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+FP16

Batch Size 1

```
=== Performance summary ===
Throughput: 348.531 qps

Latency: min = 2.85486 ms, max = 3.74609 ms, mean = 2.86726 ms, median = 2.86719
ms, percentile(90%) = 2.87012 ms, percentile(95%) = 2.87109 ms, percentile(99%) =
2.87256 ms

Enqueue Time: min = 0.0185547 ms, max = 0.547607 ms, mean = 0.0413729 ms, median =
0.027832 ms, percentile(90%) = 0.0722656 ms, percentile(95%) = 0.0888672 ms,
percentile(99%) = 0.185547 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 = 2.85486 ms, max = 3.74609 ms, mean = 2.86726 ms, median =
2.86719 ms, percentile(90%) = 2.87012 ms, percentile(95%) = 2.87109 ms,
percentile(99%) = 2.87256 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: 10.0077 s

Total GPU Compute Time: 10.001 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 100.731 qps

Latency: min = 9.87646 ms, max = 10.3865 ms, mean = 9.92575 ms, median = 9.92285
ms, percentile(90%) = 9.93262 ms, percentile(95%) = 9.93506 ms, percentile(99%) =
9.94336 ms

Enqueue Time: min = 0.0205078 ms, max = 0.816406 ms, mean = 0.0682629 ms, median =
0.0556641 ms, percentile(90%) = 0.10791 ms, percentile(95%) = 0.126587 ms,
percentile(99%) = 0.287109 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 = 9.87646 ms, max = 10.3865 ms, mean = 9.92575 ms, median =
9.92285 ms, percentile(90%) = 9.93262 ms, percentile(95%) = 9.93506 ms,
percentile(99%) = 9.94336 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: 10.0168 s

Total GPU Compute Time: 10.0151 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 48.9795 qps

Latency: min = 20.3469 ms, max = 20.4493 ms, mean = 20.4151 ms, median = 20.4155
ms, percentile(90%) = 20.4277 ms, percentile(95%) = 20.4316 ms, percentile(99%) =
20.4404 ms

Enqueue Time: min = 0.0214844 ms, max = 0.877686 ms, mean = 0.0798938 ms, median =
```

```
0.0661621 ms, percentile(90%) = 0.120117 ms, percentile(95%) = 0.139648 ms,
percentile(99%) = 0.259033 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 = 20.3469 ms, max = 20.4493 ms, mean = 20.4151 ms, median =
20.4155 ms, percentile(90%) = 20.4277 ms, percentile(95%) = 20.4316 ms,
percentile(99%) = 20.4404 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: 10.045 s

Total GPU Compute Time: 10.0442 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 32.2164 qps

Latency: min = 30.9985 ms, max = 31.0781 ms, mean = 31.0384 ms, median = 31.0385
ms, percentile(90%) = 31.0527 ms, percentile(95%) = 31.0566 ms, percentile(99%) =
31.0645 ms

Enqueue Time: min = 0.026123 ms, max = 1.74219 ms, mean = 0.092572 ms, median =
0.0689087 ms, percentile(90%) = 0.121948 ms, percentile(95%) = 0.154297 ms,
percentile(99%) = 0.342285 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 = 30.9985 ms, max = 31.0781 ms, mean = 31.0384 ms, median =
31.0385 ms, percentile(90%) = 31.0527 ms, percentile(95%) = 31.0566 ms,
percentile(99%) = 31.0645 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: 10.057 s

Total GPU Compute Time: 10.0564 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 24.0168 qps
```

```
Latency: min = 41.5017 ms, max = 42.3014 ms, mean = 41.6358 ms, median = 41.6362
ms, percentile(90%) = 41.6533 ms, percentile(95%) = 41.6572 ms, percentile(99%) =
41.6689 ms

Enqueue Time: min = 0.0336914 ms, max = 0.427734 ms, mean = 0.0749446 ms, median =
0.0615234 ms, percentile(90%) = 0.114746 ms, percentile(95%) = 0.125977 ms,
percentile(99%) = 0.236694 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 = 41.5017 ms, max = 42.3014 ms, mean = 41.6358 ms, median =
41.6362 ms, percentile(90%) = 41.6533 ms, percentile(95%) = 41.6572 ms,
percentile(99%) = 41.6689 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: 10.0763 s

Total GPU Compute Time: 10.0759 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 48.9537 qps

Latency: min = 20.3582 ms, max = 20.4644 ms, mean = 20.4257 ms, median = 20.4277
ms, percentile(90%) = 20.4424 ms, percentile(95%) = 20.4463 ms, percentile(99%) =
20.4546 ms

Enqueue Time: min = 0.0234375 ms, max = 0.961426 ms, mean = 0.0835144 ms, median =
0.0634766 ms, percentile(90%) = 0.123291 ms, percentile(95%) = 0.159912 ms,
percentile(99%) = 0.320312 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 = 20.3582 ms, max = 20.4644 ms, mean = 20.4257 ms, median =
20.4277 ms, percentile(90%) = 20.4424 ms, percentile(95%) = 20.4463 ms,
percentile(99%) = 20.4546 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: 10.0299 s

Total GPU Compute Time: 10.029 s
```