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 (FReLU)

Precision: FP32+INT8

Batch Size 1

```
=== Performance summary ===
Throughput: 324.003 qps

Latency: min = 2.82111 ms, max = 5.31958 ms, mean = 3.08474 ms, median = 2.83838
ms, percentile(90%) = 4.03564 ms, percentile(95%) = 4.2832 ms, percentile(99%) =
4.86816 ms

Enqueue Time: min = 0.0185547 ms, max = 0.978516 ms, mean = 0.0476129 ms, median =
0.0334473 ms, percentile(90%) = 0.0825195 ms, percentile(95%) = 0.102539 ms,
percentile(99%) = 0.228027 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 = 2.82111 ms, max = 5.31958 ms, mean = 3.08474 ms, median =
2.83838 ms, percentile(90%) = 4.03564 ms, percentile(95%) = 4.2832 ms,
percentile(99%) = 4.86816 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: 10.0092 s

Total GPU Compute Time: 10.0038 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 102.444 qps

Latency: min = 9.72186 ms, max = 9.78125 ms, mean = 9.75978 ms, median = 9.76074
ms, percentile(90%) = 9.76855 ms, percentile(95%) = 9.77051 ms, percentile(99%) =
9.77539 ms

Enqueue Time: min = 0.019043 ms, max = 0.353516 ms, mean = 0.0548984 ms, median =
0.0484619 ms, percentile(90%) = 0.0974121 ms, percentile(95%) = 0.111328 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 = 9.72186 ms, max = 9.78125 ms, mean = 9.75978 ms, median =
9.76074 ms, percentile(90%) = 9.76855 ms, percentile(95%) = 9.77051 ms,
percentile(99%) = 9.77539 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: 10.0152 s

Total GPU Compute Time: 10.0135 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 48.856 qps

Latency: min = 20.3981 ms, max = 20.5025 ms, mean = 20.4665 ms, median = 20.4688
ms, percentile(90%) = 20.478 ms, percentile(95%) = 20.4805 ms, percentile(99%) =
20.4868 ms

Enqueue Time: min = 0.0234375 ms, max = 0.841309 ms, mean = 0.0833658 ms, median =
```

```
0.0636597 ms, percentile(90%) = 0.121582 ms, percentile(95%) = 0.172852 ms,
percentile(99%) = 0.361694 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 = 20.3981 ms, max = 20.5025 ms, mean = 20.4665 ms, median =
20.4688 ms, percentile(90%) = 20.478 ms, percentile(95%) = 20.4805 ms,
percentile(99%) = 20.4868 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: 10.0295 s

Total GPU Compute Time: 10.0286 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 32.0432 qps

Latency: min = 30.8439 ms, max = 36.6274 ms, mean = 31.2061 ms, median = 30.8911
ms, percentile(90%) = 30.9141 ms, percentile(95%) = 32.9933 ms, percentile(99%) =
36.0018 ms

Enqueue Time: min = 0.0402832 ms, max = 0.389648 ms, mean = 0.080559 ms, median =
0.0639648 ms, percentile(90%) = 0.116699 ms, percentile(95%) = 0.143555 ms,
percentile(99%) = 0.291016 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 = 30.8439 ms, max = 36.6274 ms, mean = 31.2061 ms, median =
30.8911 ms, percentile(90%) = 30.9141 ms, percentile(95%) = 32.9933 ms,
percentile(99%) = 36.0018 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: 10.0801 s

Total GPU Compute Time: 10.0796 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 24.0878 qps
```

```
Latency: min = 41.4536 ms, max = 41.5469 ms, mean = 41.513 ms, median = 41.5137
ms, percentile(90%) = 41.5303 ms, percentile(95%) = 41.5332 ms, percentile(99%) =
41.541 ms

Enqueue Time: min = 0.0429688 ms, max = 0.796997 ms, mean = 0.0751173 ms, median =
0.0609131 ms, percentile(90%) = 0.114258 ms, percentile(95%) = 0.12207 ms,
percentile(99%) = 0.209473 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 = 41.4536 ms, max = 41.5469 ms, mean = 41.513 ms, median =
41.5137 ms, percentile(90%) = 41.5303 ms, percentile(95%) = 41.5332 ms,
percentile(99%) = 41.541 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: 10.0881 s

Total GPU Compute Time: 10.0877 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 48.6967 qps

Latency: min = 20.4943 ms, max = 20.5693 ms, mean = 20.5335 ms, median = 20.5333
ms, percentile(90%) = 20.5476 ms, percentile(95%) = 20.5508 ms, percentile(99%) =
20.5581 ms

Enqueue Time: min = 0.0224609 ms, max = 1.02319 ms, mean = 0.0793697 ms, median =
0.0625 ms, percentile(90%) = 0.121094 ms, percentile(95%) = 0.133301 ms,
percentile(99%) = 0.224854 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 = 20.4943 ms, max = 20.5693 ms, mean = 20.5335 ms, median =
20.5333 ms, percentile(90%) = 20.5476 ms, percentile(95%) = 20.5508 ms,
percentile(99%) = 20.5581 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: 10.0418 s

Total GPU Compute Time: 10.0409 s
```