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 (AconC)

Precision: FP32+FP16

Batch Size 1

```
=== Performance summary ===
Throughput: 345.122 qps


Latency: min = 2.8877 ms, max = 2.9165 ms, mean = 2.89582 ms, median = 2.896 ms,
percentile(90%) = 2.89844 ms, percentile(95%) = 2.89893 ms, percentile(99%) =
2.90088 ms

Enqueue Time: min = 0.019043 ms, max = 0.441895 ms, mean = 0.0432563 ms, median =
0.0305176 ms, percentile(90%) = 0.078125 ms, percentile(95%) = 0.0908203 ms,
percentile(99%) = 0.140625 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 = 2.8877 ms, max = 2.9165 ms, mean = 2.89582 ms, median =
2.896 ms, percentile(90%) = 2.89844 ms, percentile(95%) = 2.89893 ms,
percentile(99%) = 2.90088 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: 10.0081 s

Total GPU Compute Time: 10.0022 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 106.755 qps

Latency: min = 9.33789 ms, max = 9.38477 ms, mean = 9.36561 ms, median = 9.37256
ms, percentile(90%) = 9.37891 ms, percentile(95%) = 9.38086 ms, percentile(99%) =
9.38379 ms

Enqueue Time: min = 0.0212402 ms, max = 0.545898 ms, mean = 0.0676716 ms, median =
0.0610352 ms, percentile(90%) = 0.11084 ms, percentile(95%) = 0.123047 ms,
percentile(99%) = 0.262207 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 = 9.33789 ms, max = 9.38477 ms, mean = 9.36561 ms, median =
9.37256 ms, percentile(90%) = 9.37891 ms, percentile(95%) = 9.38086 ms,
percentile(99%) = 9.38379 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: 10.0229 s

Total GPU Compute Time: 10.0212 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 51.9085 qps

Latency: min = 19.1949 ms, max = 20.0335 ms, mean = 19.263 ms, median = 19.2622
ms, percentile(90%) = 19.2715 ms, percentile(95%) = 19.2744 ms, percentile(99%) =
19.2793 ms
```

```
Enqueue Time: min = 0.0205078 ms, max = 0.459473 ms, mean = 0.0643192 ms, median = 0.0537109 ms, percentile(90%) = 0.10498 ms, percentile(95%) = 0.114746 ms, percentile(99%) = 0.165527 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 = 19.1949 ms, max = 20.0335 ms, mean = 19.263 ms, median = 19.2622 ms, percentile(90%) = 19.2715 ms, percentile(95%) = 19.2744 ms, percentile(99%) = 19.2793 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: 10.0369 s

Total GPU Compute Time: 10.036 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 34.1847 qps

Latency: min = 29.141 ms, max = 29.865 ms, mean = 29.2511 ms, median = 29.2539 ms, percentile(90%) = 29.2695 ms, percentile(95%) = 29.2725 ms, percentile(99%) = 29.2822 ms

Enqueue Time: min = 0.0266113 ms, max = 0.721436 ms, mean = 0.0897106 ms, median = 0.0673828 ms, percentile(90%) = 0.126282 ms, percentile(95%) = 0.151367 ms, percentile(99%) = 0.360352 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 = 29.141 ms, max = 29.865 ms, mean = 29.2511 ms, median = 29.2539 ms, percentile(90%) = 29.2695 ms, percentile(95%) = 29.2725 ms, percentile(99%) = 29.2822 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: 10.063 s

Total GPU Compute Time: 10.0624 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 25.4787 qps
```

```
Latency: min = 39.1208 ms, max = 39.3047 ms, mean = 39.2467 ms, median = 39.252 ms, percentile(90%) = 39.2754 ms, percentile(95%) = 39.2798 ms, percentile(99%) = 39.2896 ms

Enqueue Time: min = 0.0351562 ms, max = 0.381836 ms, mean = 0.079608 ms, median = 0.0615234 ms, percentile(90%) = 0.120117 ms, percentile(95%) = 0.131104 ms, percentile(99%) = 0.241333 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 = 39.1208 ms, max = 39.3047 ms, mean = 39.2467 ms, median = 39.252 ms, percentile(90%) = 39.2754 ms, percentile(95%) = 39.2798 ms, percentile(99%) = 39.2896 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: 10.0868 s

Total GPU Compute Time: 10.0864 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 52.0082 qps

Latency: min = 19.1467 ms, max = 19.2588 ms, mean = 19.2261 ms, median = 19.2266 ms, percentile(90%) = 19.2383 ms, percentile(95%) = 19.2419 ms, percentile(99%) = 19.25 ms

Enqueue Time: min = 0.0249023 ms, max = 0.46875 ms, mean = 0.0790897 ms, median = 0.0603638 ms, percentile(90%) = 0.118164 ms, percentile(95%) = 0.149414 ms, percentile(99%) = 0.279785 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 = 19.1467 ms, max = 19.2588 ms, mean = 19.2261 ms, median = 19.2266 ms, percentile(90%) = 19.2383 ms, percentile(95%) = 19.2419 ms, percentile(99%) = 19.25 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: 10.0369 s

Total GPU Compute Time: 10.036 s
```