LA-UR-22-21898

In Situ Optimization and Late Lowering of Compute Kernels

George Stelle
Pat McCormick, Daniel Shevitz, Alexis Perry-Holby, Nirmal Prajapati
T.B. Schardl, Valentin Churavy

Los Alamos National Laboratory

March 2, 2022

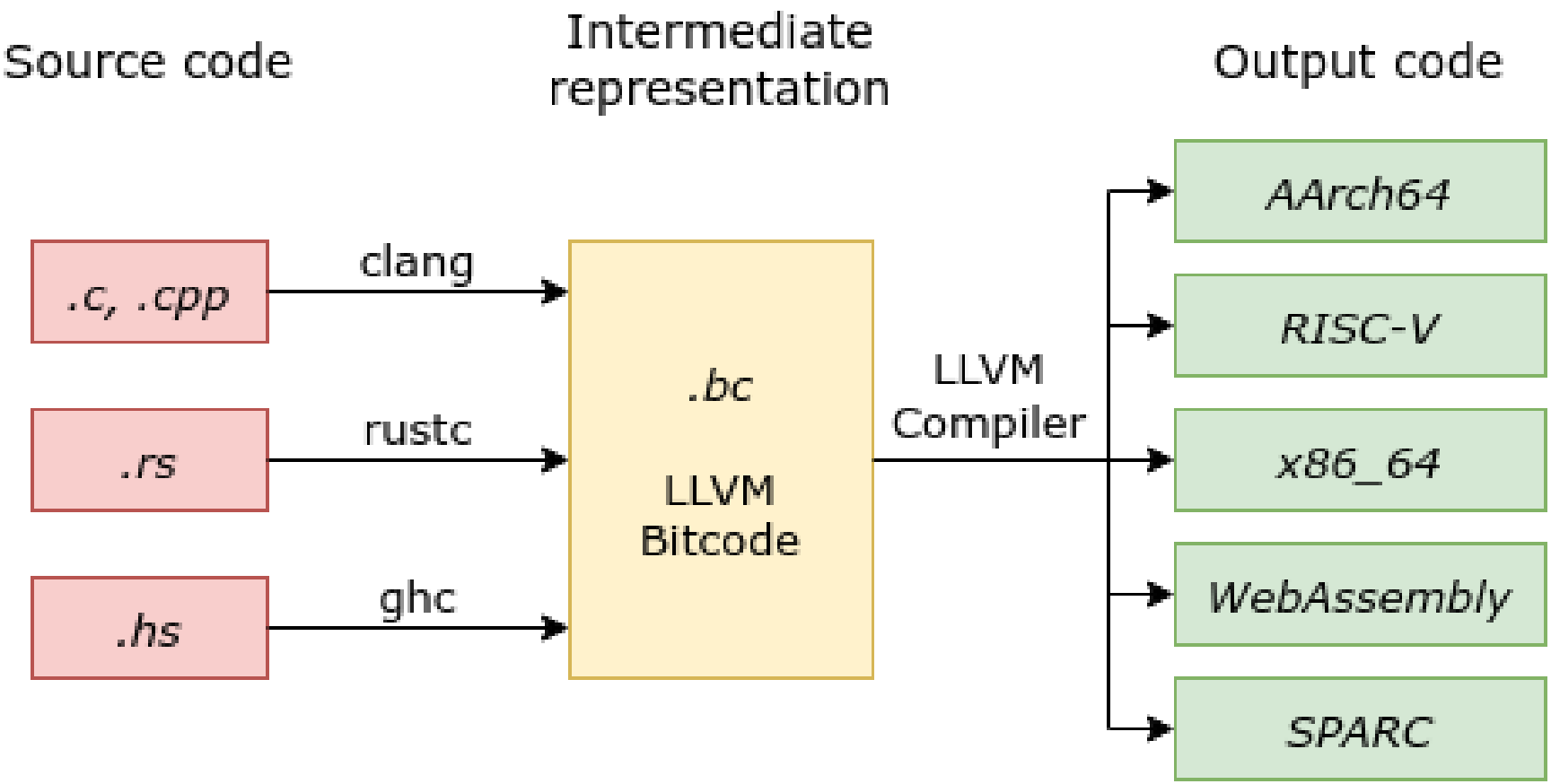
Compiler View

COMPUTE KERNEL

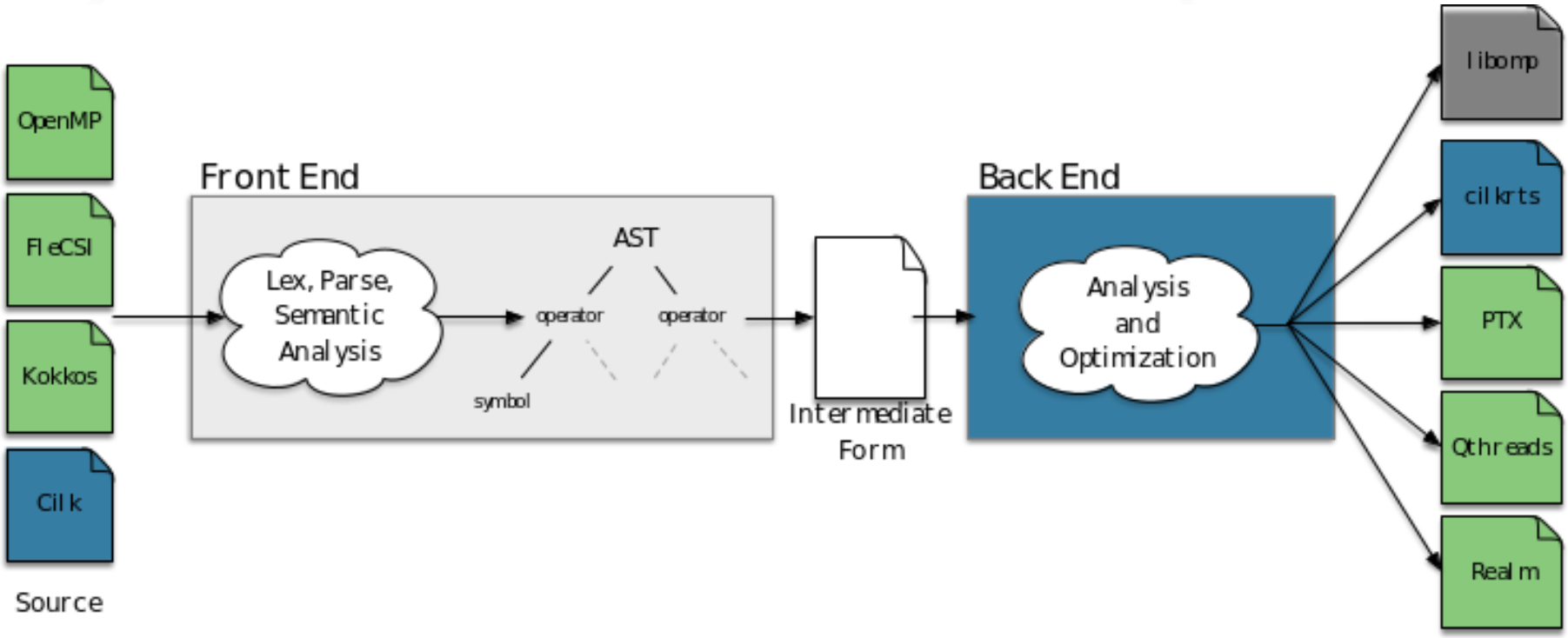
OTHER CODE

Goal

LLVM



Parallelism



Accelerators

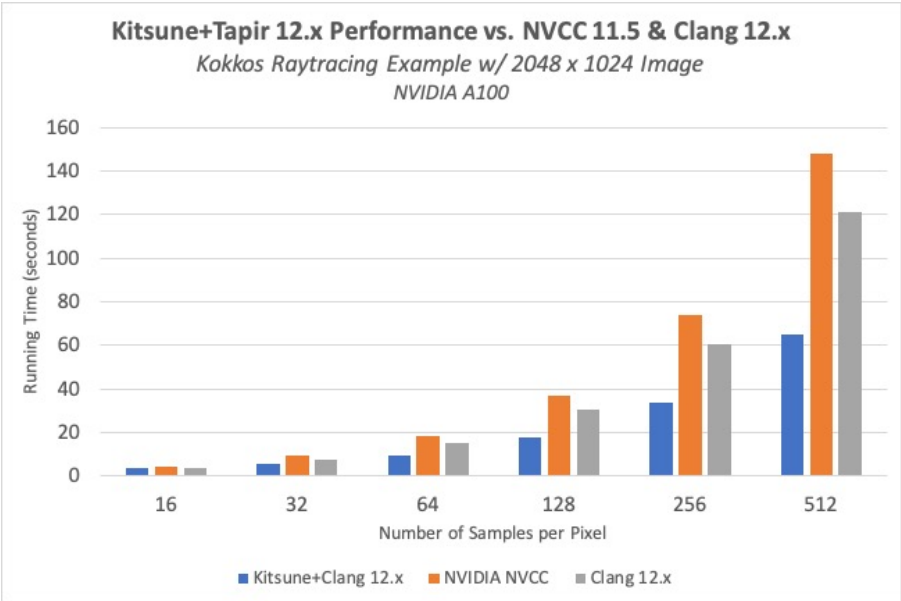
Our Approach

- Compile parallel loops into Tapir
- Optimize!
- Outline kernel, store as LLVM
- Insert calls to libllvm-gpu
- JIT compile kernel based on available devices

Legacy Approach

- Split compute kernel into separate compilation unit
- Optimize independently
- Lower to vendor ISA
- Insert calls to vendor runtime

(Very) Preliminary Results



In Progress

- Reductions
- SSA Theory Extensions
- Better JIT Utilization
- Explicit Memory Movement
- Precompilation

Questions?

https://github.com/lanl/kitsune

stelleg@lanl.gov