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 (SiLU)

Precision: FP32+INT8

Batch Size 1

```
=== Performance summary ===
Throughput: 346.988 qps

Latency: min = 2.8631 ms, max = 3.05664 ms, mean = 2.8803 ms, median = 2.87988 ms,
percentile(90%) = 2.88281 ms, percentile(95%) = 2.88379 ms, percentile(99%) =
2.88867 ms

Enqueue Time: min = 0.0175781 ms, max = 0.824951 ms, mean = 0.0448133 ms, median =
0.0302734 ms, percentile(90%) = 0.0758057 ms, percentile(95%) = 0.0932617 ms,
percentile(99%) = 0.209961 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 = 2.8631 ms, max = 3.05664 ms, mean = 2.8803 ms, median =
2.87988 ms, percentile(90%) = 2.88281 ms, percentile(95%) = 2.88379 ms,
percentile(99%) = 2.88867 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: 10.0032 s

Total GPU Compute Time: 9.99753 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 102.368 qps

Latency: min = 9.71674 ms, max = 10.2256 ms, mean = 9.76698 ms, median = 9.76416
ms, percentile(90%) = 9.77637 ms, percentile(95%) = 9.77832 ms, percentile(99%) =
9.78418 ms

Enqueue Time: min = 0.0175781 ms, max = 0.660278 ms, mean = 0.0552926 ms, median =
0.0473633 ms, percentile(90%) = 0.09375 ms, percentile(95%) = 0.106934 ms,
percentile(99%) = 0.186035 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 = 9.71674 ms, max = 10.2256 ms, mean = 9.76698 ms, median =
9.76416 ms, percentile(90%) = 9.77637 ms, percentile(95%) = 9.77832 ms,
percentile(99%) = 9.78418 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: 10.0129 s

Total GPU Compute Time: 10.0112 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 49.0846 qps

Latency: min = 20.2957 ms, max = 20.4023 ms, mean = 20.3712 ms, median = 20.375
ms, percentile(90%) = 20.3867 ms, percentile(95%) = 20.3889 ms, percentile(99%) =
20.3975 ms

Enqueue Time: min = 0.0236816 ms, max = 0.725586 ms, mean = 0.0821416 ms, median =
```

```
0.0622559 ms, percentile(90%) = 0.119141 ms, percentile(95%) = 0.147461 ms,
percentile(99%) = 0.457581 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 = 20.2957 ms, max = 20.4023 ms, mean = 20.3712 ms, median =
20.375 ms, percentile(90%) = 20.3867 ms, percentile(95%) = 20.3889 ms,
percentile(99%) = 20.3975 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: 10.0439 s

Total GPU Compute Time: 10.043 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 32.4005 qps

Latency: min = 30.7794 ms, max = 30.917 ms, mean = 30.862 ms, median = 30.8623 ms,
percentile(90%) = 30.8916 ms, percentile(95%) = 30.8992 ms, percentile(99%) =
30.9062 ms

Enqueue Time: min = 0.0253906 ms, max = 0.358398 ms, mean = 0.0745282 ms, median =
0.0585938 ms, percentile(90%) = 0.12207 ms, percentile(95%) = 0.138672 ms,
percentile(99%) = 0.251953 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 = 30.7794 ms, max = 30.917 ms, mean = 30.862 ms, median =
30.8623 ms, percentile(90%) = 30.8916 ms, percentile(95%) = 30.8992 ms,
percentile(99%) = 30.9062 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: 10.0616 s

Total GPU Compute Time: 10.061 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 24.0907 qps
```

```
Latency: min = 41.4126 ms, max = 41.5654 ms, mean = 41.5081 ms, median = 41.5068
ms, percentile(90%) = 41.54 ms, percentile(95%) = 41.5454 ms, percentile(99%) =
41.5537 ms

Enqueue Time: min = 0.0307617 ms, max = 0.371704 ms, mean = 0.0734341 ms, median =
0.0585938 ms, percentile(90%) = 0.112793 ms, percentile(95%) = 0.129883 ms,
percentile(99%) = 0.26416 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 = 41.4126 ms, max = 41.5654 ms, mean = 41.5081 ms, median =
41.5068 ms, percentile(90%) = 41.54 ms, percentile(95%) = 41.5454 ms,
percentile(99%) = 41.5537 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: 10.0869 s

Total GPU Compute Time: 10.0865 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 49.059 qps

Latency: min = 20.2968 ms, max = 21.1804 ms, mean = 20.3819 ms, median = 20.3796
ms, percentile(90%) = 20.3899 ms, percentile(95%) = 20.3945 ms, percentile(99%) =
20.4043 ms

Enqueue Time: min = 0.0219727 ms, max = 0.488647 ms, mean = 0.0767863 ms, median =
0.0605469 ms, percentile(90%) = 0.119141 ms, percentile(95%) = 0.132324 ms,
percentile(99%) = 0.34375 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 = 20.2968 ms, max = 21.1804 ms, mean = 20.3819 ms, median =
20.3796 ms, percentile(90%) = 20.3899 ms, percentile(95%) = 20.3945 ms,
percentile(99%) = 20.4043 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: 10.0287 s

Total GPU Compute Time: 10.0279 s
```