Page: Details | Process: [20744] QRFactor.exe | Launch: 5 - 12450 - csrqr_upper_direct_kernel | Add Baseline | Apply Rules | Save as PDF

Current 12450 - csrqr_upper_direct_kernel (235776, 1, 1) **Time:** 202.78 usecond **Cycles:** 194,235 **Regs:** 20 **GPU:** Quadro RTX 3000 **SM Frequency:** 955,040,831.62 cycle **CC:** 7.5 **Process:** [20744] QRFactor.exe

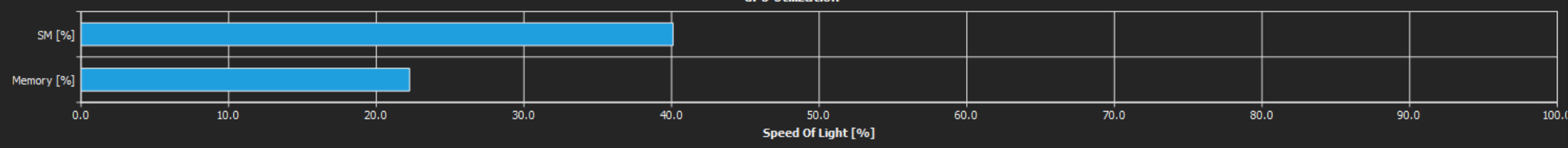
GPU Speed Of Light

SOL Chart

High-level overview of the utilization for compute and memory resources of the GPU. For each unit, the Speed Of Light (SOL) reports the achieved percentage of utilization with respect to the theoretical maximum.

SOL SM [%]	40.10	Duration [usecond]	202.78
SOL Memory [%]	22.26	Elapsed Cycles [cycle]	194,235
SOL TEX [%]	26.49	SM Active Cycles [cycle]	165,668
SOL L2 [%]	11.60	SM Frequency [cycle]	955,040,831.62
SOL FB [%]	2.68	Memory Frequency [cycle]	1,708,694,966.07

GPU Utilization



Recommendations

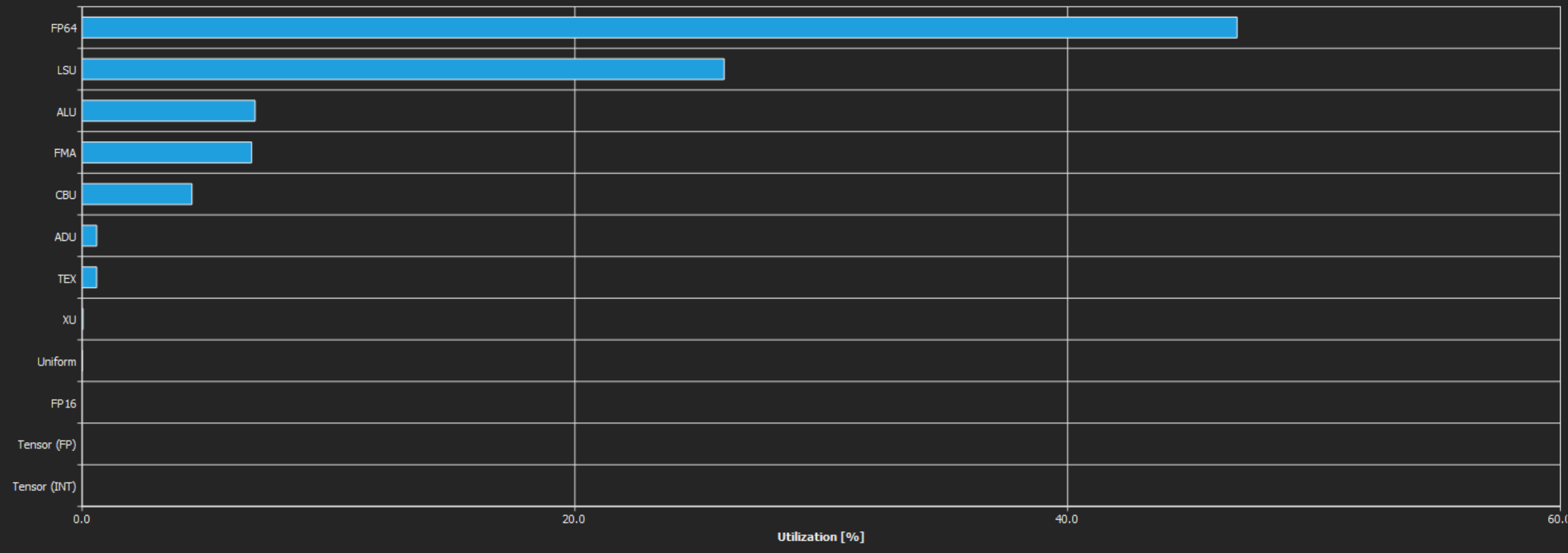
Bottleneck High-level bottleneck detection — Apply

Compute Workload Analysis

Detailed analysis of the compute resources of the streaming multiprocessors (SM), including the achieved instructions per clock (IPC) and the utilization of each available pipeline. Pipelines with very high utilization might limit the overall performance.

Executed Ipc Elapsed [cycle]	0.62	SM Busy [%]	46.88
Executed Ipc Active [cycle]	0.73	Issue Slots Busy [%]	20.39
Issued Ipc Active [cycle]	0.82	-	-

Pipe Utilization



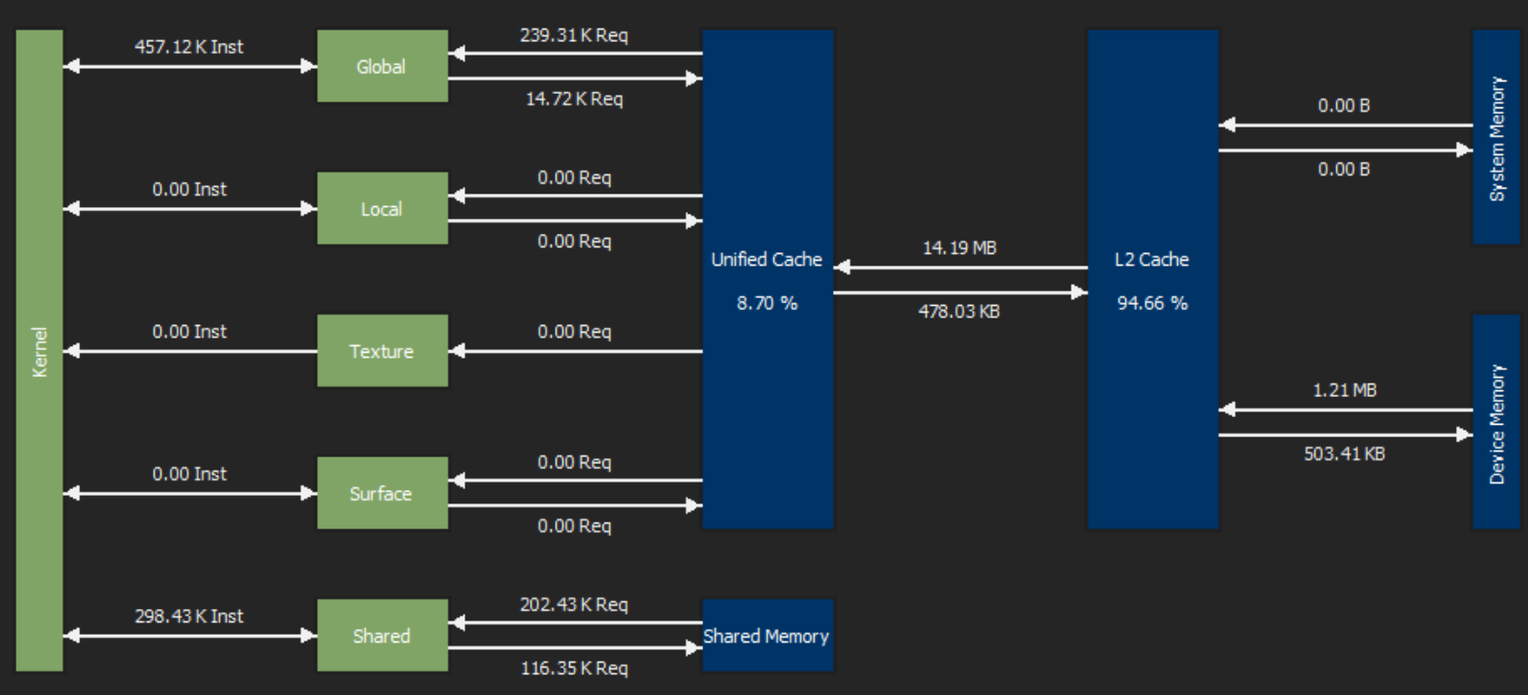
Recommendations

Slow Pipe Limiter Slow pipe limiting compute utilization — Apply

Memory Workload Analysis Chart

Detailed chart of the memory units.

Memory Chart

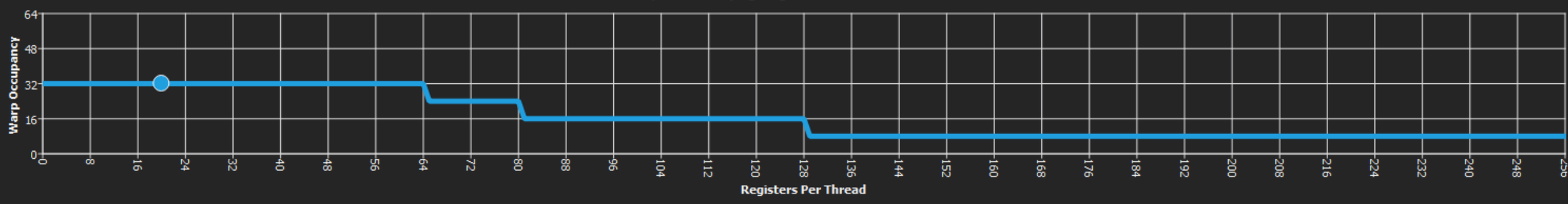


Occupancy

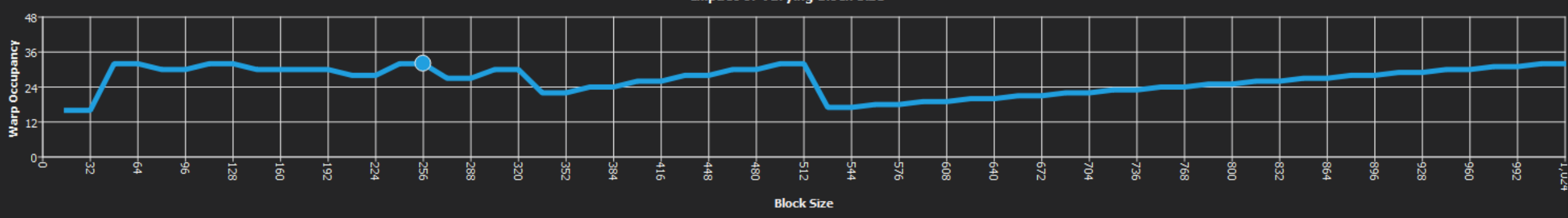
Occupancy is the ratio of the number of active warps per multiprocessor to the maximum number of possible active warps. Another way to view occupancy is the percentage of the hardware's ability to process warps that is actively in use. Higher occupancy does not always result in higher performance, however, low occupancy always reduces the ability to hide latencies, resulting in overall performance degradation. Large discrepancies between the theoretical and the achieved occupancy during execution typically indicates highly imbalanced workloads.

Theoretical Occupancy [%]	100	Block Limit Registers [block]	16
Theoretical Active Warps per SM [warp/cycle]	32	Block Limit Shared Mem [block]	15
Achieved Occupancy [%]	75.69	Block Limit Warps [block]	4
Achieved Active Warps Per SM [cycle]	24.22	Block Limit SM [block]	16

Impact of Varying Register Count Per Thread



Impact of Varying Block Size



Impact of Varying Shared Memory Usage Per Block