Automated High-Level Loop Fusion for FLAME Algorithms

Krzysztof A. Drewniak

Carnegie Mellon University

June TODO, 2018

High-level loop fusion

- Problems often are a series of subproblems
- Combining subalgorithms often helps performance
- Goal: find all the fused algorithms for a problem
- Compilers know too many details - need a high level approach

FLAME algorithms, loop invariants

- FLAME = Formal Linear Algebra Methods Eenvironments
- Provably correct algorithms from spec
- Algorithms $\Leftrightarrow$ loop invariants
- We know how to:
  - Autogenerate algorithm/code from loop invariant
  - Autogenerate all possible loop invariants
  - Identify when fusion is possible (in theory)

What we add

- Autogenerate all sets of fusable loop invariants
- Input is *partitioned matrix expression* — indicates needed computations
- Can be used to generate code

Goal

Want to compute

$$\widetilde{A} = \mathcal{F}(\hat{(}A), \underbrace{\ldots}_{O})$$

$\hat{A}$ and $\widetilde{A}$ share memory (A).
Initially, $A = \hat{A}$.
At termination, $A = \widetilde{A}$.

$$\widetilde{A} = CHOL(\hat{A})$$

Algorithm structure

partition $A \rightarrow \left(\begin{array}{c||c} A_{TL} & A_{TR} \\ \hline\hline A_{BL} & A_{BR} \end{array}\right)$

where $\dim(A_{TL}) = 0 \times 0$

do until $\dim(A_{TL}) = n \times n$

repartition $\left(\begin{array}{c||c} A_{TL} & A_{TR} \\ \hline\hline A_{BL} & A_{BR} \end{array}\right) \rightarrow \left(\begin{array}{c||c|c} A_{00} & a_{01} & A_{02} \\ \hline\hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline A_{02} & a_{21} & A_{22} \end{array}\right)$

$\vdots$] loop body

continue with $\left(\begin{array}{c||c} A_{TL} & A_{TR} \\ \hline\hline A_{BL} & A_{BR} \end{array}\right) \leftarrow \left(\begin{array}{c|c||c} A_{00} & a_{01} & A_{02} \\ \hline a_{10}^T & \alpha_{11} & a_{12}^T \\ \hline\hline A_{20} & a_{21} & A_{22} \end{array}\right)$

enddo

Algoriithm example

partition $A \rightarrow \left(\begin{array}{c||c} A_{TL} & * \\ \hline\hline A_{BL} & A_{BR} \end{array}\right)$

where $\dim(A_{TL}) = 0 \times 0$

do until $\dim(A_{TL}) = n \times n$

repartition $\left(\begin{array}{c||c} A_{TL} & * \\ \hline\hline A_{BL} & A_{BR} \end{array}\right) \rightarrow \left(\begin{array}{c||c|c} A_{00} & * & * \\ \hline\hline a_{10}^T & \alpha_{11} & * \\ \hline A_{02} & a_{21} & A_{22} \end{array}\right)$

$\alpha_{11} := \sqrt{\alpha_{11}}$

$a_{21} := a_{21}/\alpha_{11}$

$A_{22} := A_{22} - a_{21}a_{21}^T$

continue with $\left(\begin{array}{c||c} A_{TL} & * \\ \hline\hline A_{BL} & A_{BR} \end{array}\right) \leftarrow \left(\begin{array}{c|c||c} A_{00} & * & * \\ \hline a_{10}^T & \alpha_{11} & * \\ \hline\hline A_{20} & a_{21} & A_{22} \end{array}\right)$

enddo

Partitioned Matrix Expressions

- Take A (and maybe other stuff), split it into regions.
- Lines between regions move during algorithm

$$\left(\begin{array}{c||c} \widetilde{A}_{TL} = \mathcal{F}_{TL}(\hat{A}, \ldots) & \widetilde{A}_{TR} = \mathcal{F}_{TR}(\hat{A}, \ldots) \\ \hline\hline \widetilde{A}_{BL} = \mathcal{F}_{BL}(\hat{A}, \ldots) & \widetilde{A}_{BR} = \mathcal{F}_{BR}(\hat{A}, \ldots) \end{array}\right)$$

$$\left(\begin{array}{c||c} \widetilde{A}_{TL} = CHOL(\hat{A}_{TL}) & * \\ \hline\hline \widetilde{A}_{BL} = \hat{A}_{BL}\widetilde{A}_{TL}^{-T} & \widetilde{A}_{BR} = CHOL(\hat{A}_{BR} - \widetilde{A}_{BL}\widetilde{A}_{BL}^{T}) \end{array}\right)$$

Loop invariants

- Find f_R and f'_R so $\mathcal{F}_R(\hat{A}) = f'(f(\hat{A}))$.
- f_R is loop invariant for R, f'_R is remainder
- Invariant for algorithm is an invariant per region
- Completely determine algorithm

This is a loop invariant

Starting from Cholesky's PME:

$$\left(\begin{array}{c||c} \widetilde{A}_{TL} = CHOL(\hat{A}_{TL}) & * \\ \hline\hline \widetilde{A}_{BL} = \hat{A}_{BL}\widetilde{A}_{TL}^{-T} & \widetilde{A}_{BR} = CHOL(\hat{A}_{BR} - \widetilde{A}_{BL}\widetilde{A}_{BL}^{T}) \end{array}\right)$$

We obtain

$$\left(\begin{array}{c||c} A_{TL} = CHOL(\hat{A}_{TL}) & * \\ \hline\hline A_{BL} = \hat{A}_{BL}\widetilde{A}_{TL}^{-T} & A_{BR} = \hat{A}_{BR} - \widetilde{A}_{BL}\widetilde{A}_{BL}^{T} \end{array}\right)$$

As are these

$$\left(\begin{array}{c||c} A_{TL} = CHOL(\hat{A}_{TL}) & * \\ \hline\hline A_{BL} = \hat{A}_{BL}\widetilde{A}_{TL}^{-T} & A_{BR} = \hat{A}_{BR} \end{array}\right)$$

$$\left(\begin{array}{c||c} A_{TL} = CHOL(\hat{A}_{TL}) & * \\ \hline\hline A_{BL} = \hat{A}_{BL} & A_{BR} = \hat{A}_{BR} \end{array}\right)$$

But not these

$$\left(\begin{array}{c||c} A_{TL} = CHOL(\hat{A}_{TL}) & * \\ \hline\hline A_{BL} = \hat{A}_{BL}\widetilde{A}_{TL}^{-T} & A_{BR} = CHOL(\hat{A}_{BR} - \widetilde{A}_{BL}\widetilde{A}_{BL}^{T}) \end{array}\right)$$

$$\left(\begin{array}{c||c} A_{TL} = \hat{A}_{TL} & * \\ \hline\hline A_{BL} = \hat{A}_{BL} & A_{BR} = \hat{A}_{BR} \end{array}\right)$$

Or this

$$\left(\begin{array}{c||c} A_{TL} = \hat{A}_{TL} & * \\ \hline\hline A_{BL} = \hat{A}_{BL}\widetilde{A}_{TL}^{-T} & A_{BR} = \hat{A}_{BR} \end{array}\right)$$