HOW NOT TO PROGRAM YOUR GPUS.

MUHAMMAD OSAMA

This seminar is a direct result of my dissertation work titled “Load Balancing on the GPU,” a GPU-based load-balancing programming model for fine-grained computations. This talk is intended as a tutorial, in which I will present lessons learned on how *not* to program a GPU. I will focus on three key aspects: performance, program abstractions and software engineering. Together we will build the fastest General Matrix-Multiplication (GEMM) kernel, ever.