Device

```
=== Device Information ===
Available Devices:
Device 0: "NVIDIA GeForce RTX 4050 Laptop GPU"
Selected Device: NVIDIA GeForce RTX 4050 Laptop GPU
Selected Device ID: 0
Compute Capability: 8.9
SMs: 20
Device Global Memory: 6140 MiB
Shared Memory per SM: 100 KiB
Memory Bus Width: 96 bits (ECC disabled)
Application Compute Clock Rate: 2.355 GHz
Application Memory Clock Rate: 7.825 GHz
```

Output Details

- `Latency`: refers to the [min, max, mean, median, 99% percentile] of the engine latency measurements, when timing the engine w/o profiling layers.
- `Throughput`: is measured in query (inference) per second (QPS).
- `Enqueue Time`: Time taken to enqueue inference requests.
- `H2D Latency`: Host-to-Device latency (data transfer time).
- `GPU Compute Time`: Time spent computing on the GPU.
- `D2H Latency`: Device-to-Host latency (data transfer time).
- `Total Host Walltime`
- `Total GPU Compute Time`

YOLOv9-C QAT (Mish)

Precision: FP32+INT8

Batch Size 1

```
=== Performance summary ===
Throughput: 330.117 qps

Latency: min = 3.01257 ms, max = 3.72827 ms, mean = 3.0276 ms, median = 3.02588
ms, percentile(90%) = 3.03418 ms, percentile(95%) = 3.03613 ms, percentile(99%) =
3.04053 ms

Enqueue Time: min = 0.0195312 ms, max = 0.615234 ms, mean = 0.0454716 ms, median =
0.0340576 ms, percentile(90%) = 0.0766602 ms, percentile(95%) = 0.09375 ms,
percentile(99%) = 0.19043 ms

H2D Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms
```

```
GPU Compute Time: min = 3.01257 ms, max = 3.72827 ms, mean = 3.0276 ms, median =
3.02588 ms, percentile(90%) = 3.03418 ms, percentile(95%) = 3.03613 ms,
percentile(99%) = 3.04053 ms

D2H Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

Total Host Walltime: 10.0055 s

Total GPU Compute Time: 10.0002 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 91.3616 qps

Latency: min = 10.9025 ms, max = 11.0664 ms, mean = 10.9438 ms, median = 10.9443
ms, percentile(90%) = 10.9556 ms, percentile(95%) = 10.958 ms, percentile(99%) =
10.9629 ms

Enqueue Time: min = 0.020752 ms, max = 0.532959 ms, mean = 0.0580063 ms, median =
0.050293 ms, percentile(90%) = 0.101562 ms, percentile(95%) = 0.11377 ms,
percentile(99%) = 0.177246 ms

H2D Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

GPU Compute Time: min = 10.9025 ms, max = 11.0664 ms, mean = 10.9438 ms, median =
10.9443 ms, percentile(90%) = 10.9556 ms, percentile(95%) = 10.958 ms,
percentile(99%) = 10.9629 ms

D2H Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

Total Host Walltime: 10.0152 s

Total GPU Compute Time: 10.0136 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 48.9507 qps

Latency: min = 20.3643 ms, max = 21.1794 ms, mean = 20.4271 ms, median = 20.4258
ms, percentile(90%) = 20.437 ms, percentile(95%) = 20.4409 ms, percentile(99%) =
20.4473 ms

Enqueue Time: min = 0.0214844 ms, max = 0.42627 ms, mean = 0.075538 ms, median =
```

```
0.0620117 ms, percentile(90%) = 0.117676 ms, percentile(95%) = 0.127441 ms,
percentile(99%) = 0.233887 ms

H2D Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

GPU Compute Time: min = 20.3643 ms, max = 21.1794 ms, mean = 20.4271 ms, median =
20.4258 ms, percentile(90%) = 20.437 ms, percentile(95%) = 20.4409 ms,
percentile(99%) = 20.4473 ms

D2H Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

Total Host Walltime: 10.0305 s

Total GPU Compute Time: 10.0297 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 32.2187 qps

Latency: min = 30.9883 ms, max = 31.0742 ms, mean = 31.0362 ms, median = 31.0371
ms, percentile(90%) = 31.0498 ms, percentile(95%) = 31.0537 ms, percentile(99%) =
31.063 ms

Enqueue Time: min = 0.0253906 ms, max = 0.40625 ms, mean = 0.0721937 ms, median =
0.0598145 ms, percentile(90%) = 0.114075 ms, percentile(95%) = 0.131348 ms,
percentile(99%) = 0.256836 ms

H2D Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

GPU Compute Time: min = 30.9883 ms, max = 31.0742 ms, mean = 31.0362 ms, median =
31.0371 ms, percentile(90%) = 31.0498 ms, percentile(95%) = 31.0537 ms,
percentile(99%) = 31.063 ms

D2H Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

Total Host Walltime: 10.0563 s

Total GPU Compute Time: 10.0557 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 24.0033 qps
```

```
Latency: min = 41.6174 ms, max = 41.7002 ms, mean = 41.6591 ms, median = 41.6594
ms, percentile(90%) = 41.6748 ms, percentile(95%) = 41.679 ms, percentile(99%) =
41.6892 ms

Enqueue Time: min = 0.0283203 ms, max = 1.18127 ms, mean = 0.0669195 ms, median =
0.057312 ms, percentile(90%) = 0.0908203 ms, percentile(95%) = 0.109375 ms,
percentile(99%) = 0.132812 ms

H2D Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

GPU Compute Time: min = 41.6174 ms, max = 41.7002 ms, mean = 41.6591 ms, median =
41.6594 ms, percentile(90%) = 41.6748 ms, percentile(95%) = 41.679 ms,
percentile(99%) = 41.6892 ms

D2H Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

Total Host Walltime: 10.0819 s

Total GPU Compute Time: 10.0815 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 48.8766 qps

Latency: min = 20.4316 ms, max = 20.4822 ms, mean = 20.4578 ms, median = 20.458
ms, percentile(90%) = 20.4697 ms, percentile(95%) = 20.4736 ms, percentile(99%) =
20.4795 ms

Enqueue Time: min = 0.0227051 ms, max = 0.307617 ms, mean = 0.0747259 ms, median =
0.0644531 ms, percentile(90%) = 0.120117 ms, percentile(95%) = 0.136719 ms,
percentile(99%) = 0.188965 ms

H2D Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

GPU Compute Time: min = 20.4316 ms, max = 20.4822 ms, mean = 20.4578 ms, median =
20.458 ms, percentile(90%) = 20.4697 ms, percentile(95%) = 20.4736 ms,
percentile(99%) = 20.4795 ms

D2H Latency: min = 0 ms, max = 0 ms, mean = 0 ms, median = 0 ms, percentile(90%) =
0 ms, percentile(95%) = 0 ms, percentile(99%) = 0 ms

Total Host Walltime: 10.0457 s

Total GPU Compute Time: 10.0448 s
```