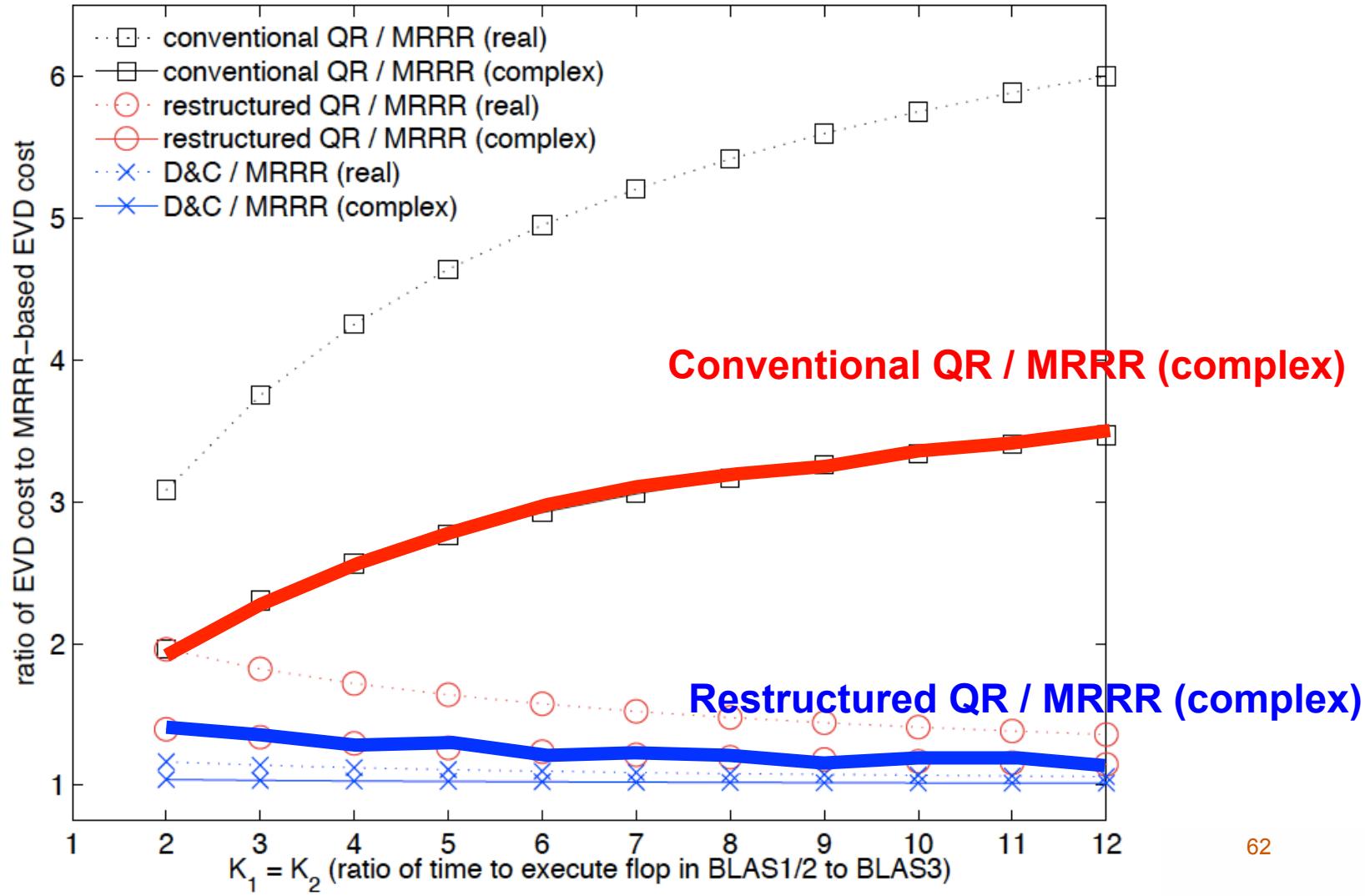
Overview

- 50+ years of progress
- The hidden costs of MRRR and D&C
- QR algorithm basics
- Accumulating and applying rotations
- **Performance**
- Conclusion

Predicted Performance



Predicted performance (EVD)

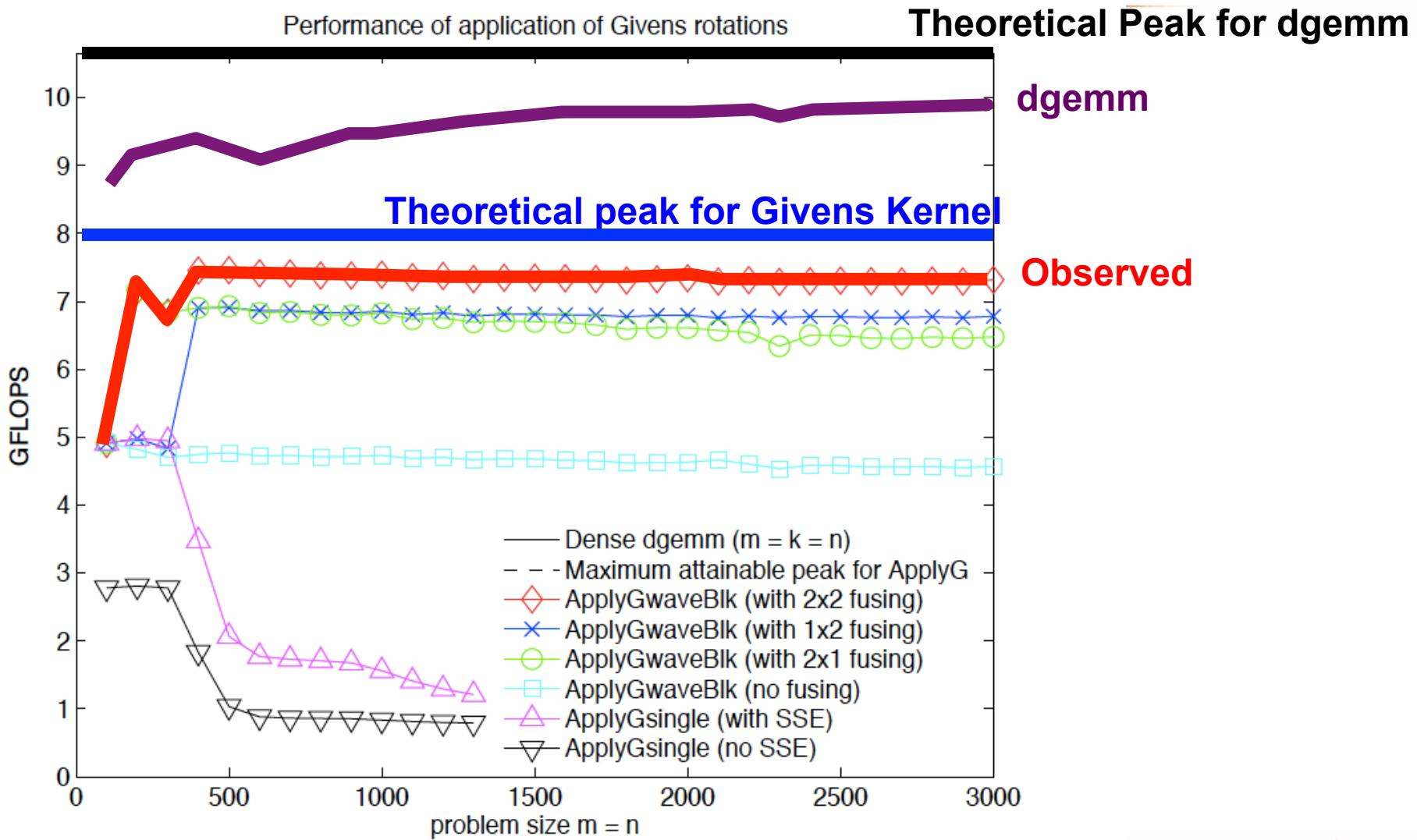




Observed Performance

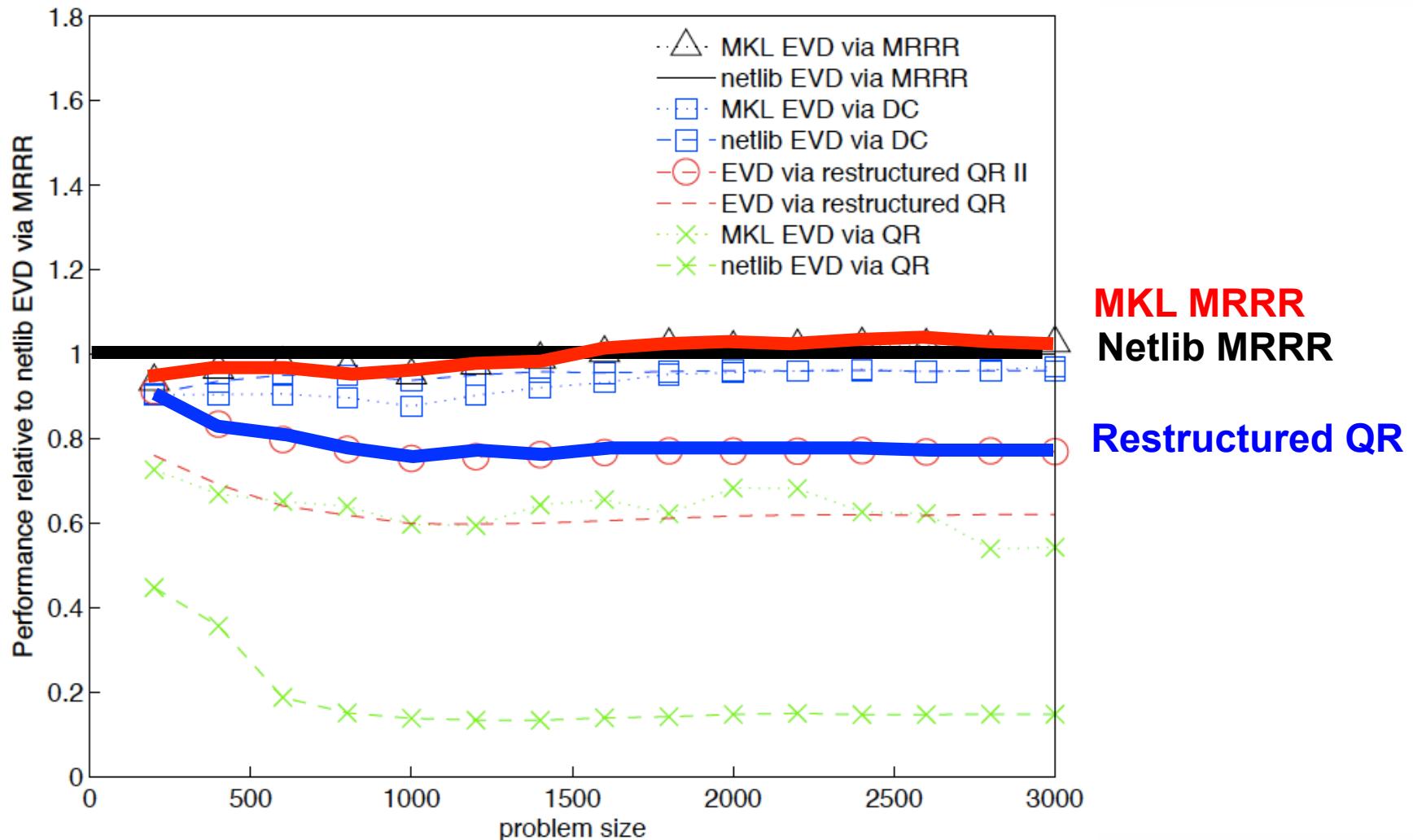
- Target architecture:
- Single core of a Dell PowerEdge R900 server
- 16 megabyte L2 cache/core.
- Single core peak of 10.64 GFLOPS.

Application of Givens rotations

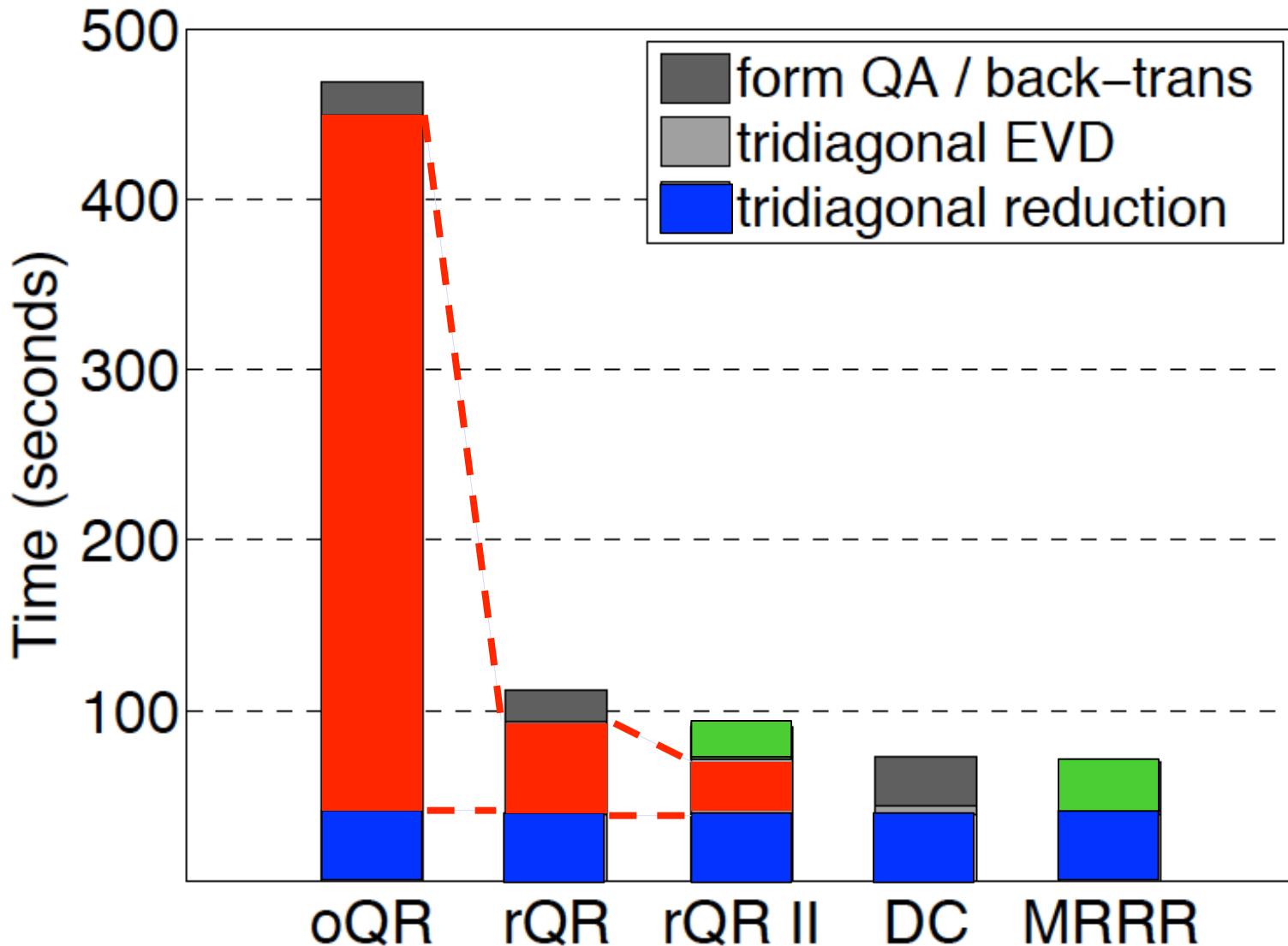


EVD performance (relative to netlib MRRR)

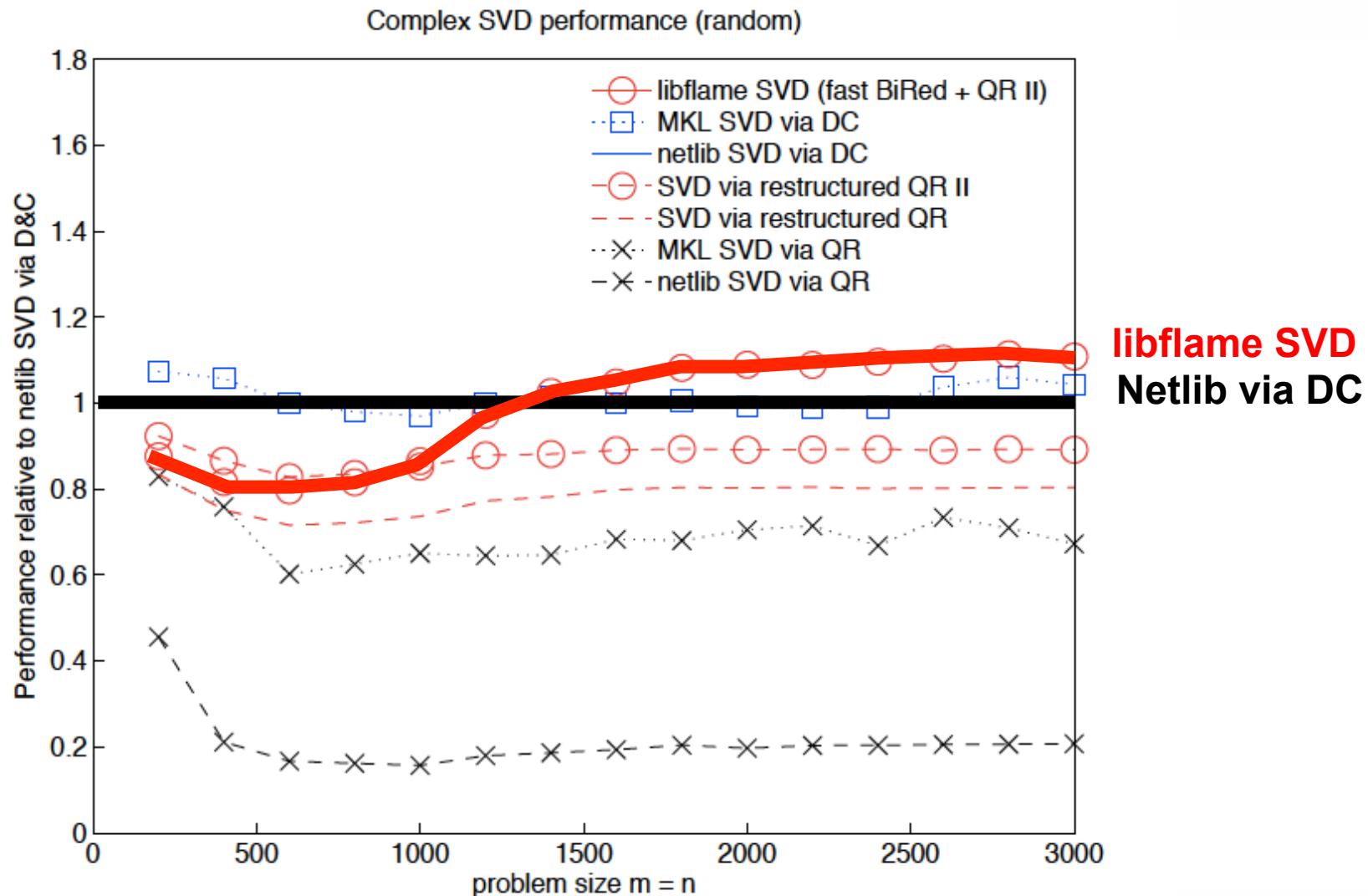
Complex Hermitian EVD performance (linear)



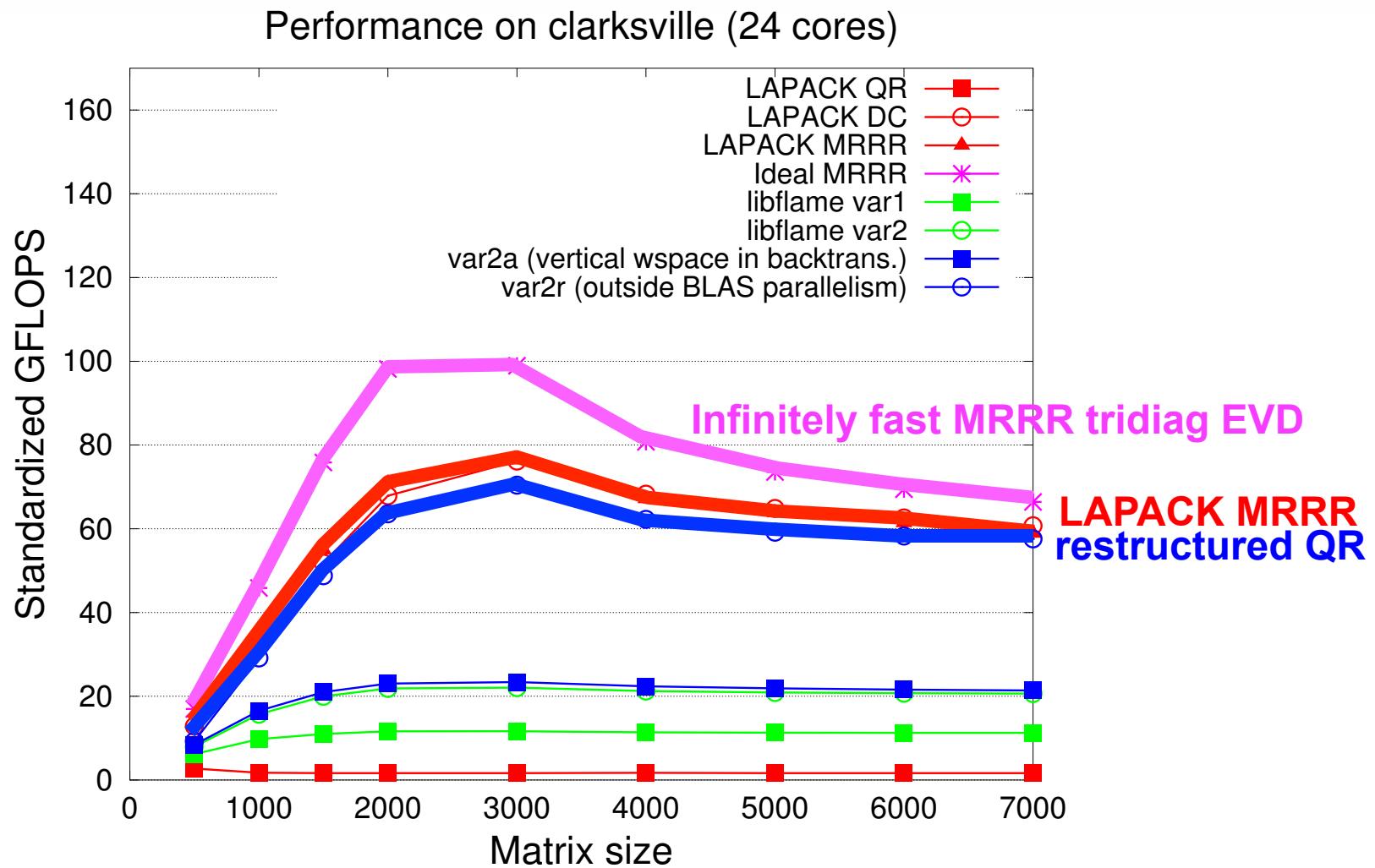
Breakdown of EVD run time (3000x3000)



libflame SVD Performance



EVD Parallel Performance (24 cores)





Is your favorite graph missing?

- The paper has an electronic appendix with tons of performance graphs.



Overview

- 50+ years of progress
- The hidden costs of MRRR and D&C
- QR algorithm basics
- Accumulating and applying rotations
- Performance
- Conclusion



Conclusion

- The QR algorithm lives!
- Future directions:
 - Parallelization
 - (multi)GPU
 - Aggressive early deflation