

Private to every GPU

Private to every SM

Private to every processing block

64 KiB registers

16 KiB L0 instruction cache

96 KiB L1 data cache/shared memory

2 KiB L1 constant cache

~46 KiB L1.5 constant cache/L1 instruction cache

4096 KiB L2 data cache/L2 constant cache/L2 instruction cache

~16 GB GDDR6 global memory

