OMPRTF

OpenMP RunTime Fabric

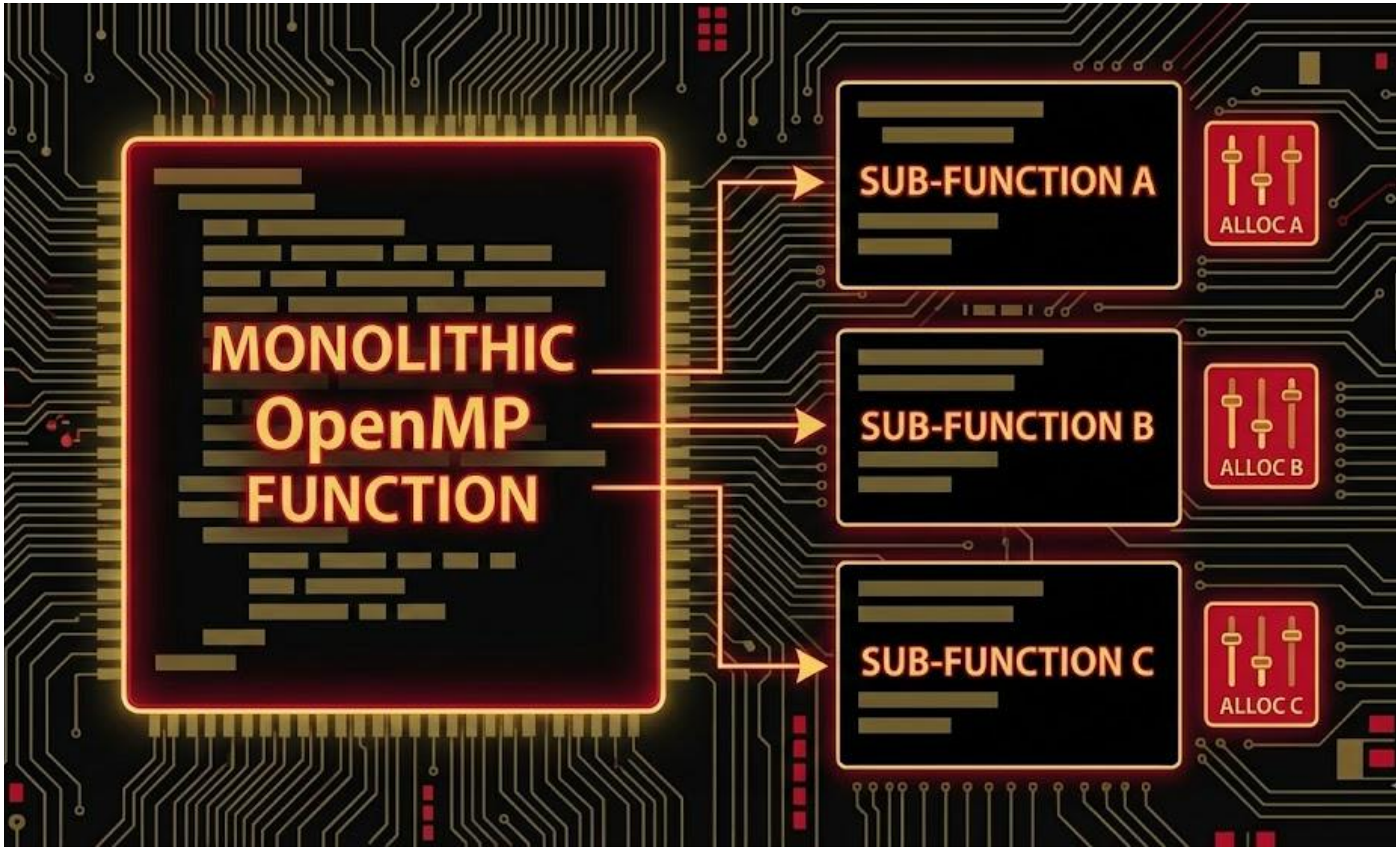
Brady Bangasser

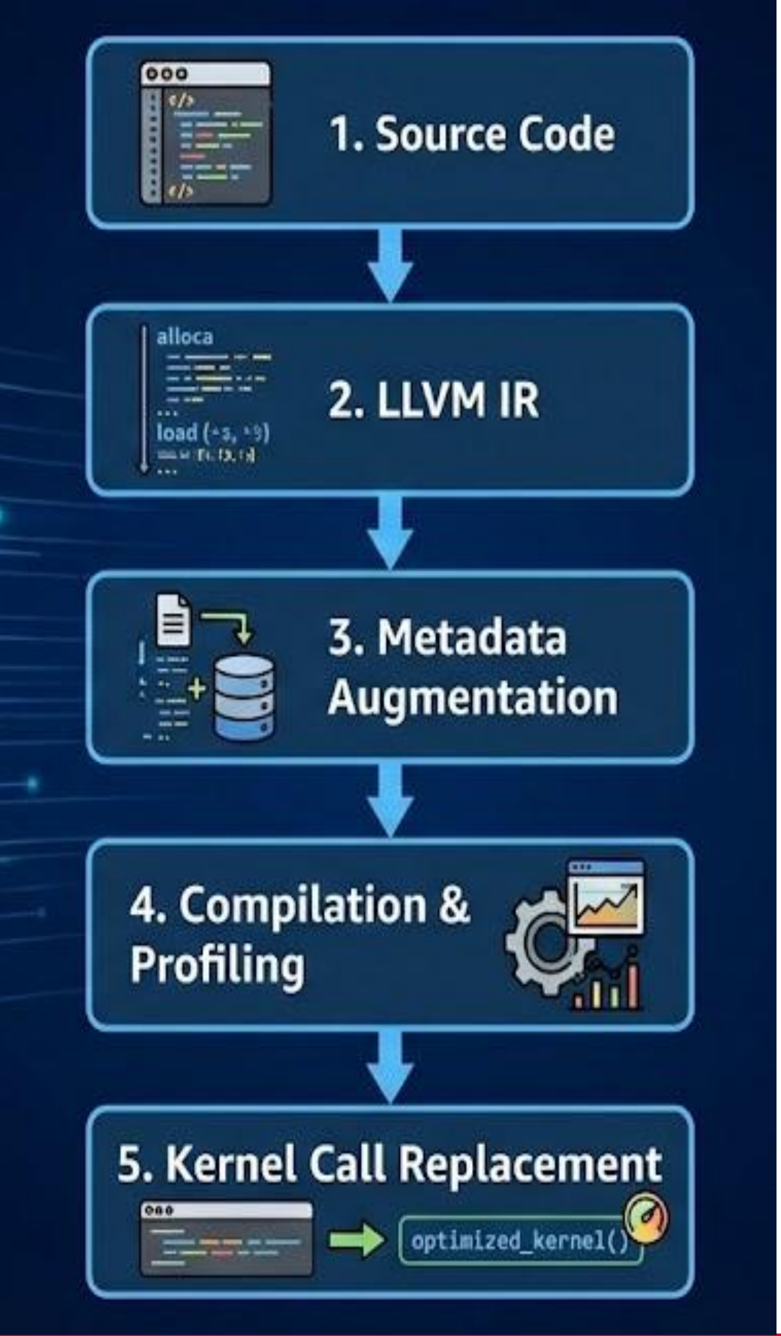
IOWA STATE UNIVERSITY

The Problem

- NVIDIA shipped almost 4 million GPUs to data centers in 2023
- Tools like OpenMP have increased access to GPU Offloading
- The main bottleneck in GPU based computation is the transfer speeds
- OpenMP uses Monolithic functions to allocate memory, copy data, and run the kernel
- Generally these transfers are plagued with problems such as duplicate data transfer

The Solution





The Process

Using augmented metadata, the profiler will flag problematic Open MP Offload calls. These calls are then replaced by finer tuned kernel calls, which are able to be optimized much further than the monolithic call we began with.

Why LLVM?

LLVM was chosen for it's ease of in memory representation and modification, as well as it's speedy compile time. Though it does introduce significant issues when it comes to input files.

Users need to provide a .ll IR file, which can create challenges for multi-file projects, as there is no well maintained Machine Code to LLVM IR