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 (AconC)

Precision: FP32+FP16+INT8

Batch Size 1

```
=== Performance summary ===
Throughput: 344.474 qps

Latency: min = 2.89282 ms, max = 2.92041 ms, mean = 2.9013 ms, median = 2.90137
ms, percentile(90%) = 2.90405 ms, percentile(95%) = 2.90479 ms, percentile(99%) =
2.90625 ms

Enqueue Time: min = 0.0195312 ms, max = 0.907837 ms, mean = 0.0503547 ms, median =
0.03125 ms, percentile(90%) = 0.0878906 ms, percentile(95%) = 0.109375 ms,
percentile(99%) = 0.259888 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 = 2.89282 ms, max = 2.92041 ms, mean = 2.9013 ms, median =
2.90137 ms, percentile(90%) = 2.90405 ms, percentile(95%) = 2.90479 ms,
percentile(99%) = 2.90625 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: 10.0066 s

Total GPU Compute Time: 10.0008 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 106.321 qps

Latency: min = 9.32959 ms, max = 10.3383 ms, mean = 9.40385 ms, median = 9.34619
ms, percentile(90%) = 9.37988 ms, percentile(95%) = 9.67163 ms, percentile(99%) =
10.1929 ms

Enqueue Time: min = 0.0195312 ms, max = 0.427734 ms, mean = 0.0606582 ms, median =
0.0527344 ms, percentile(90%) = 0.103516 ms, percentile(95%) = 0.118164 ms,
percentile(99%) = 0.230469 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 = 9.32959 ms, max = 10.3383 ms, mean = 9.40385 ms, median =
9.34619 ms, percentile(90%) = 9.37988 ms, percentile(95%) = 9.67163 ms,
percentile(99%) = 10.1929 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: 10.0169 s

Total GPU Compute Time: 10.0151 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 51.8766 qps

Latency: min = 19.2358 ms, max = 20.0448 ms, mean = 19.2749 ms, median = 19.269
ms, percentile(90%) = 19.2793 ms, percentile(95%) = 19.2822 ms, percentile(99%) =
19.292 ms

Enqueue Time: min = 0.0234375 ms, max = 0.332031 ms, mean = 0.0658474 ms, median =
0.057373 ms, percentile(90%) = 0.104492 ms, percentile(95%) = 0.113281 ms,
percentile(99%) = 0.161621 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 = 19.2358 ms, max = 20.0448 ms, mean = 19.2749 ms, median =
19.269 ms, percentile(90%) = 19.2793 ms, percentile(95%) = 19.2822 ms,
percentile(99%) = 19.292 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: 34.1904 qps

Latency: min = 29.2045 ms, max = 29.3232 ms, mean = 29.2463 ms, median = 29.2446
ms, percentile(90%) = 29.2578 ms, percentile(95%) = 29.2637 ms, percentile(99%) =
29.3164 ms

Enqueue Time: min = 0.0336914 ms, max = 0.956543 ms, mean = 0.0883362 ms, median =
0.067627 ms, percentile(90%) = 0.123047 ms, percentile(95%) = 0.165039 ms,
percentile(99%) = 0.335938 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 = 29.2045 ms, max = 29.3232 ms, mean = 29.2463 ms, median =
29.2446 ms, percentile(90%) = 29.2578 ms, percentile(95%) = 29.2637 ms,
percentile(99%) = 29.3164 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: 10.0613 s

Total GPU Compute Time: 10.0607 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 25.4242 qps
```

```
Latency: min = 39.1772 ms, max = 39.4248 ms, mean = 39.3309 ms, median = 39.311
ms, percentile(90%) = 39.4033 ms, percentile(95%) = 39.4092 ms, percentile(99%) =
39.418 ms

Enqueue Time: min = 0.0317383 ms, max = 0.17041 ms, mean = 0.0666009 ms, median =
0.059082 ms, percentile(90%) = 0.106201 ms, percentile(95%) = 0.114197 ms,
percentile(99%) = 0.141479 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 = 39.1772 ms, max = 39.4248 ms, mean = 39.3309 ms, median =
39.311 ms, percentile(90%) = 39.4033 ms, percentile(95%) = 39.4092 ms,
percentile(99%) = 39.418 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: 10.1085 s

Total GPU Compute Time: 10.108 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 45.6912 qps

Latency: min = 19.2113 ms, max = 40.3887 ms, mean = 21.8843 ms, median = 22.3374
ms, percentile(90%) = 24.1133 ms, percentile(95%) = 24.3425 ms, percentile(99%) =
27.8108 ms

Enqueue Time: min = 0.0280762 ms, max = 0.408203 ms, mean = 0.0773624 ms, median =
0.060791 ms, percentile(90%) = 0.111633 ms, percentile(95%) = 0.154785 ms,
percentile(99%) = 0.307373 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 = 19.2113 ms, max = 40.3887 ms, mean = 21.8843 ms, median =
22.3374 ms, percentile(90%) = 24.1133 ms, percentile(95%) = 24.3425 ms,
percentile(99%) = 27.8108 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: 10.0457 s

Total GPU Compute Time: 10.0449 s
```