

Multi-Device GPU-Code Generation with LLVM

Philipp Schaad

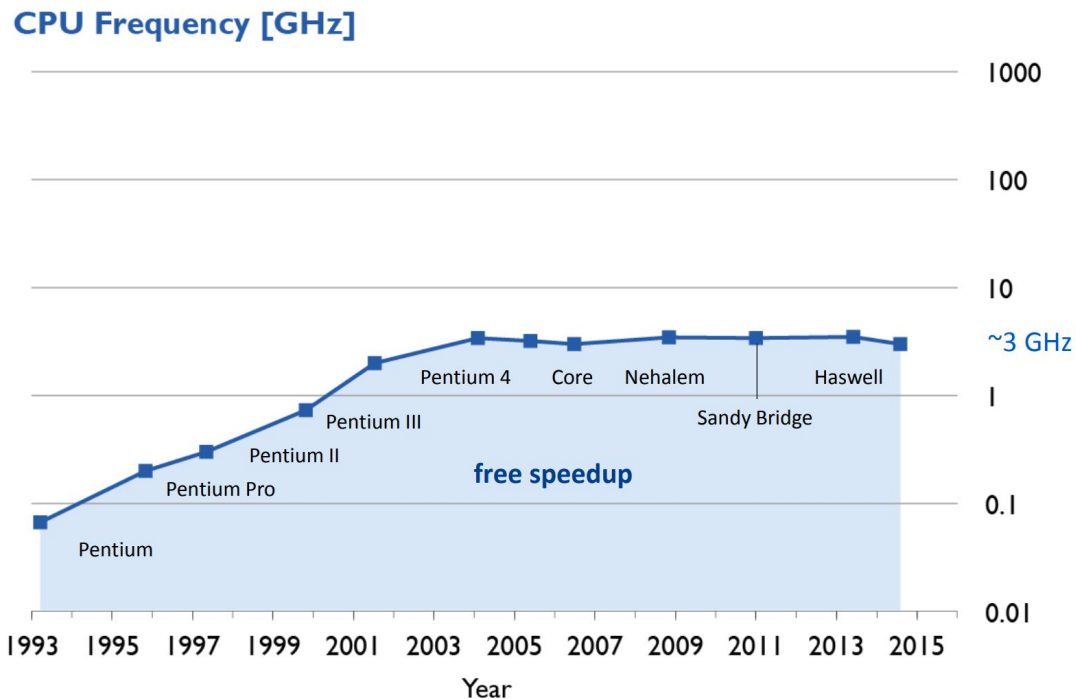
Bachelor Thesis

April - September 2017

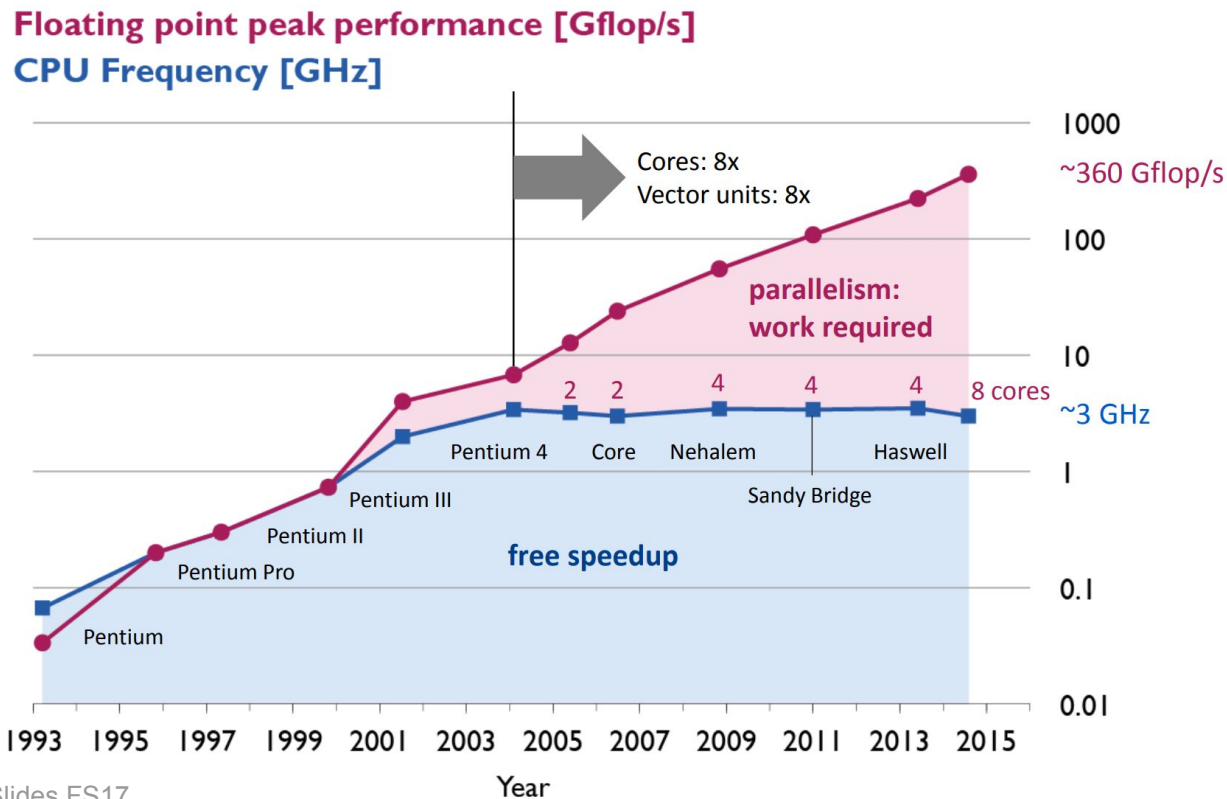
Dr. Tobias Grosser

Prof. Dr. Torsten Hoefler

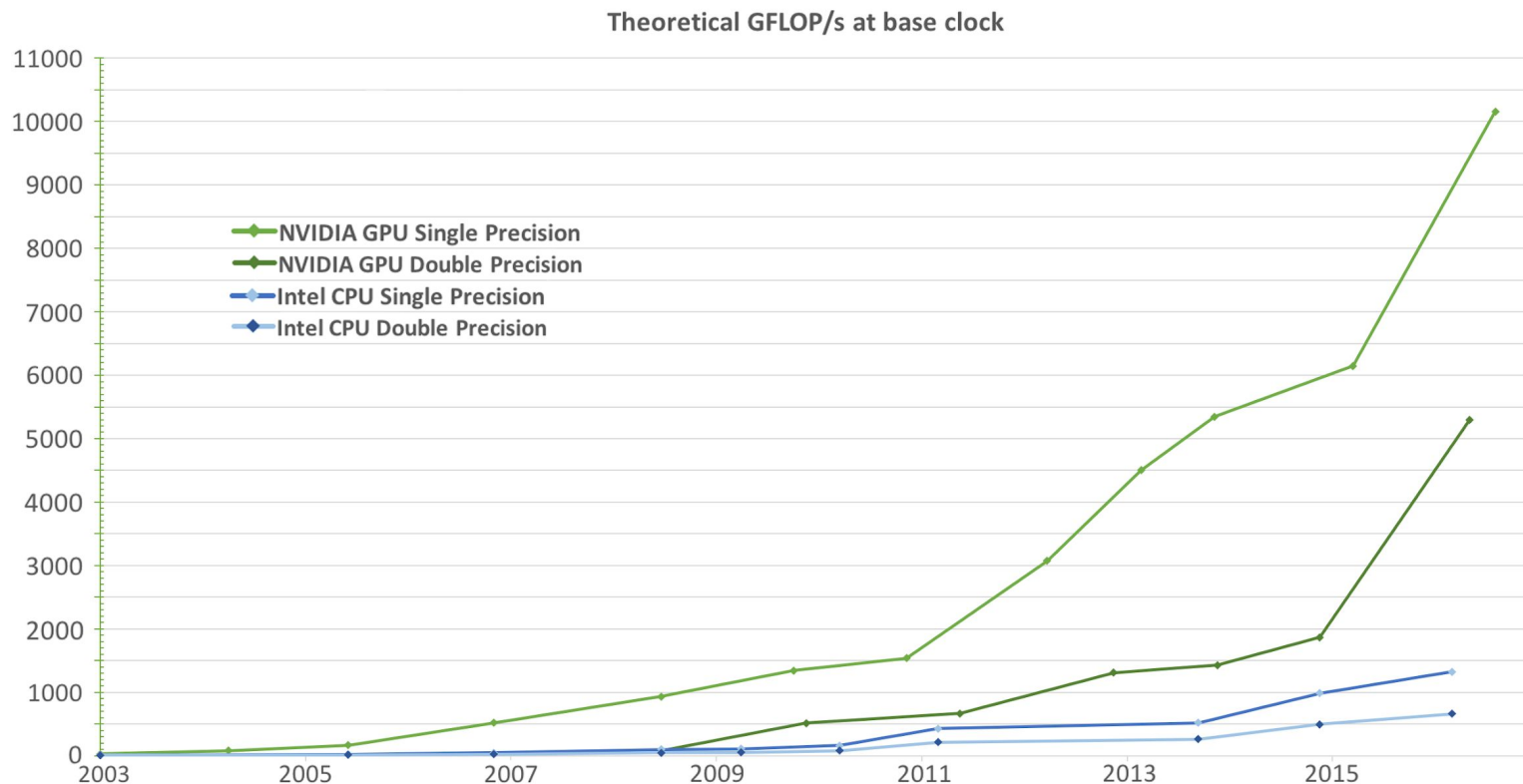
Why GPU-Code Generation?



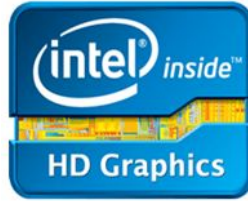
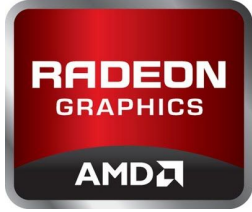
Why GPU-Code Generation?



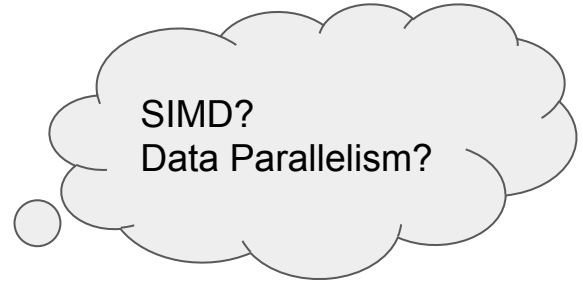
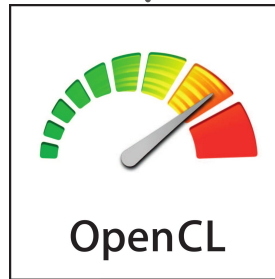
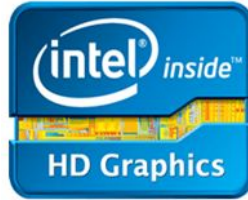
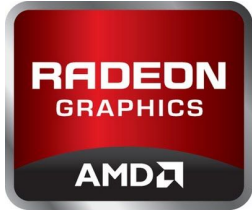
Why GPU-Code Generation?



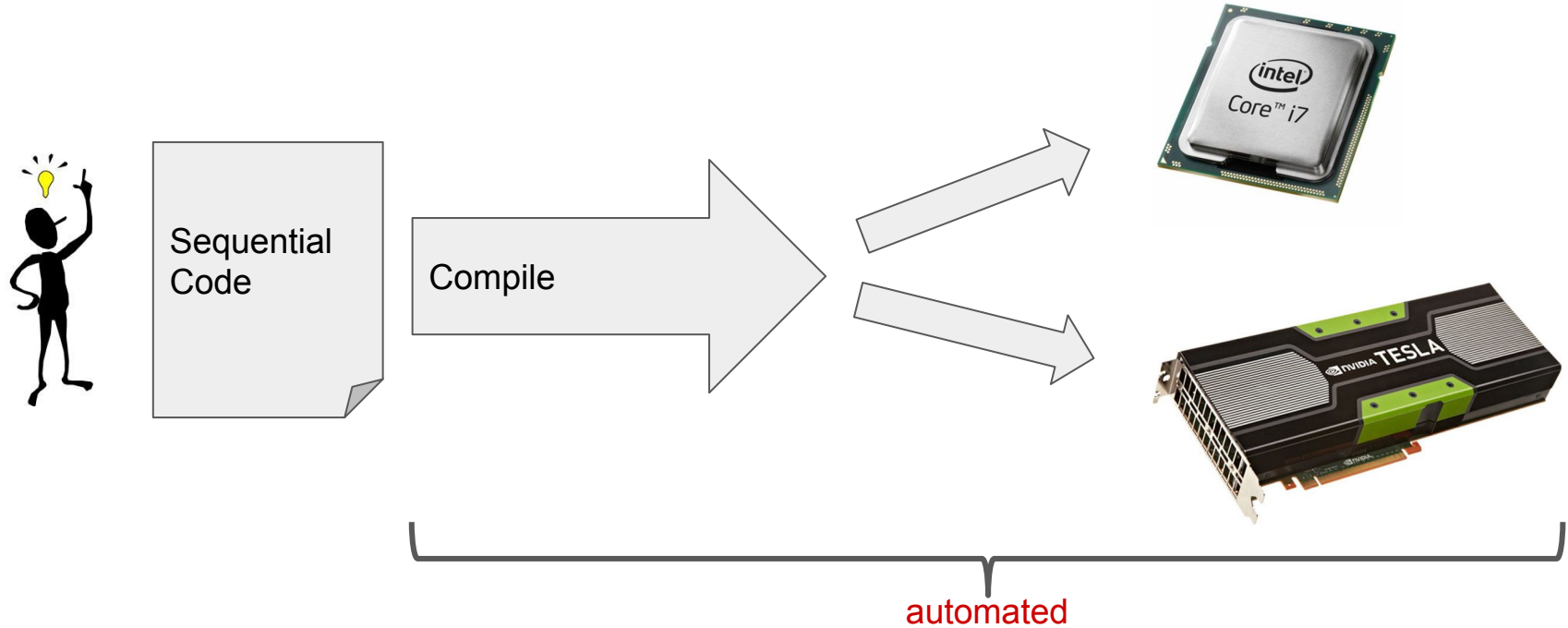
Why GPU-Code Generation?



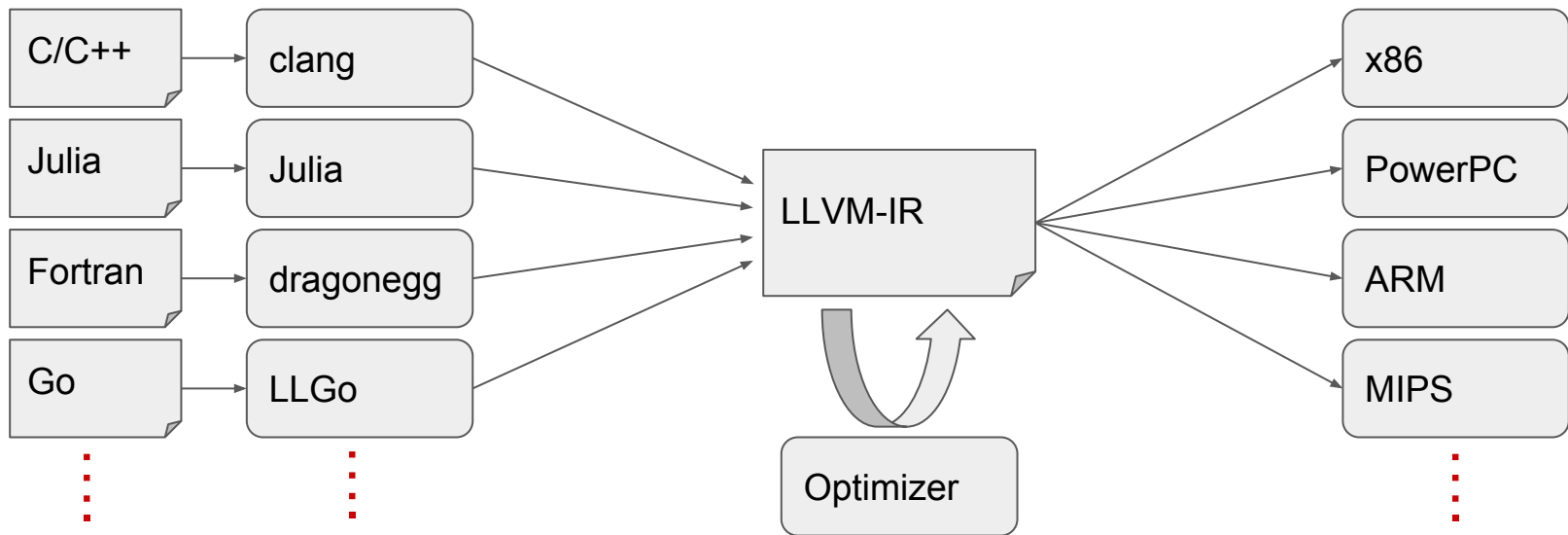
Why GPU-Code Generation?



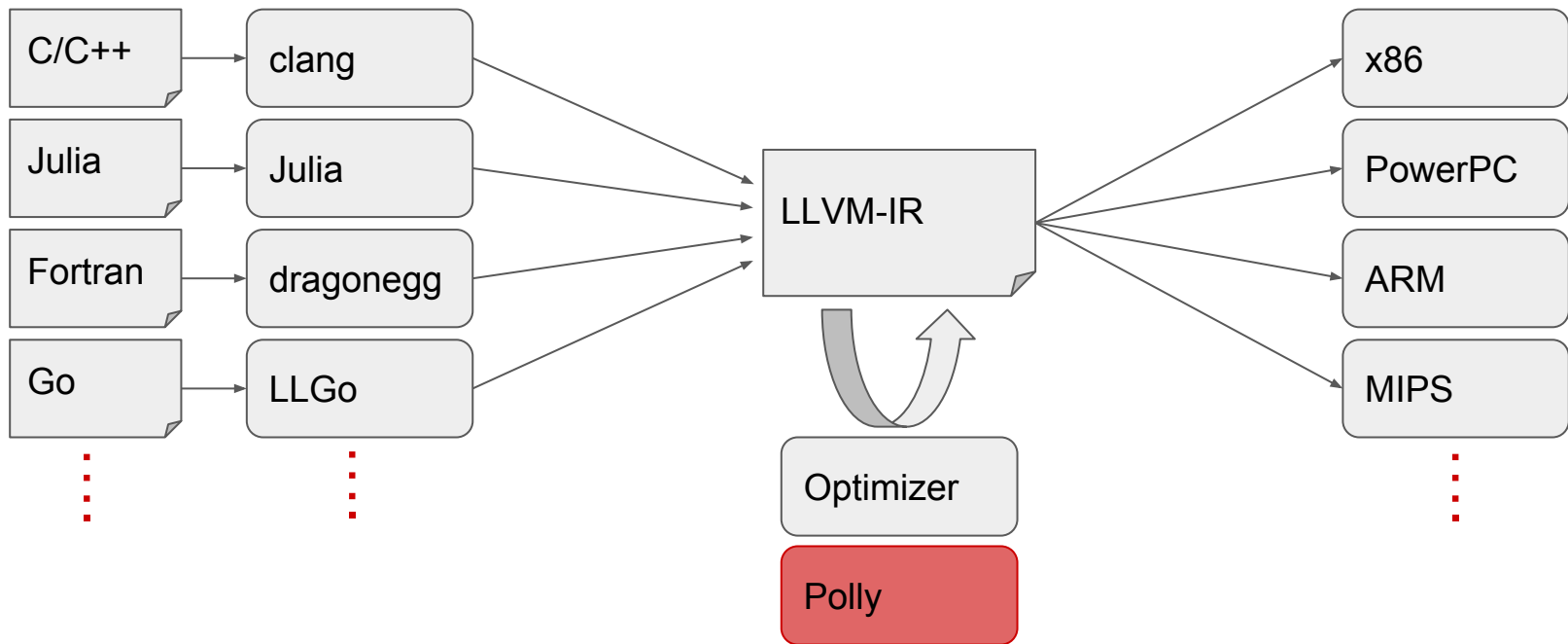
Why GPU-Code Generation?



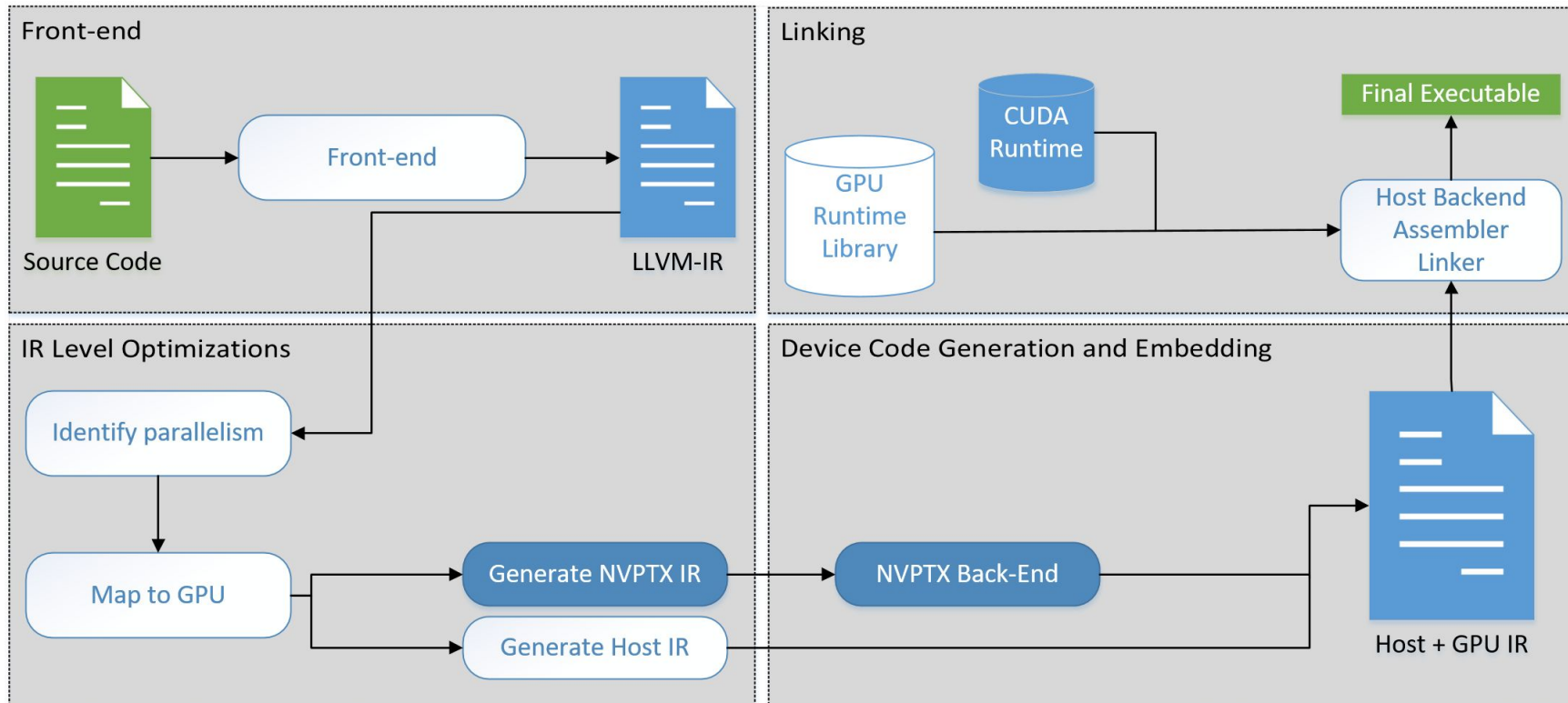
Why LLVM?



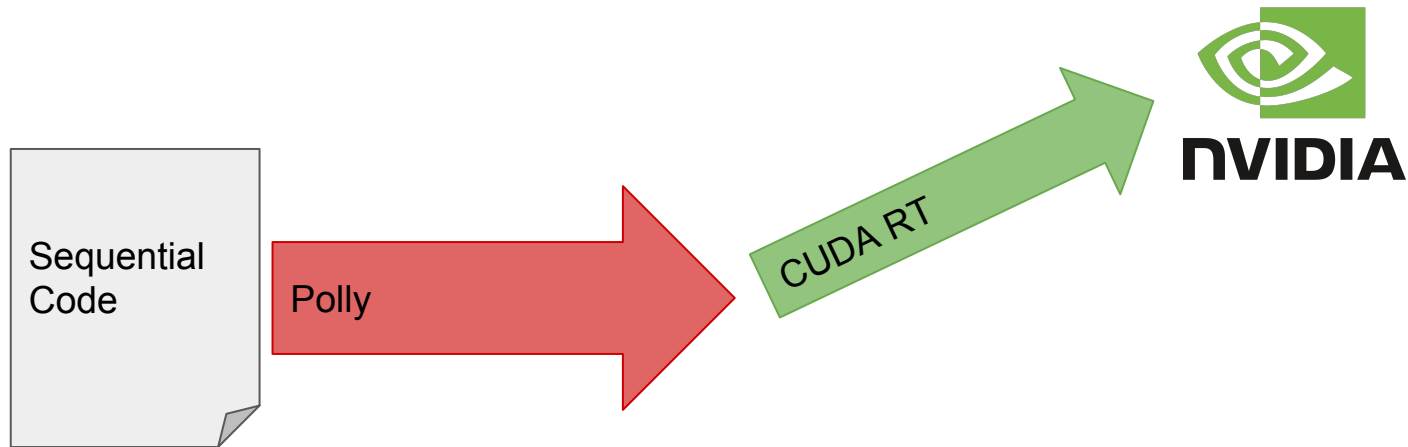
Why LLVM?



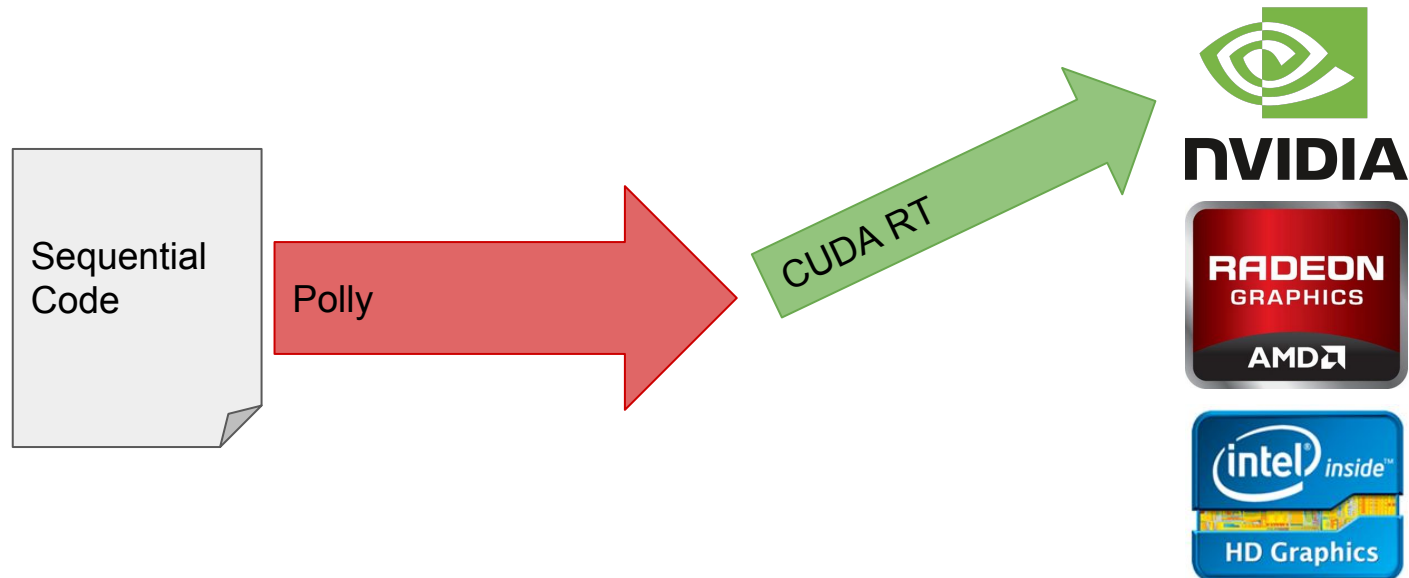
Polly's Architecture



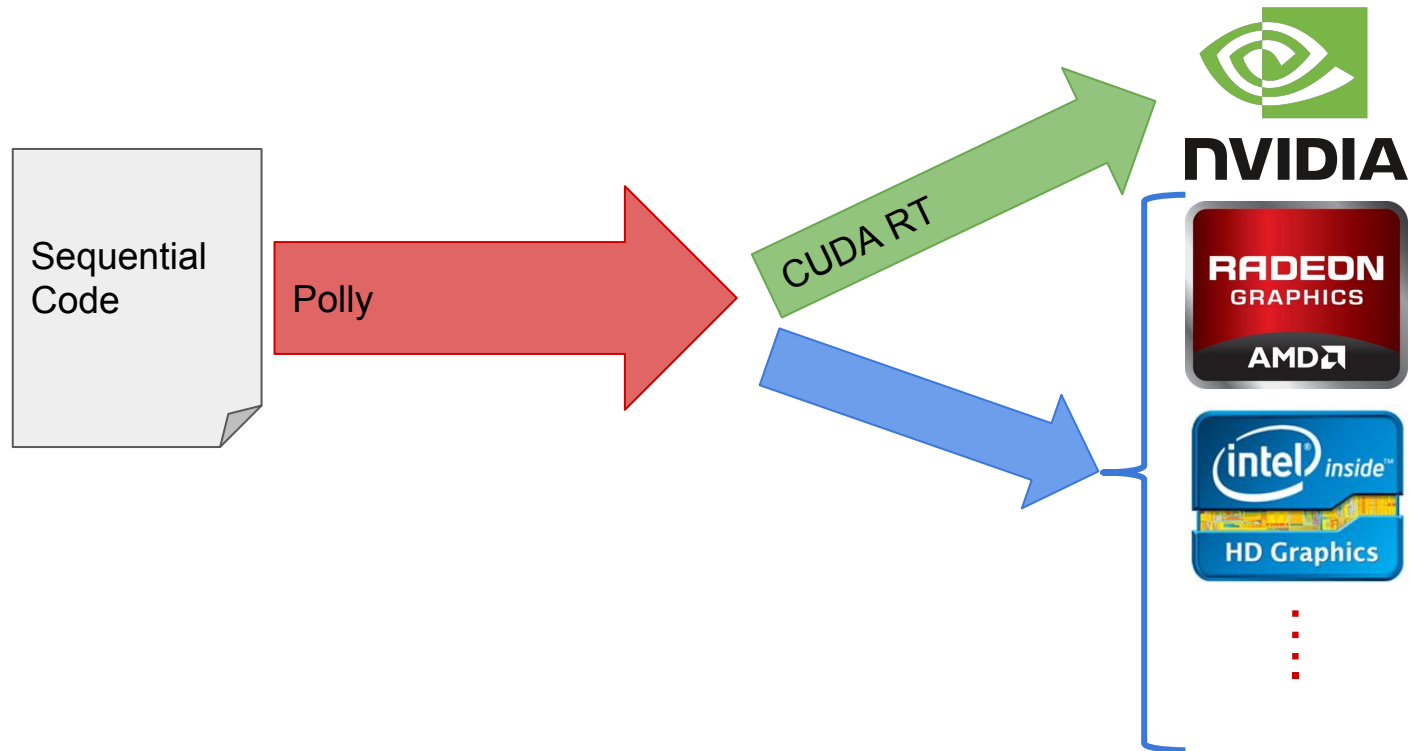
Extending Polly



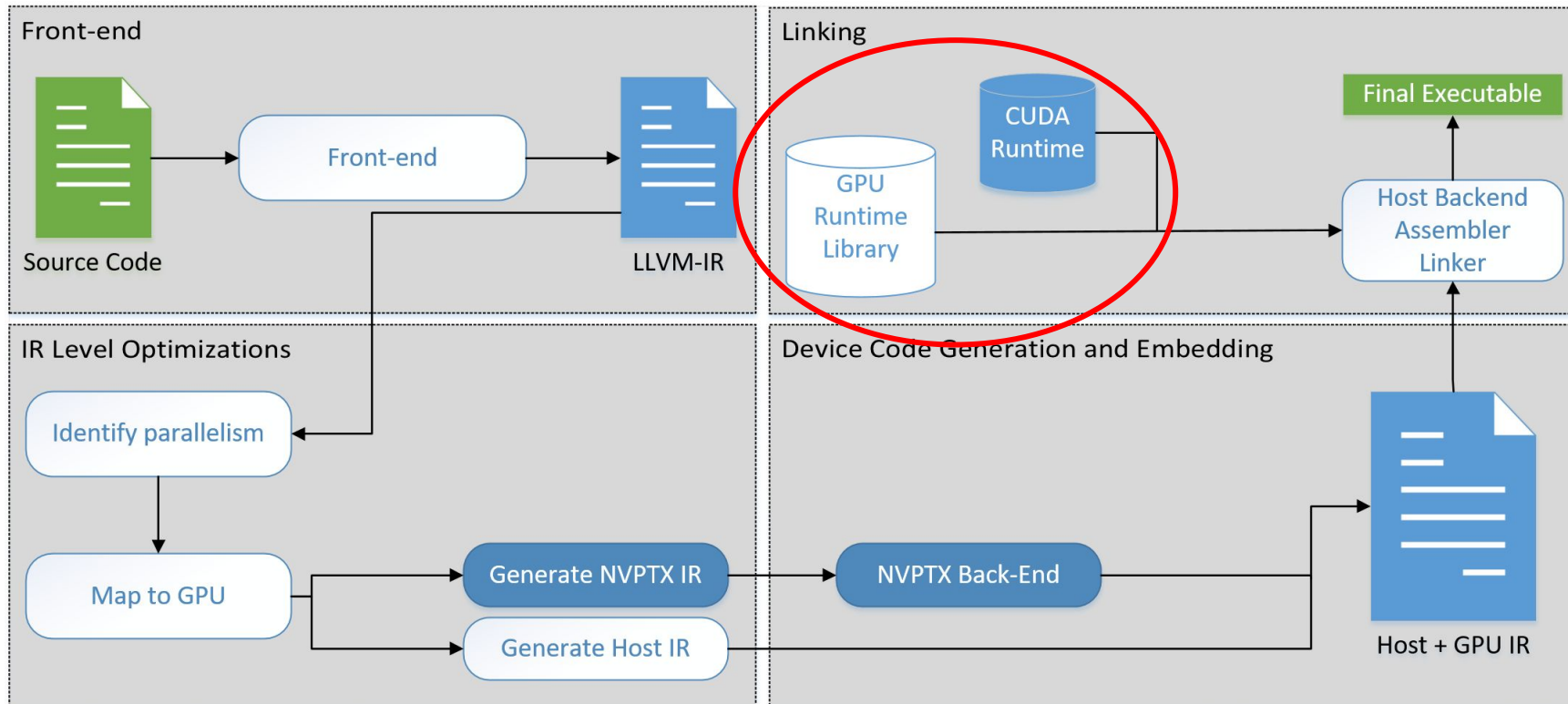
Extending Polly



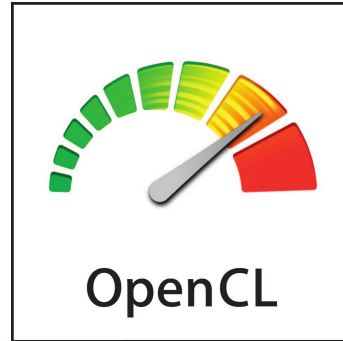
Extending Polly



CUDA as a Limiting Factor



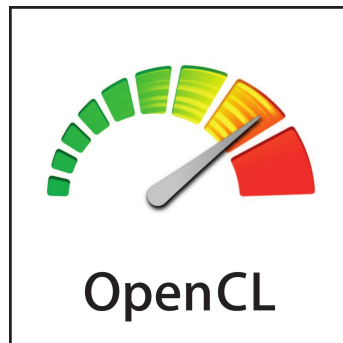
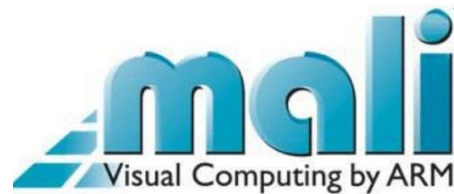
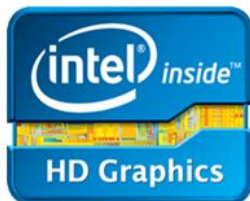
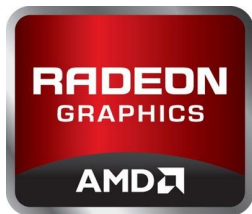
OpenCL Interface



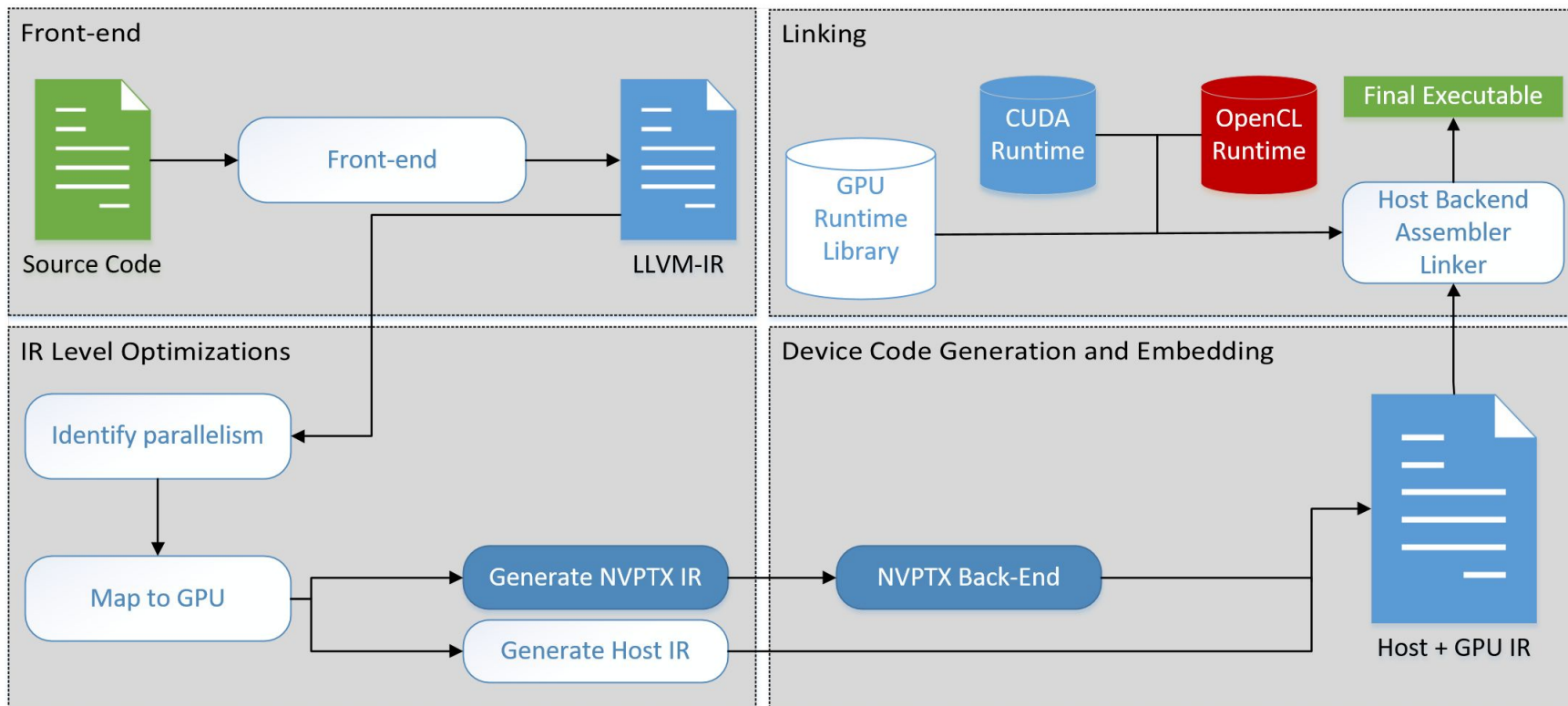
OpenCL Interface



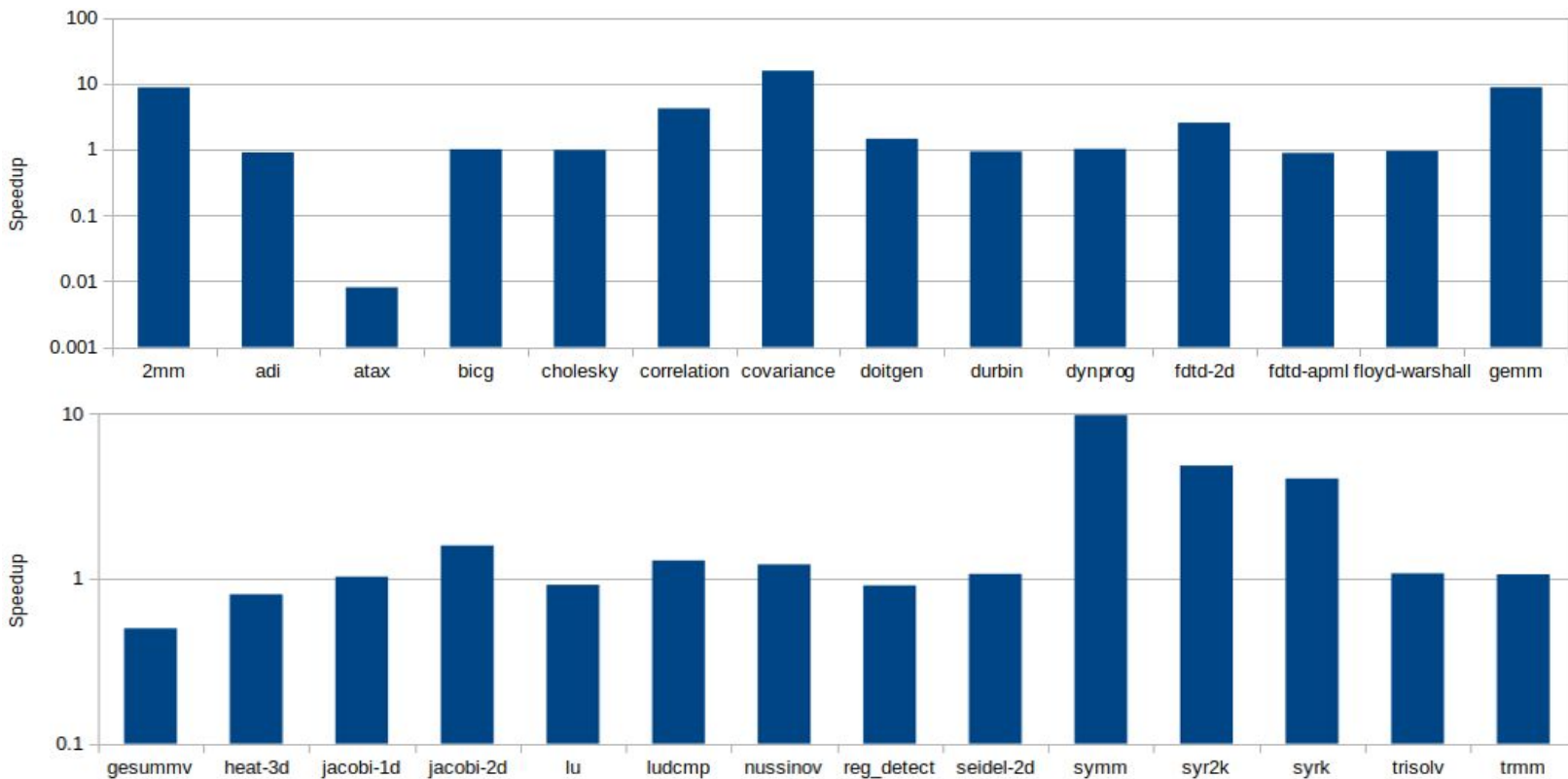
nVIDIA



OpenCL Interface

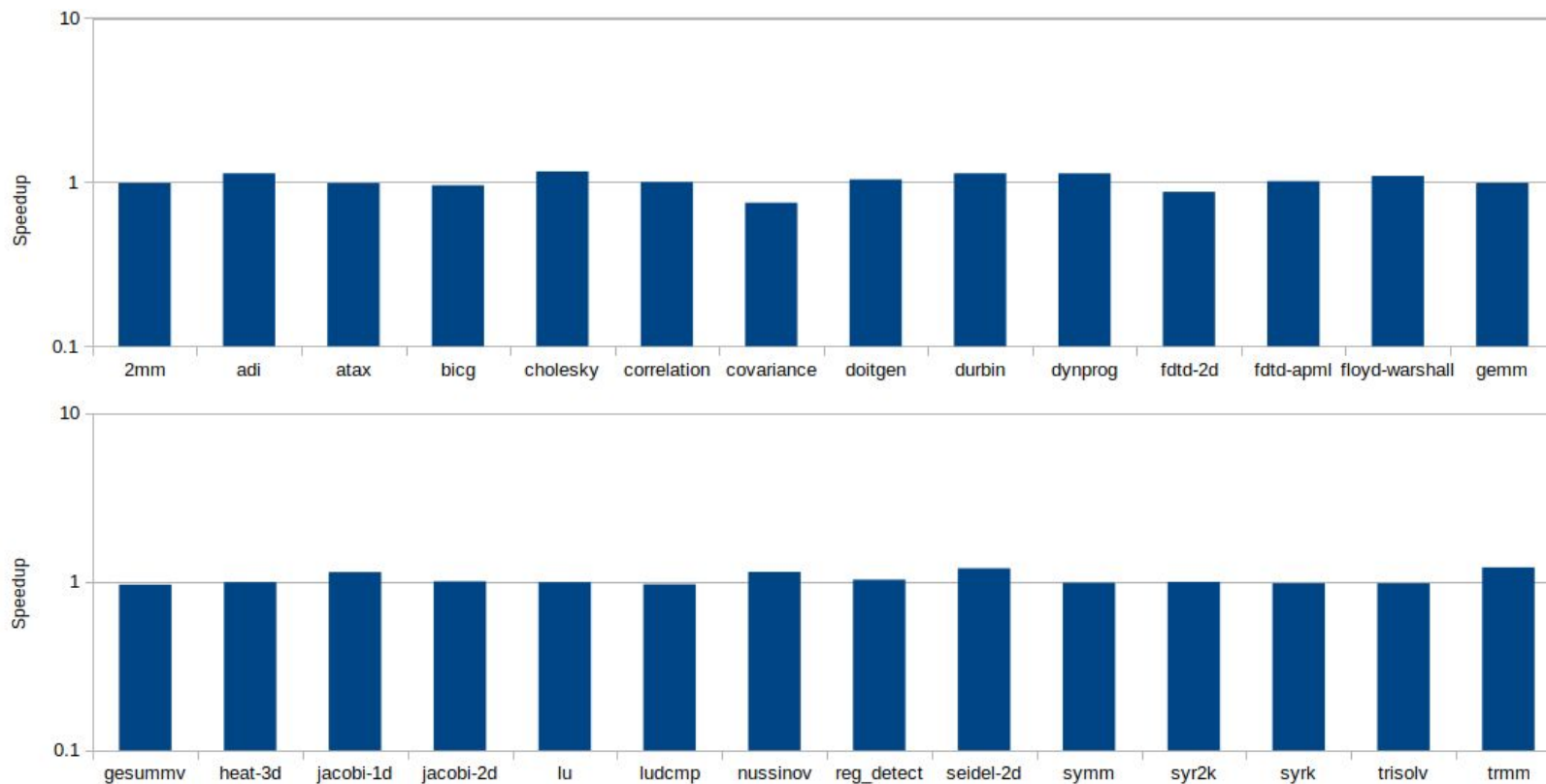


GPU Code Generation (OpenCL) Compared to -O3

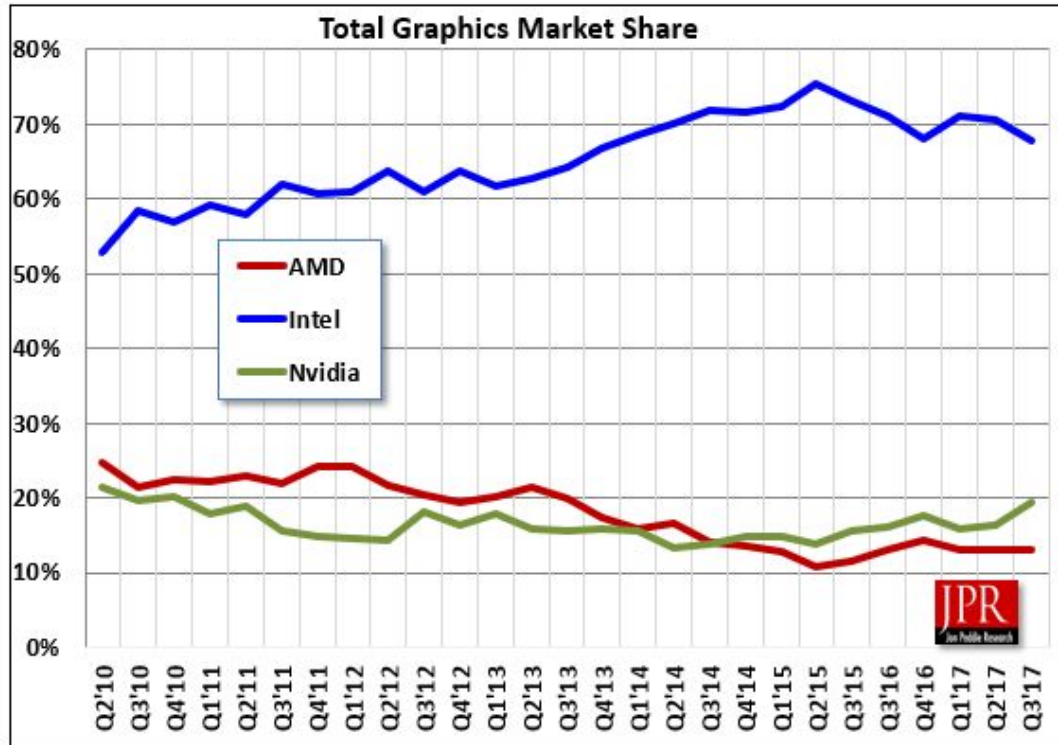


Geom. mean: 2.1x

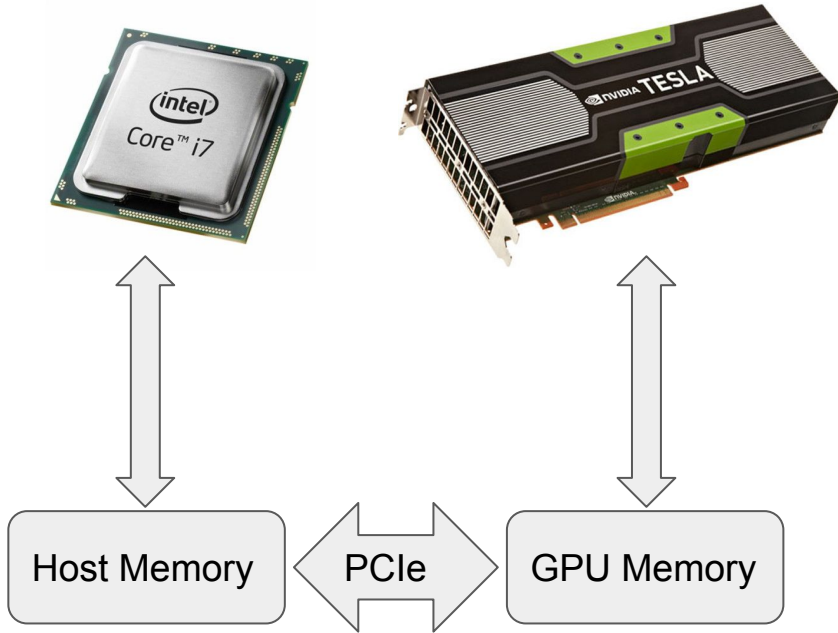
How Does It Compare to CUDA?



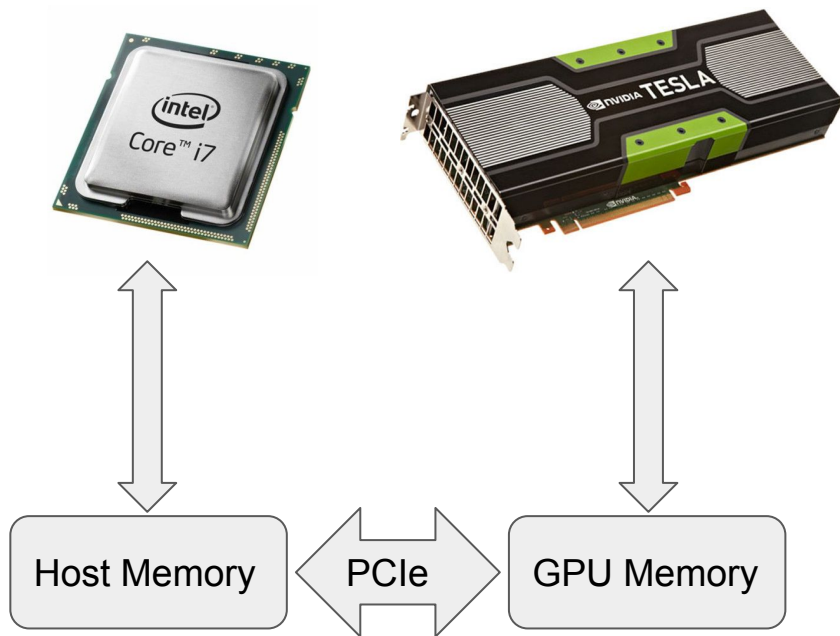
Why Intel?



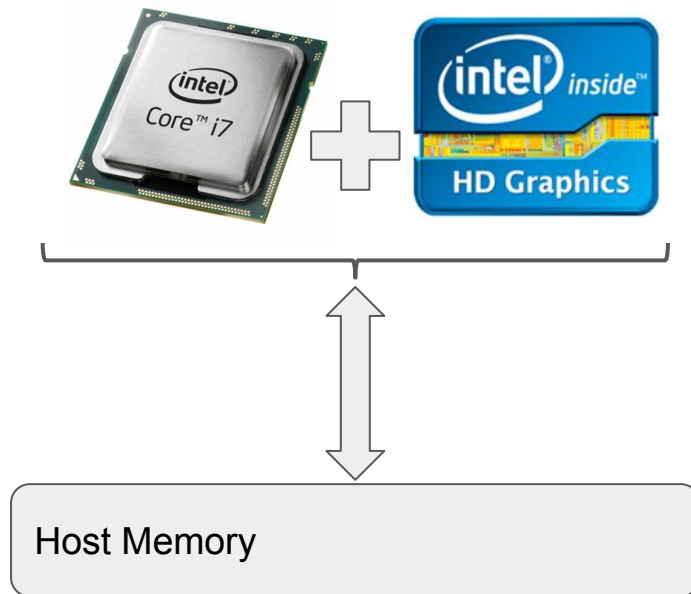
Why Intel?



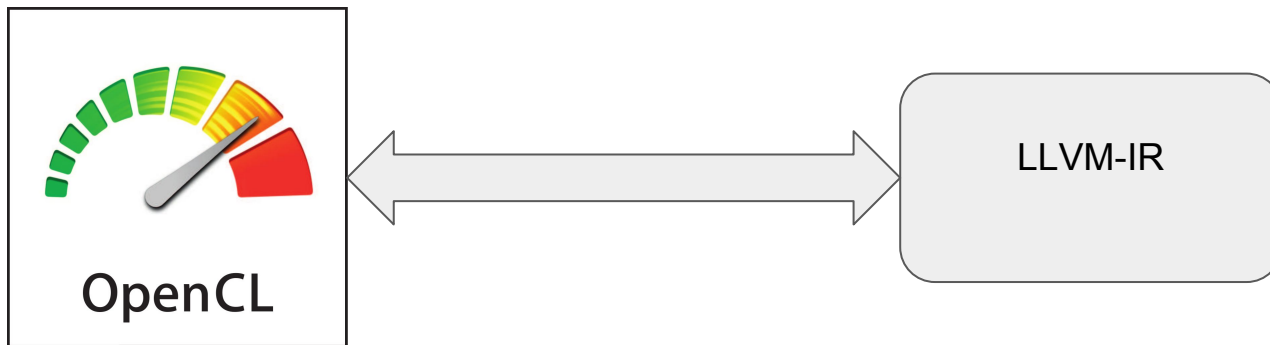
Why Intel?



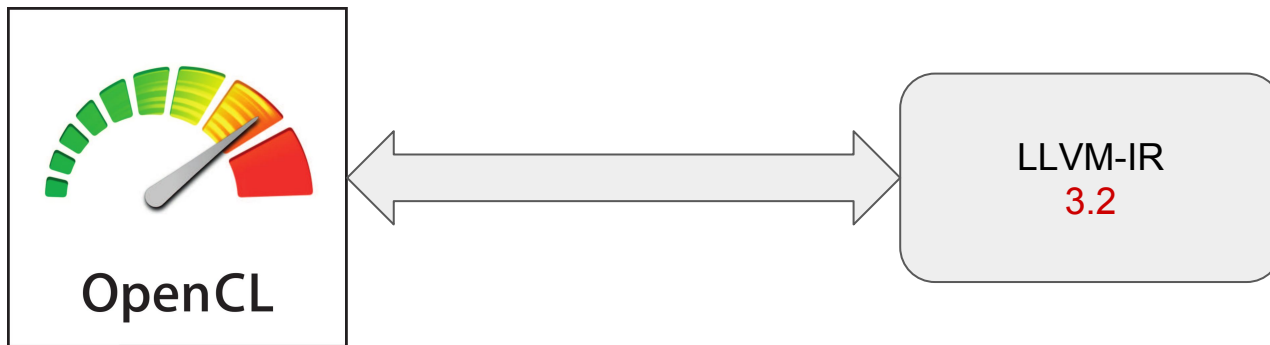
Unified Memory



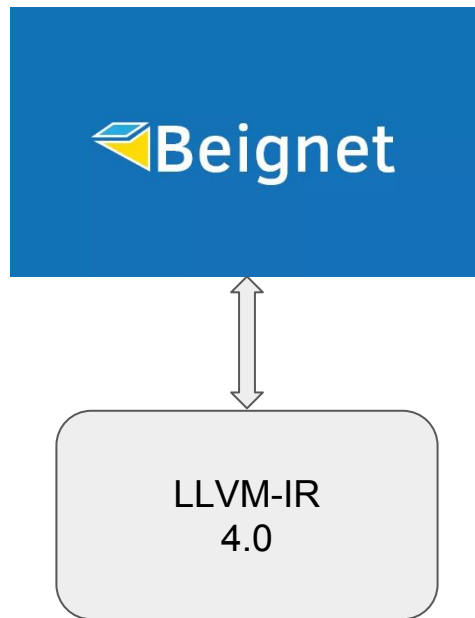
Generating Intel GPU Code



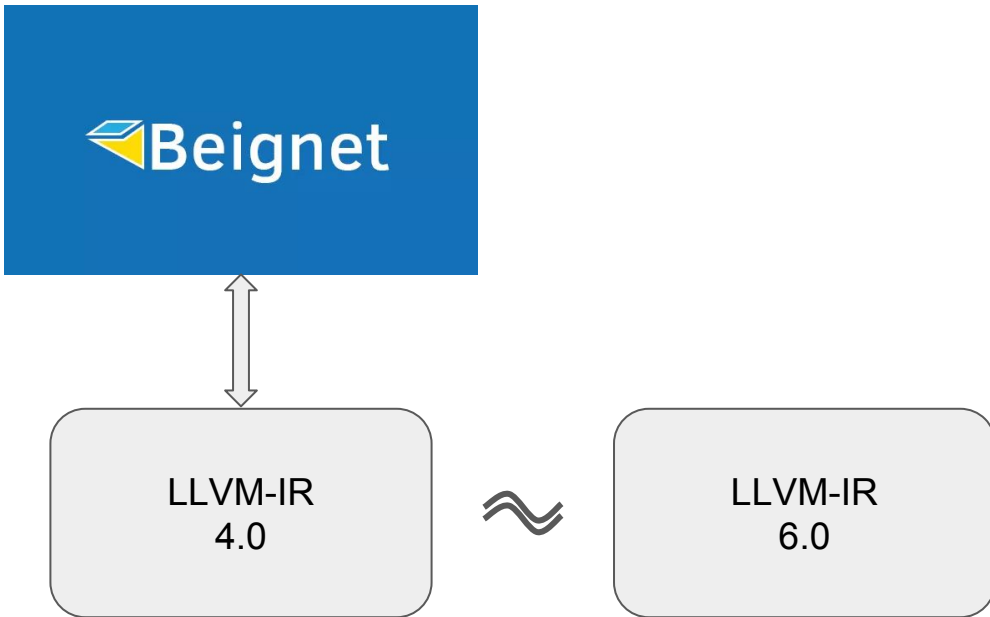
Generating Intel GPU Code



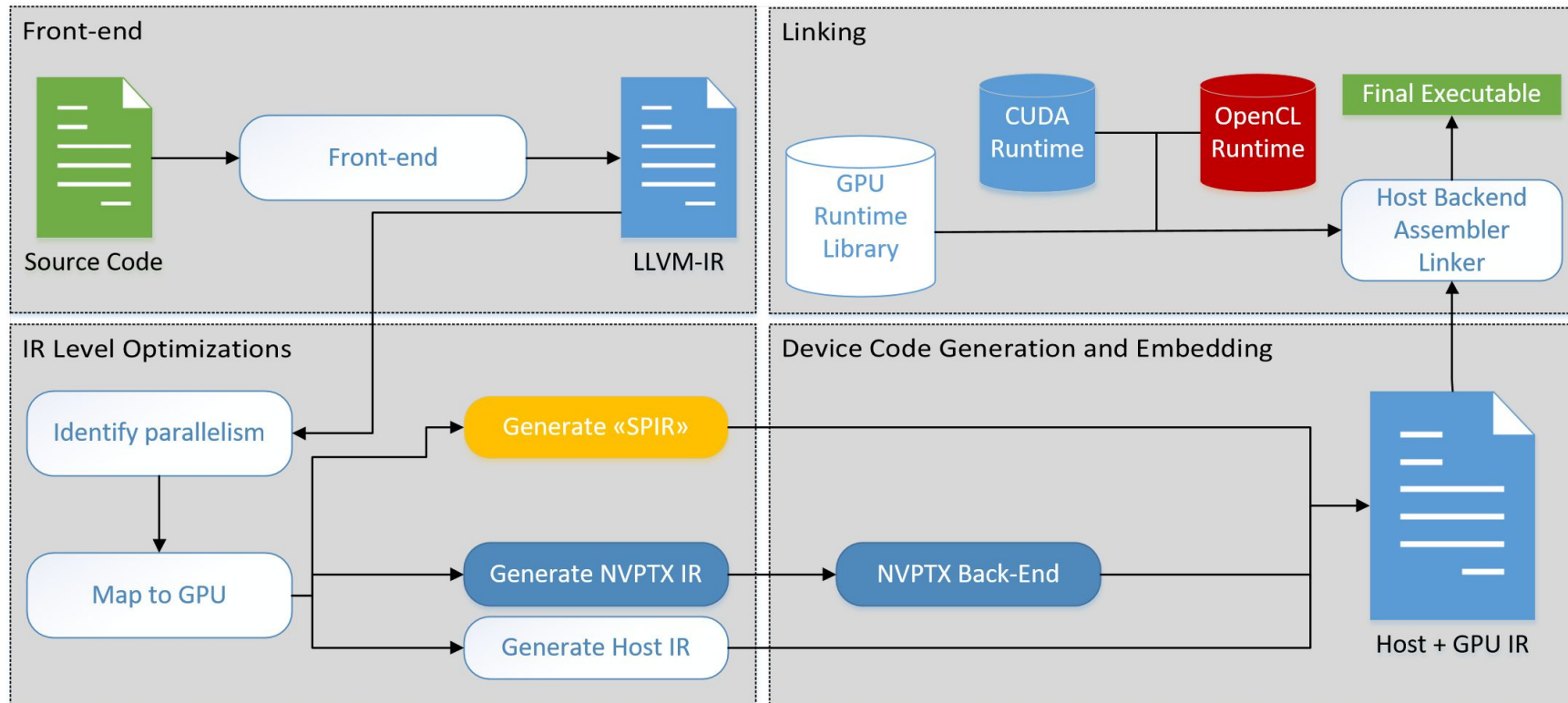
Running Intel GPU Code



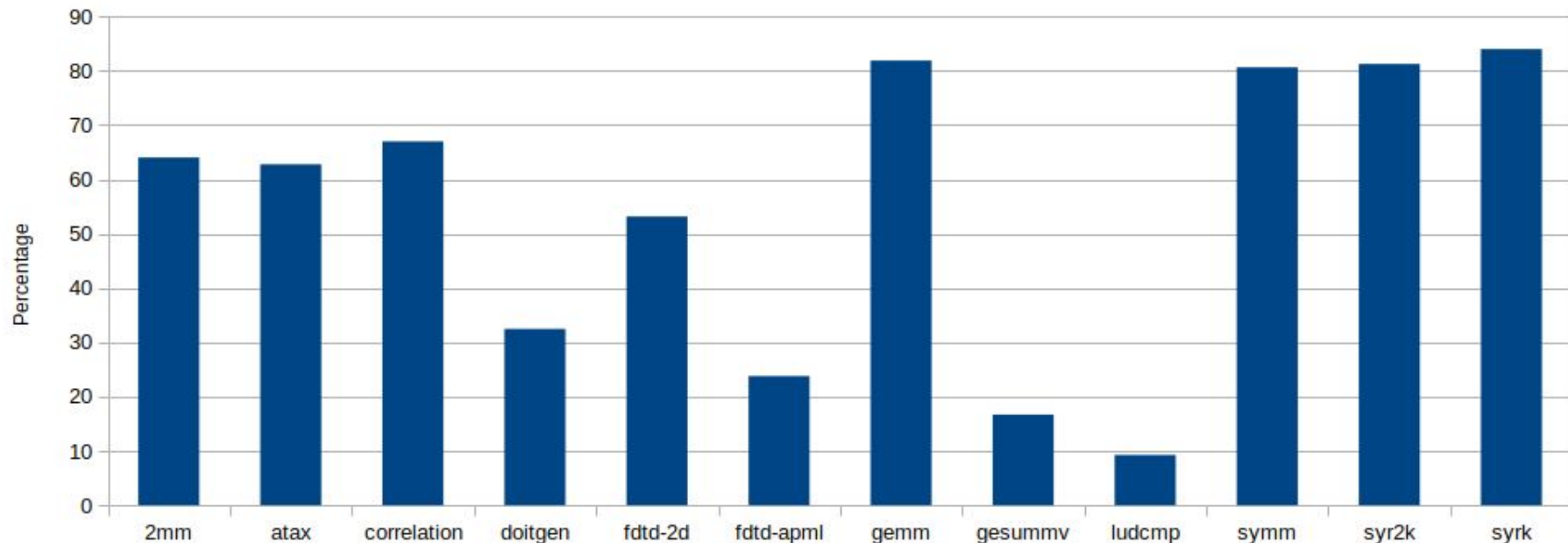
Running Intel GPU Code



Intel Through SPIR

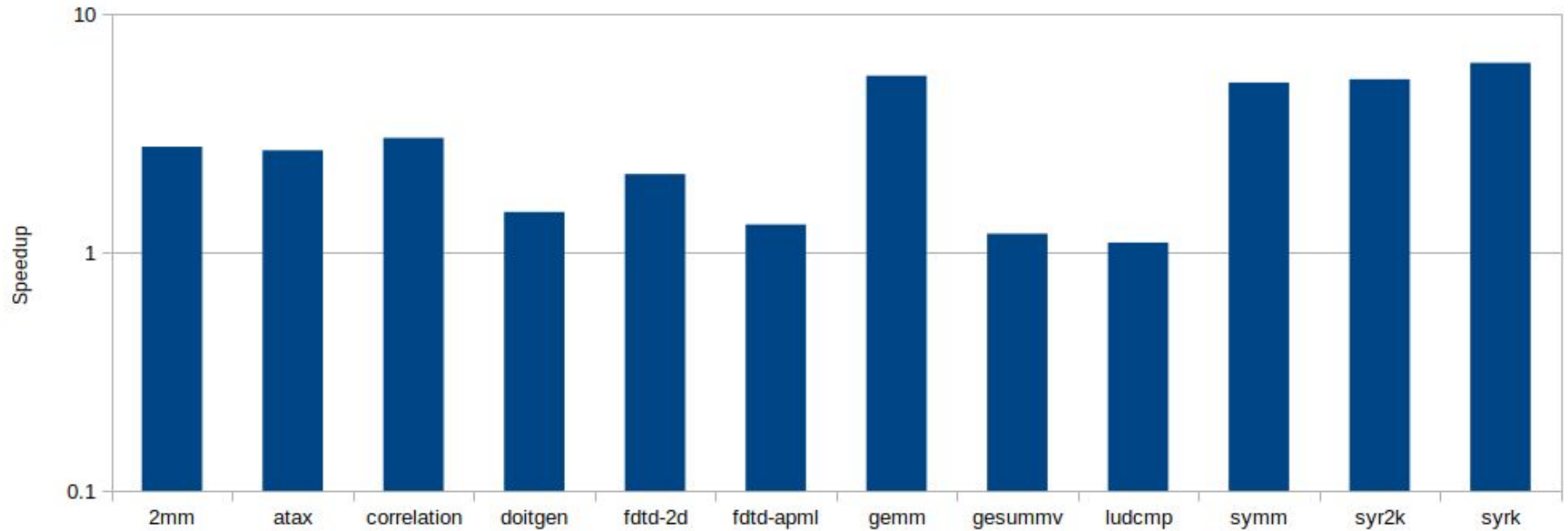


Percentage of Time Spent Copying Data



Geom. mean: 46%

This Is 'Why Intel'!

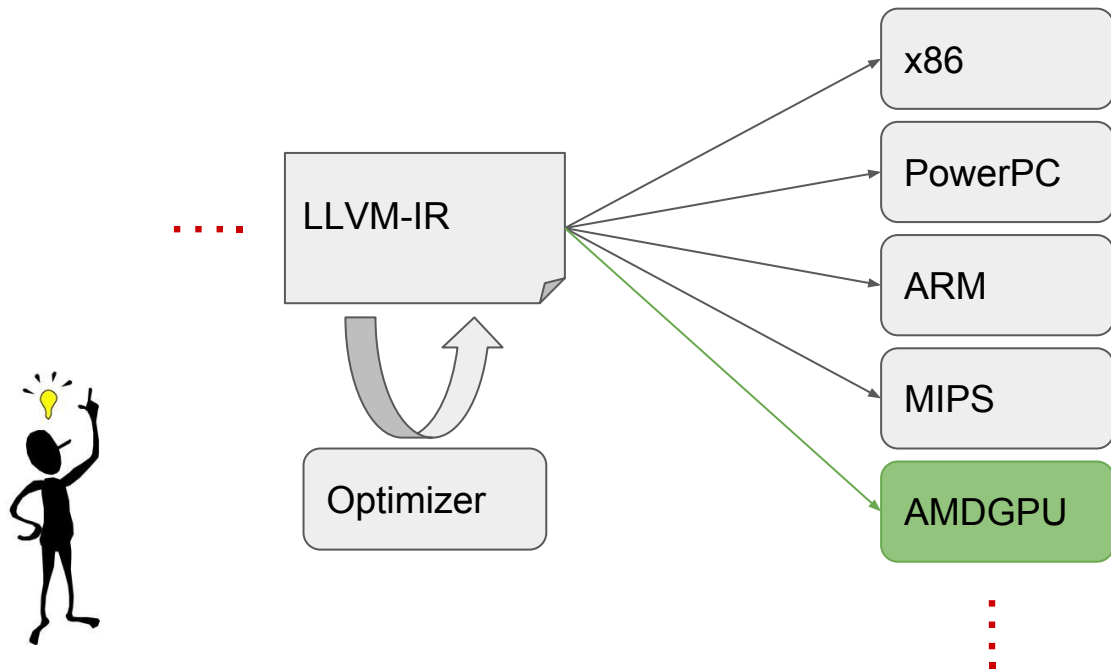


Geom. mean: 3.2x

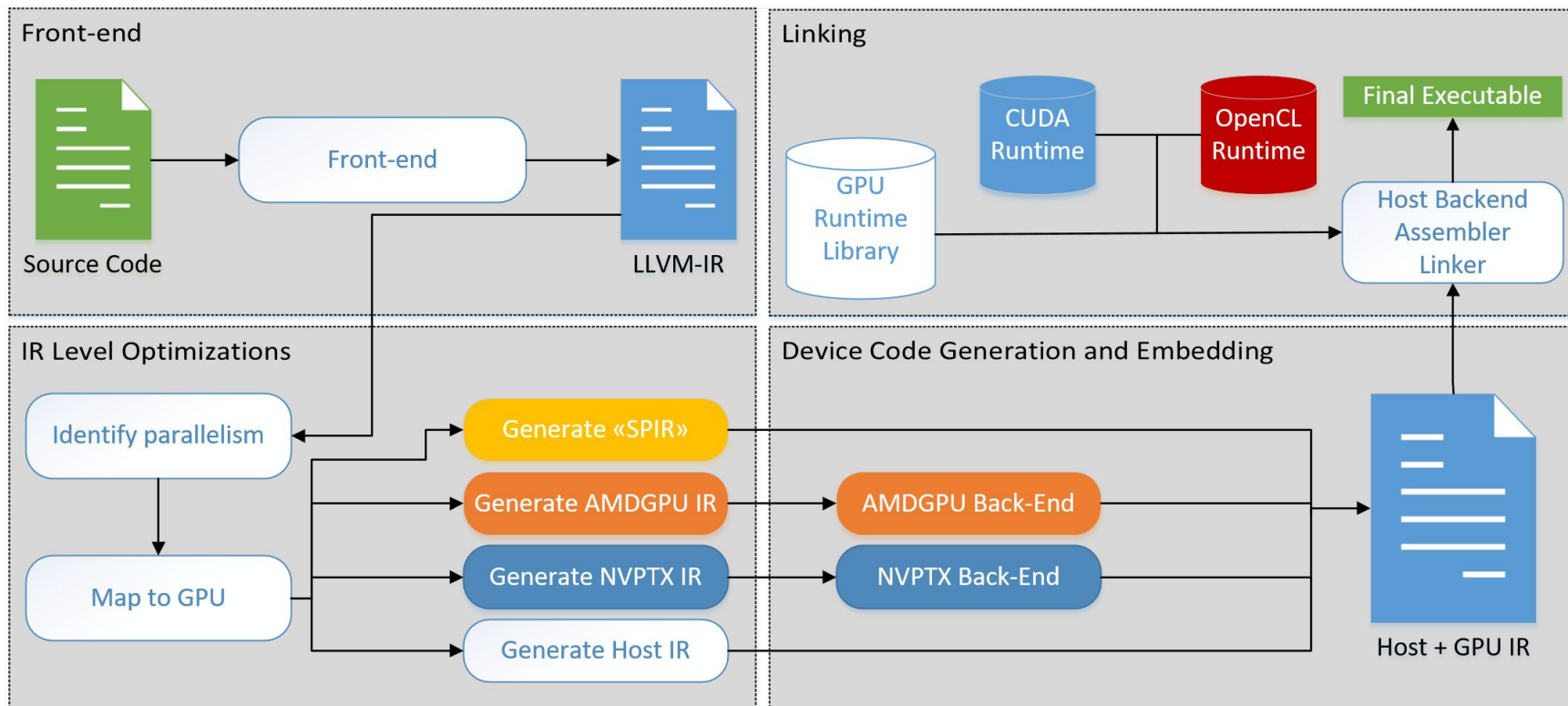
AMD Code Generation



AMD Code Generation



AMD Code Generation

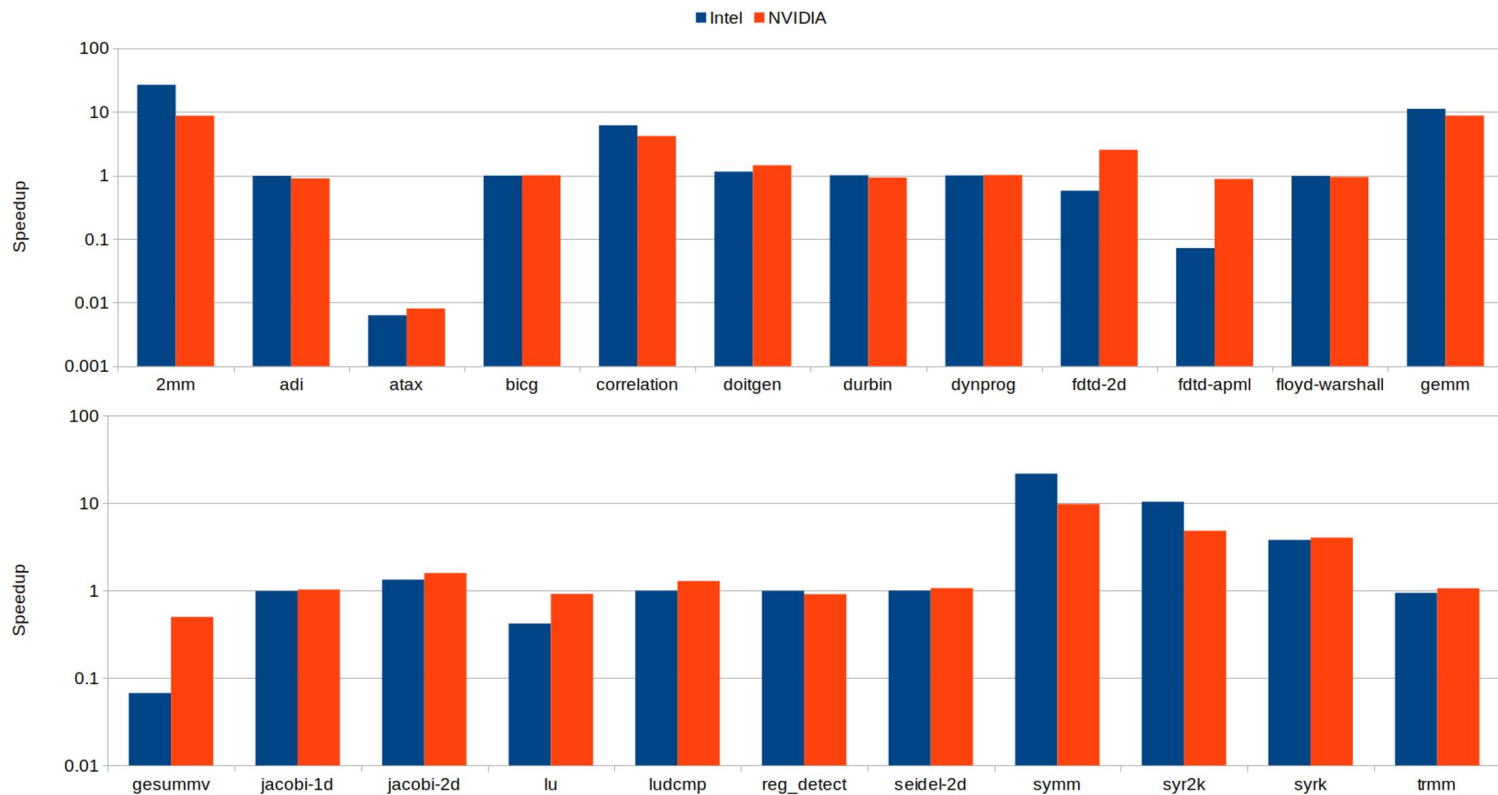


Running AMD ISA

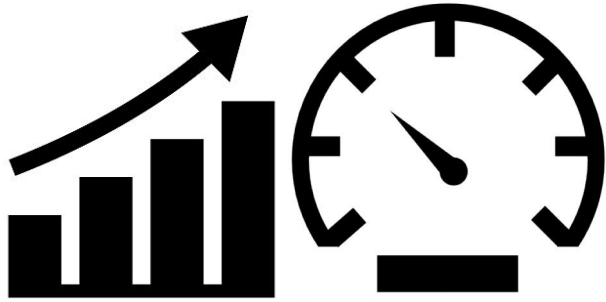


Radeon Open Compute

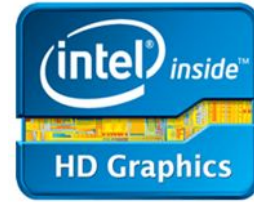
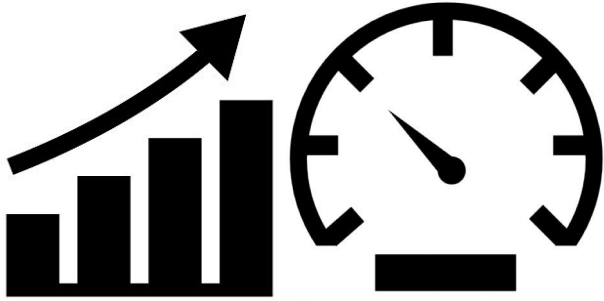
Runtime Improvement by Platform over -O3



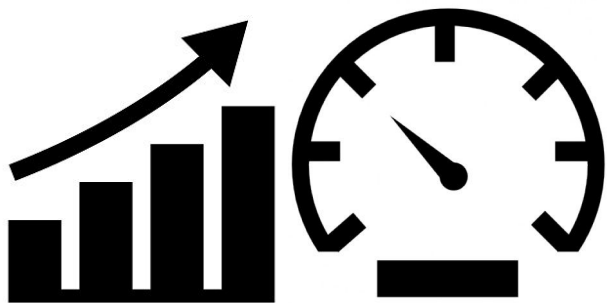
What can we do now?



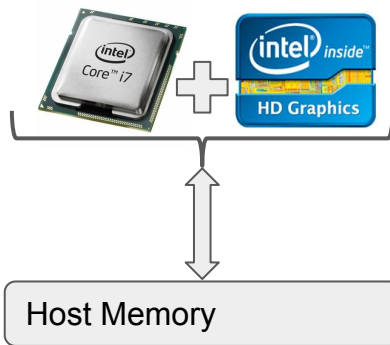
What can we do now?



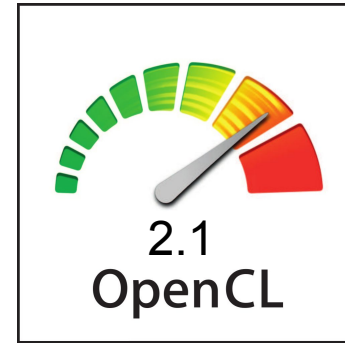
What can we do now?



Unified Memory



SPIR-V



We're One Step Closer

