Current	Result	Time	Cycles	GPU	Dispatch	Process	Attributes
1 fn gemm_gemm_kernel	1.20 ms (100.00 %)	0.29 ms	0.29 ms	0.29 ms	1	Metal Application [97] 60443 capturing GPU	1 ALU Limiter

Summary | Details | Metal Validation | Counters | Shaders | Bandwidth | Memory Size | Dependencies | Performance State | Compare | Hide | Export

Kernel Execution

GPU Speed of Light Throughput

Speed of Light provides a high-level overview of compute and memory resources of the GPU. For each unit, the throughput reports the achieved percentage of utilization with respect to the theoretical maximum. Breakdowns show the throughput for each individual sub-metric of Compute and Memory to clearly identify the highest contributor. High level overview of the utilization for compute and memory resources of the GPU presented as a roofline chart.

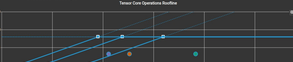
Compute SOL Throughput	64.04 %
Memory Throughput [%]	64.04 %
L1/TEX Cache Throughput [%]	64.04 %
L2 Cache Throughput [%]	64.04 %
DRAM Throughput [%]	2.44 %
High Memory Throughput	Memory is more heavily utilized than Compute. Look at the Memory Workload Analysis section to identify the L1 bottleneck. Check memory replay (coalescing) metrics to make sure you're efficiently utilizing the data transferred to/from device memory. Consider improving cache hit rates or moving to shared memory.

▸ Key Performance Indicators

Memory Matrix	1	Compute Matrix	1
gm_compute_memory_throughput avg.pct_of_peak_sustained_elapsed	64.04 %	Duration	10.560 us

Roofline Analysis

The roofline chart plots Peak GFlop/s (vs. Peak GFlop/s) of the kernel. Roofline chart shows the relation between achieved performance and the theoretical limits of the device. Roofline chart helps to identify if the kernel is limited by memory bandwidth or by compute. Hovering over the chart shows the exact values for the selected kernel. Click on a kernel in the table below to see its performance in the roofline chart. Roofline chart is not available if the profiled data is incomplete or if the kernel's performance is limited by latency. Roofline charts are not available for Metal 3 and Metal 4 kernels.



Tensor Core Operations Roofline

Arithmetic Intensity [OP/Byte]	Achieved Performance [TOP/s]
0x1,0x1,0x1,