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+FP16

Batch Size 1

```
=== Performance summary ===
Throughput: 359.95 qps

Latency: min = 2.75562 ms, max = 2.96136 ms, mean = 2.7764 ms, median = 2.77588
ms, percentile(90%) = 2.7793 ms, percentile(95%) = 2.78027 ms, percentile(99%) =
2.78638 ms

Enqueue Time: min = 0.0185547 ms, max = 0.469238 ms, mean = 0.0413242 ms, median =
0.0273438 ms, percentile(90%) = 0.0737305 ms, percentile(95%) = 0.0908203 ms,
percentile(99%) = 0.167969 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 = 2.75562 ms, max = 2.96136 ms, mean = 2.7764 ms, median = 2.77588 ms, percentile(90%) = 2.7793 ms, percentile(95%) = 2.78027 ms, percentile(99%) = 2.78638 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: 10.0042 s

Total GPU Compute Time: 9.99782 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 106.034 qps

Latency: min = 9.38904 ms, max = 9.45663 ms, mean = 9.42928 ms, median = 9.43066 ms, percentile(90%) = 9.43604 ms, percentile(95%) = 9.4375 ms, percentile(99%) = 9.44092 ms

Enqueue Time: min = 0.0214844 ms, max = 0.550049 ms, mean = 0.0702165 ms, median = 0.0578003 ms, percentile(90%) = 0.109375 ms, percentile(95%) = 0.124512 ms, percentile(99%) = 0.296875 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 = 9.38904 ms, max = 9.45663 ms, mean = 9.42928 ms, median = 9.43066 ms, percentile(90%) = 9.43604 ms, percentile(95%) = 9.4375 ms, percentile(99%) = 9.44092 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: 10.0156 s

Total GPU Compute Time: 10.0139 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 51.7271 qps

Latency: min = 19.2368 ms, max = 19.3574 ms, mean = 19.3306 ms, median = 19.3311 ms, percentile(90%) = 19.3413 ms, percentile(95%) = 19.3442 ms, percentile(99%) = 19.3486 ms

Enqueue Time: min = 0.0244141 ms, max = 0.547363 ms, mean = 0.0788795 ms, median =
0.0615234 ms, percentile(90%) = 0.118164 ms, percentile(95%) = 0.138184 ms,
percentile(99%) = 0.274414 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 = 19.2368 ms, max = 19.3574 ms, mean = 19.3306 ms, median =
19.3311 ms, percentile(90%) = 19.3413 ms, percentile(95%) = 19.3442 ms,
percentile(99%) = 19.3486 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: 10.0334 s

Total GPU Compute Time: 10.0326 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 34.1621 qps

Latency: min = 29.225 ms, max = 29.2969 ms, mean = 29.2704 ms, median = 29.271 ms,
percentile(90%) = 29.2852 ms, percentile(95%) = 29.2881 ms, percentile(99%) =
29.2949 ms

Enqueue Time: min = 0.0263672 ms, max = 0.62207 ms, mean = 0.0789147 ms, median =
0.0615234 ms, percentile(90%) = 0.118164 ms, percentile(95%) = 0.135498 ms,
percentile(99%) = 0.286194 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 = 29.225 ms, max = 29.2969 ms, mean = 29.2704 ms, median =
29.271 ms, percentile(90%) = 29.2852 ms, percentile(95%) = 29.2881 ms,
percentile(99%) = 29.2949 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: 10.0404 s

Total GPU Compute Time: 10.0398 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 25.3869 qps
```

```
Latency: min = 39.3145 ms, max = 39.4238 ms, mean = 39.3886 ms, median = 39.3898 ms, percentile(90%) = 39.4033 ms, percentile(95%) = 39.4106 ms, percentile(99%) = 39.4199 ms

Enqueue Time: min = 0.0322266 ms, max = 0.499512 ms, mean = 0.0784879 ms, median = 0.0630493 ms, percentile(90%) = 0.118774 ms, percentile(95%) = 0.138184 ms, percentile(99%) = 0.253906 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 = 39.3145 ms, max = 39.4238 ms, mean = 39.3886 ms, median = 39.3898 ms, percentile(90%) = 39.4033 ms, percentile(95%) = 39.4106 ms, percentile(99%) = 39.4199 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: 10.0839 s

Total GPU Compute Time: 10.0835 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 51.8812 qps

Latency: min = 19.2051 ms, max = 19.3047 ms, mean = 19.2731 ms, median = 19.2734 ms, percentile(90%) = 19.2849 ms, percentile(95%) = 19.2871 ms, percentile(99%) = 19.2939 ms

Enqueue Time: min = 0.0234375 ms, max = 0.709961 ms, mean = 0.0736384 ms, median = 0.0594482 ms, percentile(90%) = 0.110352 ms, percentile(95%) = 0.127441 ms, percentile(99%) = 0.261719 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 = 19.3047 ms, mean = 19.2731 ms, median = 19.2734 ms, percentile(90%) = 19.2849 ms, percentile(95%) = 19.2871 ms, percentile(99%) = 19.2939 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: 10.0422 s

Total GPU Compute Time: 10.0413 s
```