GPU Solutions for PSCAD: IT17112

Reporting Period	June 18, 2020 - June 25, 2020
Activities	• Profiled GPU use during the execution of `QRFactor` on the Quadro RTX 3000 card • During execution, there were three primary kernel calls to the GPU that were of particular importance: (a) "csrqr_leftLooking_cta_byLevels_core", called once during the factoring of the matrix A into the product QR, (b) "csrqr_solve_Qtb_cta_core", called to compute $Q^{-1}\mathbf{b}(t)$ at each time step, and (c) "csrqr_upper_direct_kernel", called to solve $R\mathbf{x}(t) = Q^{-1}\mathbf{b}(t)$ for $\mathbf{x}(t)$ at each time step • Using NVIDIA Compute profiling tools, each kernel can be examined in great detail. See figure 1 for a sample profile page • Extracted Total GPU Utilization and Achieved Occupancy values from each kernel. See table 1 for values • High values for Achieved Occupancy signal efficient use of available warps in the Streaming Multiprocessor (SM) • While total GPU Utilization never exceeded 66%, this figure can be misleading. In the Compute Workload Analysis portion of figure 1 we see a breakdown of utilization of the each type of resource that the SM is capable of using. Since the Quadro RTX 3000 is specialized towards graphics rendering, there are fewer resources available at the hardware level for the double-precision operations required by `QRFactor`. Depending on the choice of hardware, more FP64 resources may be available, leading to a lower total GPU Utilization.
Issues	• None
Milestones Accomplished	• Profiling of `QRFactor` on Quadro RTX 3000
Milestones Not Accomplished	• None
Next Week's Milestones	• Examine effects of scaling the size of the problem on P100 and V100 hardware
Forwarded Issues	• None

Kernel	**Total GPU Utilization %**	**Achieved Occupancy %**
csrqr_leftLooking_cta_byLevels_core	65.7	92.7
csrqr_solve_Qtb_cta_core	14.5	99.2
csrqr_upper_direct_kernel	40.1	75.7

Table 1: Profiling information for the three major kernel calls in `QRFactor`.

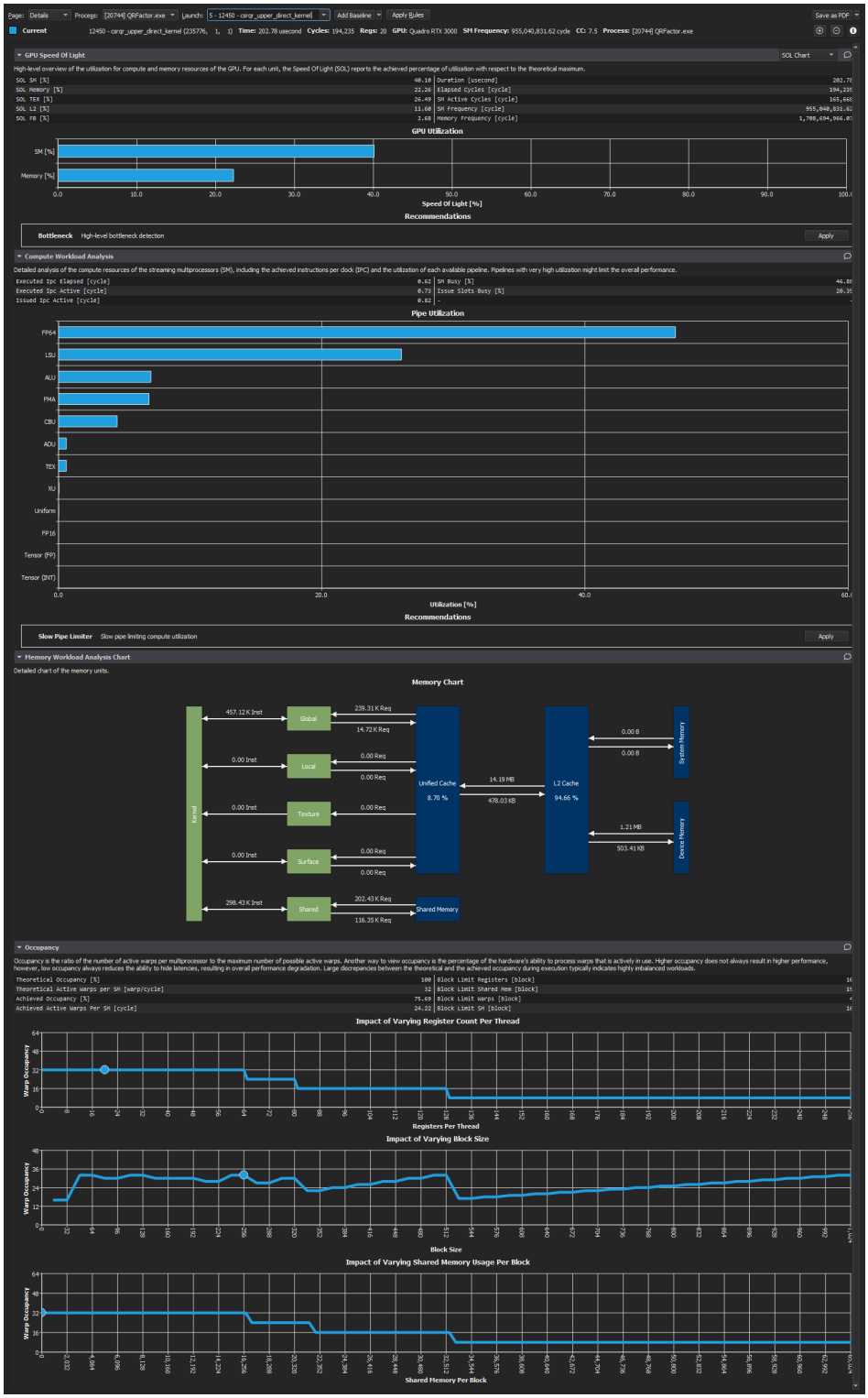

Figure 1: An example of the full profile for a kernel call using NVIDIA Compute.