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+FP16+INT8

Batch Size 1

```
=== Performance summary ===
Throughput: 364.037 qps

Latency: min = 2.73096 ms, max = 3.59717 ms, mean = 2.74528 ms, median = 2.74512
ms, percentile(90%) = 2.74805 ms, percentile(95%) = 2.74854 ms, percentile(99%) =
2.75098 ms

Enqueue Time: min = 0.0175781 ms, max = 0.571777 ms, mean = 0.0383388 ms, median =
0.0263672 ms, percentile(90%) = 0.0683594 ms, percentile(95%) = 0.0827637 ms,
percentile(99%) = 0.141846 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 = 2.73096 ms, max = 3.59717 ms, mean = 2.74528 ms, median =
2.74512 ms, percentile(90%) = 2.74805 ms, percentile(95%) = 2.74854 ms,
percentile(99%) = 2.75098 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: 10.0072 s

Total GPU Compute Time: 10.001 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 105.756 qps

Latency: min = 9.38989 ms, max = 10.3414 ms, mean = 9.45407 ms, median = 9.40234
ms, percentile(90%) = 9.41309 ms, percentile(95%) = 9.75049 ms, percentile(99%) =
10.2922 ms

Enqueue Time: min = 0.0195312 ms, max = 0.985352 ms, mean = 0.0542712 ms, median =
0.0444336 ms, percentile(90%) = 0.0932617 ms, percentile(95%) = 0.107422 ms,
percentile(99%) = 0.186279 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 = 9.38989 ms, max = 10.3414 ms, mean = 9.45407 ms, median =
9.40234 ms, percentile(90%) = 9.41309 ms, percentile(95%) = 9.75049 ms,
percentile(99%) = 10.2922 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: 10.0231 s

Total GPU Compute Time: 10.0213 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 52.0511 qps

Latency: min = 19.1805 ms, max = 19.2422 ms, mean = 19.2102 ms, median = 19.2104
ms, percentile(90%) = 19.2188 ms, percentile(95%) = 19.2214 ms, percentile(99%) =
19.2246 ms

Enqueue Time: min = 0.0229492 ms, max = 0.384277 ms, mean = 0.0693368 ms, median =
0.0612793 ms, percentile(90%) = 0.113281 ms, percentile(95%) = 0.12323 ms,
percentile(99%) = 0.191742 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 = 19.1805 ms, max = 19.2422 ms, mean = 19.2102 ms, median =
19.2104 ms, percentile(90%) = 19.2188 ms, percentile(95%) = 19.2214 ms,
percentile(99%) = 19.2246 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: 10.0478 s

Total GPU Compute Time: 10.0469 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 34.0751 qps

Latency: min = 29.2905 ms, max = 29.4287 ms, mean = 29.3453 ms, median = 29.3447
ms, percentile(90%) = 29.3571 ms, percentile(95%) = 29.3623 ms, percentile(99%) =
29.417 ms

Enqueue Time: min = 0.0224609 ms, max = 0.274902 ms, mean = 0.0678564 ms, median =
0.0586548 ms, percentile(90%) = 0.109619 ms, percentile(95%) = 0.118652 ms,
percentile(99%) = 0.217773 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 = 29.2905 ms, max = 29.4287 ms, mean = 29.3453 ms, median =
29.3447 ms, percentile(90%) = 29.3571 ms, percentile(95%) = 29.3623 ms,
percentile(99%) = 29.417 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: 10.0367 s

Total GPU Compute Time: 10.0361 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 25.0544 qps
```

```
Latency: min = 39.4558 ms, max = 46.7661 ms, mean = 39.9114 ms, median = 39.4937 ms, percentile(90%) = 39.5151 ms, percentile(95%) = 42.2358 ms, percentile(99%) = 46.1701 ms

Enqueue Time: min = 0.0257568 ms, max = 0.327148 ms, mean = 0.0756002 ms, median = 0.0595703 ms, percentile(90%) = 0.121582 ms, percentile(95%) = 0.132812 ms, percentile(99%) = 0.294434 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 = 39.4558 ms, max = 46.7661 ms, mean = 39.9114 ms, median = 39.4937 ms, percentile(90%) = 39.5151 ms, percentile(95%) = 42.2358 ms, percentile(99%) = 46.1701 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: 10.098 s

Total GPU Compute Time: 10.0976 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 51.8237 qps

Latency: min = 19.244 ms, max = 19.3208 ms, mean = 19.2945 ms, median = 19.2949 ms, percentile(90%) = 19.3057 ms, percentile(95%) = 19.3096 ms, percentile(99%) = 19.3127 ms

Enqueue Time: min = 0.0234375 ms, max = 0.484375 ms, mean = 0.0693623 ms, median = 0.0581055 ms, percentile(90%) = 0.108398 ms, percentile(95%) = 0.118896 ms, percentile(99%) = 0.192383 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 = 19.244 ms, max = 19.3208 ms, mean = 19.2945 ms, median = 19.2949 ms, percentile(90%) = 19.3057 ms, percentile(95%) = 19.3096 ms, percentile(99%) = 19.3127 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: 10.034 s

Total GPU Compute Time: 10.0331 s
```