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

Batch Size 1

```
=== Performance summary ===
Throughput: 365.25 qps

Latency: min = 2.72076 ms, max = 2.7627 ms, mean = 2.73618 ms, median = 2.7373 ms, percentile(90%) = 2.73926 ms, percentile(95%) = 2.74023 ms, percentile(99%) = 2.74121 ms

Enqueue Time: min = 0.0180664 ms, max = 0.554688 ms, mean = 0.0436635 ms, median = 0.0307922 ms, percentile(90%) = 0.0749512 ms, percentile(95%) = 0.0883789 ms, percentile(99%) = 0.168945 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 = 2.72076 ms, max = 2.7627 ms, mean = 2.73618 ms, median =
2.7373 ms, percentile(90%) = 2.73926 ms, percentile(95%) = 2.74023 ms,
percentile(99%) = 2.74121 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: 10.0041 s

Total GPU Compute Time: 9.99799 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 106.55 qps

Latency: min = 9.33887 ms, max = 9.85602 ms, mean = 9.38359 ms, median = 9.38232
ms, percentile(90%) = 9.39453 ms, percentile(95%) = 9.39648 ms, percentile(99%) =
9.4043 ms

Enqueue Time: min = 0.0185547 ms, max = 0.404541 ms, mean = 0.0530362 ms, median =
0.045166 ms, percentile(90%) = 0.0917969 ms, percentile(95%) = 0.101929 ms,
percentile(99%) = 0.182617 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 = 9.33887 ms, max = 9.85602 ms, mean = 9.38359 ms, median =
9.38232 ms, percentile(90%) = 9.39453 ms, percentile(95%) = 9.39648 ms,
percentile(99%) = 9.4043 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: 10.0141 s

Total GPU Compute Time: 10.0123 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 52.0447 qps

Latency: min = 19.1416 ms, max = 19.2432 ms, mean = 19.2125 ms, median = 19.2153
ms, percentile(90%) = 19.2285 ms, percentile(95%) = 19.2314 ms, percentile(99%) =
19.2373 ms

Enqueue Time: min = 0.0244141 ms, max = 0.469727 ms, mean = 0.0681757 ms, median =
0.0585938 ms, percentile(90%) = 0.110352 ms, percentile(95%) = 0.121094 ms,
percentile(99%) = 0.201172 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 = 19.1416 ms, max = 19.2432 ms, mean = 19.2125 ms, median =
19.2153 ms, percentile(90%) = 19.2285 ms, percentile(95%) = 19.2314 ms,
percentile(99%) = 19.2373 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: 10.0298 s

Total GPU Compute Time: 10.0289 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 34.0602 qps

Latency: min = 29.3171 ms, max = 29.3848 ms, mean = 29.358 ms, median = 29.3582
ms, percentile(90%) = 29.3711 ms, percentile(95%) = 29.3745 ms, percentile(99%) =
29.3799 ms

Enqueue Time: min = 0.0214844 ms, max = 1.28125 ms, mean = 0.0729841 ms, median =
0.0595703 ms, percentile(90%) = 0.109131 ms, percentile(95%) = 0.12207 ms,
percentile(99%) = 0.225586 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 = 29.3171 ms, max = 29.3848 ms, mean = 29.358 ms, median =
29.3582 ms, percentile(90%) = 29.3711 ms, percentile(95%) = 29.3745 ms,
percentile(99%) = 29.3799 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: 10.0704 s

Total GPU Compute Time: 10.0698 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 25.3558 qps
```

```
Latency: min = 39.3062 ms, max = 39.4702 ms, mean = 39.4369 ms, median = 39.4434
ms, percentile(90%) = 39.457 ms, percentile(95%) = 39.4609 ms, percentile(99%) =
39.4658 ms

Enqueue Time: min = 0.0280762 ms, max = 0.297852 ms, mean = 0.0752583 ms, median =
0.0595703 ms, percentile(90%) = 0.117981 ms, percentile(95%) = 0.133789 ms,
percentile(99%) = 0.247559 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 = 39.3062 ms, max = 39.4702 ms, mean = 39.4369 ms, median =
39.4434 ms, percentile(90%) = 39.457 ms, percentile(95%) = 39.4609 ms,
percentile(99%) = 39.4658 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: 10.0963 s

Total GPU Compute Time: 10.0959 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 51.8852 qps

Latency: min = 19.2103 ms, max = 19.2993 ms, mean = 19.2714 ms, median = 19.2725
ms, percentile(90%) = 19.2871 ms, percentile(95%) = 19.29 ms, percentile(99%) =
19.2964 ms

Enqueue Time: min = 0.0205078 ms, max = 0.53772 ms, mean = 0.0724351 ms, median =
0.0610352 ms, percentile(90%) = 0.114258 ms, percentile(95%) = 0.132812 ms,
percentile(99%) = 0.210449 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 = 19.2103 ms, max = 19.2993 ms, mean = 19.2714 ms, median =
19.2725 ms, percentile(90%) = 19.2871 ms, percentile(95%) = 19.29 ms,
percentile(99%) = 19.2964 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: 10.0404 s
```