New results on the I/O complexity of some Numerical Linear Algebra kernels

Julien Langou

Abstract

When designing an algorithm, one cares about arithmetic/computational complexity, but data movement (I/O) complexity is playing an increasingly important role that highly impacts performance and energy consumption. The objective of I/O complexity analysis is to compute, for a given program, its minimal I/O requirement among all valid schedules. We consider a sequential execution model with two memories, an infinite one, and a small one of size S on which a computation unit retrieves and produces data. The I/O is the number of reads and writes between the two memories. From this model, we review various Numerical Linear Algebra kernels that are increasingly complicated from matrix-matrix multiplication, to LU factorization, then to symmetric rank-k update, to Cholesky factorization, then to Modified Gram-Schmidt to Householder QR factorization. We will show practical examples of these results too.

In particular, we will focus on two recent results.

First, we present the "*hourglass pattern*" which is useful in analysing algorithm such as, for example, Modified Gram-Schmidt or Householder QR factorization. We identify a common hourglass pattern in the dependency graphs of several common linear algebra kernels. Using the properties of this pattern, we mathematically prove tighter lower bounds on their I/O complexity, which improves the previous state-of-the-art bound by a parametric ratio. This proof was integrated inside the IOLB automatic lower bound derivation tool. These results were presented in [1]. In addition to lower bound results, we will show a tiling (valid for Modified Gram-Schmidt or Householder QR factorization) which enables to reach the lower bound. We present numerical experiments on modern platforms that shows the effectiveness of the new tiling.

Second, in [6], we focus on the problem of to apply a chain of sequences of Givens rotations to a matrix A. Applying a chain of Givens rotations efficiently is an important building tool in numerical linear algebra. Some examples are the implicit QR algorithm [2] and the Jacobi method for the singular value decomposition [3]. To achieve high performance, many factorizations limit their initial calculations to a smaller submatrix of the original matrix. Updating the rest of the matrix (which often involves the bulk of the floating-point operations) can then be done efficiently with an optimized routine. In practice, we observe that a vanilla algorithm performs poorly and is memory-bound. However this algorithm has a three-loop structure reminiscent of a Level 3 BLAS subroutine, and one would want to reorganize the operations to get a compute-bound algorithm. Kågström et al. [4] and later Van Zee et al. [5] demonstrated two ways to increase efficiency: wavefront pattern and fused rotations. We present a new algorithm that is innovative in three main ways. Firstly, we introduce a kernel that is optimized for register reuse in a novel way. Secondly, we introduce a blocking and packing scheme that improves the cache efficiency of the algorithm. Finally, we thoroughly analyze the memory operations of the algorithm which leads to important theoretical insights and makes it easier to select good parameters. Numerical experiments show that our algorithm outperforms the state-of-the-art and achieves a flop rate close to the theoretical peak on modern hardware. In addition to a practical new algorithm, we use our I/O lower bound theory to prove that our tiling is optimal in terms of I/O. A technical report explaining these new findings will be released soon.

1. Lionel Eyraud-Dubois, Guillaume Iooss, Julien Langou, and Fabrice Rastello. Tightening I/O lower bounds through the hourglass dependency pattern. In *the 36th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA'24)*, Nantes, France, June 17–21, 2024. DOI: 10.1145/3626183.3659986.

2. John GF Francis. The QR transformation a unitary analogue to the LR transformation—Part 1. The Computer Journal, 4(3):265–271, 1961. DOI: 10.1093/comjnl/4.3.265.

3. Carl Gustav Jacob Jacobi. Über ein leichtes verfahren die in der theorie der säcularstörungen vorkommenden gleichungen numerisch aufzulösen. Journal für die reine und angewandte Mathematik, 30:51-94, 1846. DOI: 10.1017/CBO9781139568012.016

4. Bo Kågström, Daniel Kressner, Enrique S. Quintana-Ortí, and Gregorio Quintana-Ortí. Blocked algorithms for the reduction to Hessenberg-triangular form revisited. BIT Numerical Mathematics, 48:563–584, 2008. DOI: 10.1007/s10543-008-0180-1

5. Field G. Van Zee, Robert A. van de Geijn and Gregorio Quintana-Ortí. Restructuring the tridiagonal and bidiagonal QR algorithms for performance. ACM Transactions on Mathematical Software (TOMS), 40(3):1–34, 2014. DOI: 10.1145/2535371.

6. Thijs Steel and Julien Langou. Communication efficient application of chains of sequences of planar rotations to a matrix. Technical Report to be released late 2024 or early 2025.