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+FP16+INT8

Batch Size 1

```
=== Performance summary ===
Throughput: 353.35 qps

Latency: min = 2.81299 ms, max = 3.00446 ms, mean = 2.82833 ms, median = 2.82812
ms, percentile(90%) = 2.83105 ms, percentile(95%) = 2.83203 ms, percentile(99%) =
2.83984 ms

Enqueue Time: min = 0.0180664 ms, max = 0.638184 ms, mean = 0.0440733 ms, median =
0.0322266 ms, percentile(90%) = 0.0759888 ms, percentile(95%) = 0.0914307 ms,
percentile(99%) = 0.172363 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 = 2.81299 ms, max = 3.00446 ms, mean = 2.82833 ms, median =
2.82812 ms, percentile(90%) = 2.83105 ms, percentile(95%) = 2.83203 ms,
percentile(99%) = 2.83984 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: 10.0043 s

Total GPU Compute Time: 9.99813 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 105.978 qps

Latency: min = 9.4126 ms, max = 9.45801 ms, mean = 9.4341 ms, median = 9.43408 ms,
percentile(90%) = 9.44043 ms, percentile(95%) = 9.44336 ms, percentile(99%) =
9.44873 ms

Enqueue Time: min = 0.0179443 ms, max = 0.487793 ms, mean = 0.0609208 ms, median =
0.0522461 ms, percentile(90%) = 0.101562 ms, percentile(95%) = 0.114746 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 = 9.4126 ms, max = 9.45801 ms, mean = 9.4341 ms, median =
9.43408 ms, percentile(90%) = 9.44043 ms, percentile(95%) = 9.44336 ms,
percentile(99%) = 9.44873 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: 10.0209 s

Total GPU Compute Time: 10.019 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 51.9609 qps

Latency: min = 19.1601 ms, max = 20.0898 ms, mean = 19.2436 ms, median = 19.2441
ms, percentile(90%) = 19.2549 ms, percentile(95%) = 19.2588 ms, percentile(99%) =
19.2646 ms

Enqueue Time: min = 0.0224609 ms, max = 0.500488 ms, mean = 0.0793705 ms, median =
```

```
0.0615234 ms, percentile(90%) = 0.120972 ms, percentile(95%) = 0.150635 ms,
percentile(99%) = 0.366333 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 = 19.1601 ms, max = 20.0898 ms, mean = 19.2436 ms, median =
19.2441 ms, percentile(90%) = 19.2549 ms, percentile(95%) = 19.2588 ms,
percentile(99%) = 19.2646 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: 10.0268 s

Total GPU Compute Time: 10.0259 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 34.1497 qps

Latency: min = 29.184 ms, max = 29.3555 ms, mean = 29.2811 ms, median = 29.2659
ms, percentile(90%) = 29.3379 ms, percentile(95%) = 29.3428 ms, percentile(99%) =
29.3477 ms

Enqueue Time: min = 0.0268555 ms, max = 0.478027 ms, mean = 0.0878111 ms, median =
0.0643921 ms, percentile(90%) = 0.127258 ms, percentile(95%) = 0.196289 ms,
percentile(99%) = 0.312012 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 = 29.184 ms, max = 29.3555 ms, mean = 29.2811 ms, median =
29.2659 ms, percentile(90%) = 29.3379 ms, percentile(95%) = 29.3428 ms,
percentile(99%) = 29.3477 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: 10.0733 s

Total GPU Compute Time: 10.0727 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 25.3653 qps
```

```
Latency: min = 39.3718 ms, max = 39.4531 ms, mean = 39.4221 ms, median = 39.4221
ms, percentile(90%) = 39.4346 ms, percentile(95%) = 39.4404 ms, percentile(99%) =
39.4482 ms

Enqueue Time: min = 0.0361328 ms, max = 0.253418 ms, mean = 0.0759963 ms, median =
0.0615234 ms, percentile(90%) = 0.116211 ms, percentile(95%) = 0.126953 ms,
percentile(99%) = 0.21582 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 = 39.3718 ms, max = 39.4531 ms, mean = 39.4221 ms, median =
39.4221 ms, percentile(90%) = 39.4346 ms, percentile(95%) = 39.4404 ms,
percentile(99%) = 39.4482 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: 10.0925 s

Total GPU Compute Time: 10.0921 s
```

BatchSize Resilience

```
 === Performance summary ===
Throughput: 51.5089 qps

Latency: min = 19.2051 ms, max = 22.4696 ms, mean = 19.4124 ms, median = 19.2358
ms, percentile(90%) = 19.293 ms, percentile(95%) = 20.3141 ms, percentile(99%) =
21.8358 ms

Enqueue Time: min = 0.0234375 ms, max = 0.817383 ms, mean = 0.0787169 ms, median =
0.0625 ms, percentile(90%) = 0.117065 ms, percentile(95%) = 0.129395 ms,
percentile(99%) = 0.319092 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 = 19.2051 ms, max = 22.4696 ms, mean = 19.4124 ms, median =
19.2358 ms, percentile(90%) = 19.293 ms, percentile(95%) = 20.3141 ms,
percentile(99%) = 21.8358 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: 10.0371 s

Total GPU Compute Time: 10.0362 s
```