Optimizing Code Generation for Matrix Multiplication

Henri Willems
henri.willems@campus.tu-berlin.de
WS 23/24, TU Berlin

March 04, 2024

Matrix Multiplication

What is it
How often is it used
What is the challenge? I.e. memory or cache or computation bound

BLAS

What is it

Code Generation

What is it and why
When is it preferrable to BLAS or not? (combine with prev slide)

MLIR Code rewriting

Lower code from Daphne dialect to LLVM through affine loops
Maybe show syntax or maybe not
Effect: Enabled matmuls on these additional value types

Optimizations enabled

Tiling effect Vectorization effect

Optimizations enabled

Combined effect

Further improvements

Packing effect

Bibliography

Henri Willems
henri.willems@campus.tu-berlin.de

Appendix

Further experiments