Determinism and Low-Latency GPU Scheduling in OpenCL

Balázs Keszthelyi[1]

[1] V-Nova Ltd., London, United Kingdom

Given today's successful GPU compute applications are primarily focusing on high-throughput massively sized workloads, latency and determinism requirements are things that are mostly addressable via sacrificing compute headroom. In this case-study, I am going to present some of the common pitfalls of trying to tame a throughput-oriented GPU architecture in a way to deliver deterministic output performance on a number of parallel processing threads, and also some techniques relying only merely on standard OpenCL 1.x features.