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 (ReLU)

Precision: FP32+INT8

Batch Size 1

```
=== Performance summary ===
Throughput: 357.896 qps

Latency: min = 2.77502 ms, max = 2.9624 ms, mean = 2.79247 ms, median = 2.79248
ms, percentile(90%) = 2.79492 ms, percentile(95%) = 2.7959 ms, percentile(99%) =
2.80164 ms

Enqueue Time: min = 0.0180664 ms, max = 0.642578 ms, mean = 0.041482 ms, median =
0.0273438 ms, percentile(90%) = 0.0732422 ms, percentile(95%) = 0.0891113 ms,
percentile(99%) = 0.166504 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 = 2.77502 ms, max = 2.9624 ms, mean = 2.79247 ms, median =
2.79248 ms, percentile(90%) = 2.79492 ms, percentile(95%) = 2.7959 ms,
percentile(99%) = 2.80164 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: 10.0029 s

Total GPU Compute Time: 9.99703 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 102.953 qps

Latency: min = 9.67065 ms, max = 10.1796 ms, mean = 9.71152 ms, median = 9.71051
ms, percentile(90%) = 9.72266 ms, percentile(95%) = 9.72461 ms, percentile(99%) =
9.72998 ms

Enqueue Time: min = 0.0200195 ms, max = 0.439331 ms, mean = 0.0587875 ms, median =
0.0517578 ms, percentile(90%) = 0.0991211 ms, percentile(95%) = 0.112793 ms,
percentile(99%) = 0.170898 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 = 9.67065 ms, max = 10.1796 ms, mean = 9.71152 ms, median =
9.71051 ms, percentile(90%) = 9.72266 ms, percentile(95%) = 9.72461 ms,
percentile(99%) = 9.72998 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: 10.024 s

Total GPU Compute Time: 10.0223 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 49.097 qps

Latency: min = 20.2906 ms, max = 20.4004 ms, mean = 20.3661 ms, median = 20.3662
ms, percentile(90%) = 20.3857 ms, percentile(95%) = 20.3887 ms, percentile(99%) =
20.396 ms

Enqueue Time: min = 0.0222168 ms, max = 0.419922 ms, mean = 0.0720532 ms, median =
```

```
0.0566406 ms, percentile(90%) = 0.115234 ms, percentile(95%) = 0.128418 ms,
percentile(99%) = 0.242676 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 = 20.2906 ms, max = 20.4004 ms, mean = 20.3661 ms, median =
20.3662 ms, percentile(90%) = 20.3857 ms, percentile(95%) = 20.3887 ms,
percentile(99%) = 20.396 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: 10.0414 s

Total GPU Compute Time: 10.0405 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 32.3057 qps

Latency: min = 30.8419 ms, max = 30.9863 ms, mean = 30.9525 ms, median = 30.9546
ms, percentile(90%) = 30.9658 ms, percentile(95%) = 30.9707 ms, percentile(99%) =
30.9761 ms

Enqueue Time: min = 0.0283203 ms, max = 0.390625 ms, mean = 0.0739696 ms, median =
0.0610352 ms, percentile(90%) = 0.116455 ms, percentile(95%) = 0.128418 ms,
percentile(99%) = 0.209717 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 = 30.8419 ms, max = 30.9863 ms, mean = 30.9525 ms, median =
30.9546 ms, percentile(90%) = 30.9658 ms, percentile(95%) = 30.9707 ms,
percentile(99%) = 30.9761 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: 10.0601 s

Total GPU Compute Time: 10.0596 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 23.8169 qps
```

```
Latency: min = 41.5479 ms, max = 49.151 ms, mean = 41.9852 ms, median = 41.5879
ms, percentile(90%) = 41.6113 ms, percentile(95%) = 44.6689 ms, percentile(99%) =
49.1274 ms

Enqueue Time: min = 0.0297852 ms, max = 0.597656 ms, mean = 0.0753755 ms, median =
0.0606689 ms, percentile(90%) = 0.123535 ms, percentile(95%) = 0.135742 ms,
percentile(99%) = 0.263672 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 = 41.5479 ms, max = 49.151 ms, mean = 41.9852 ms, median =
41.5879 ms, percentile(90%) = 41.6113 ms, percentile(95%) = 44.6689 ms,
percentile(99%) = 49.1274 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: 10.0769 s

Total GPU Compute Time: 10.0765 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 49.1891 qps

Latency: min = 20.2598 ms, max = 20.395 ms, mean = 20.328 ms, median = 20.3286 ms,
percentile(90%) = 20.3398 ms, percentile(95%) = 20.3428 ms, percentile(99%) =
20.3506 ms

Enqueue Time: min = 0.0234375 ms, max = 0.386719 ms, mean = 0.0715278 ms, median =
0.0595093 ms, percentile(90%) = 0.112793 ms, percentile(95%) = 0.127441 ms,
percentile(99%) = 0.239258 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 = 20.2598 ms, max = 20.395 ms, mean = 20.328 ms, median =
20.3286 ms, percentile(90%) = 20.3398 ms, percentile(95%) = 20.3428 ms,
percentile(99%) = 20.3506 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: 10.0429 s

Total GPU Compute Time: 10.042 s
```