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

Batch Size 1

```
=== Performance summary ===
Throughput: 353.015 qps

Latency: min = 2.80988 ms, max = 3.0116 ms, mean = 2.83097 ms, median = 2.83057
ms, percentile(90%) = 2.83545 ms, percentile(95%) = 2.83649 ms, percentile(99%) =
2.83984 ms

Enqueue Time: min = 0.0180664 ms, max = 0.687744 ms, mean = 0.0383459 ms, median =
0.0258789 ms, percentile(90%) = 0.0683594 ms, percentile(95%) = 0.0825195 ms,
percentile(99%) = 0.160156 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 = 2.80988 ms, max = 3.0116 ms, mean = 2.83097 ms, median =
2.83057 ms, percentile(90%) = 2.83545 ms, percentile(95%) = 2.83649 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: 10.0052 s

Total GPU Compute Time: 9.999 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 106.051 qps

Latency: min = 9.38904 ms, max = 9.45166 ms, mean = 9.42776 ms, median = 9.42969
ms, percentile(90%) = 9.43604 ms, percentile(95%) = 9.4375 ms, percentile(99%) =
9.44141 ms

Enqueue Time: min = 0.019043 ms, max = 0.645996 ms, mean = 0.0529916 ms, median =
0.0449219 ms, percentile(90%) = 0.0913086 ms, percentile(95%) = 0.100586 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 = 9.38904 ms, max = 9.45166 ms, mean = 9.42776 ms, median =
9.42969 ms, percentile(90%) = 9.43604 ms, percentile(95%) = 9.4375 ms,
percentile(99%) = 9.44141 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: 51.9759 qps

Latency: min = 19.1724 ms, max = 20.1001 ms, mean = 19.238 ms, median = 19.2393
ms, percentile(90%) = 19.2502 ms, percentile(95%) = 19.2549 ms, percentile(99%) =
19.2627 ms

Enqueue Time: min = 0.0227051 ms, max = 0.54834 ms, mean = 0.0778302 ms, median =
0.0629883 ms, percentile(90%) = 0.117188 ms, percentile(95%) = 0.133789 ms,
percentile(99%) = 0.245605 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 = 19.1724 ms, max = 20.1001 ms, mean = 19.238 ms, median =
19.2393 ms, percentile(90%) = 19.2502 ms, percentile(95%) = 19.2549 ms,
percentile(99%) = 19.2627 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: 10.0431 s

Total GPU Compute Time: 10.0422 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 34.1773 qps

Latency: min = 29.2024 ms, max = 29.2959 ms, mean = 29.2575 ms, median = 29.2583
ms, percentile(90%) = 29.2725 ms, percentile(95%) = 29.2754 ms, percentile(99%) =
29.2832 ms

Enqueue Time: min = 0.0249023 ms, max = 0.720703 ms, mean = 0.0767499 ms, median =
0.0595703 ms, percentile(90%) = 0.116211 ms, percentile(95%) = 0.132324 ms,
percentile(99%) = 0.318359 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 = 29.2024 ms, max = 29.2959 ms, mean = 29.2575 ms, median =
29.2583 ms, percentile(90%) = 29.2725 ms, percentile(95%) = 29.2754 ms,
percentile(99%) = 29.2832 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: 10.0652 s

Total GPU Compute Time: 10.0646 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 25.3801 qps
```

```
Latency: min = 39.2693 ms, max = 39.4312 ms, mean = 39.3991 ms, median = 39.4033
ms, percentile(90%) = 39.416 ms, percentile(95%) = 39.4219 ms, percentile(99%) =
39.4282 ms

Enqueue Time: min = 0.0224609 ms, max = 0.902344 ms, mean = 0.0781488 ms, median =
0.0612793 ms, percentile(90%) = 0.116699 ms, percentile(95%) = 0.128418 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 = 39.2693 ms, max = 39.4312 ms, mean = 39.3991 ms, median =
39.4033 ms, percentile(90%) = 39.416 ms, percentile(95%) = 39.4219 ms,
percentile(99%) = 39.4282 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: 10.0866 s

Total GPU Compute Time: 10.0862 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 52.0234 qps

Latency: min = 19.1498 ms, max = 19.2461 ms, mean = 19.2205 ms, median = 19.2227
ms, percentile(90%) = 19.2344 ms, percentile(95%) = 19.2383 ms, percentile(99%) =
19.2422 ms

Enqueue Time: min = 0.0229492 ms, max = 2 ms, mean = 0.0917844 ms, median =
0.0683594 ms, percentile(90%) = 0.128906 ms, percentile(95%) = 0.172852 ms,
percentile(99%) = 0.34668 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 = 19.1498 ms, max = 19.2461 ms, mean = 19.2205 ms, median =
19.2227 ms, percentile(90%) = 19.2344 ms, percentile(95%) = 19.2383 ms,
percentile(99%) = 19.2422 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: 10.0532 s

Total GPU Compute Time: 10.0523 s
```