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+INT8

Batch Size 1

```
=== Performance summary ===
Throughput: 337.733 qps

Latency: min = 2.93579 ms, max = 4.30487 ms, mean = 2.95929 ms, median = 2.94385
ms, percentile(90%) = 2.97058 ms, percentile(95%) = 2.99933 ms, percentile(99%) =
3.45093 ms

Enqueue Time: min = 0.0187988 ms, max = 2.05127 ms, mean = 0.0561888 ms, median =
0.0449219 ms, percentile(90%) = 0.0927734 ms, percentile(95%) = 0.12207 ms,
percentile(99%) = 0.289062 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 = 2.93579 ms, max = 4.30487 ms, mean = 2.95929 ms, median =
2.94385 ms, percentile(90%) = 2.97058 ms, percentile(95%) = 2.99933 ms,
percentile(99%) = 3.45093 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: 10.0049 s

Total GPU Compute Time: 9.99945 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 91.1228 qps

Latency: min = 9.73212 ms, max = 17.251 ms, mean = 10.9725 ms, median = 9.78418
ms, percentile(90%) = 14.7383 ms, percentile(95%) = 15.232 ms, percentile(99%) =
15.9211 ms

Enqueue Time: min = 0.0209961 ms, max = 0.863403 ms, mean = 0.0601527 ms, median =
0.0527344 ms, percentile(90%) = 0.0957031 ms, percentile(95%) = 0.10791 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 = 9.73212 ms, max = 17.251 ms, mean = 10.9725 ms, median =
9.78418 ms, percentile(90%) = 14.7383 ms, percentile(95%) = 15.232 ms,
percentile(99%) = 15.9211 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: 10.0194 s

Total GPU Compute Time: 10.0179 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 29.8483 qps

Latency: min = 30.2837 ms, max = 43.2568 ms, mean = 33.5009 ms, median = 33.0537
ms, percentile(90%) = 35.6905 ms, percentile(95%) = 36.5486 ms, percentile(99%) =
38.5537 ms

Enqueue Time: min = 0.0400391 ms, max = 0.246094 ms, mean = 0.0806334 ms, median =
```

```
0.0678711 ms, percentile(90%) = 0.109741 ms, percentile(95%) = 0.121094 ms,
percentile(99%) = 0.171143 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 = 30.2837 ms, max = 43.2568 ms, mean = 33.5009 ms, median =
33.0537 ms, percentile(90%) = 35.6905 ms, percentile(95%) = 36.5486 ms,
percentile(99%) = 38.5537 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: 10.0843 s

Total GPU Compute Time: 10.0838 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 32.3757 qps

Latency: min = 30.8633 ms, max = 30.9238 ms, mean = 30.8858 ms, median = 30.8851
ms, percentile(90%) = 30.8965 ms, percentile(95%) = 30.9001 ms, percentile(99%) =
30.9065 ms

Enqueue Time: min = 0.0249023 ms, max = 0.415527 ms, mean = 0.0910913 ms, median =
0.0742188 ms, percentile(90%) = 0.130371 ms, percentile(95%) = 0.162476 ms,
percentile(99%) = 0.232422 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 = 30.8633 ms, max = 30.9238 ms, mean = 30.8858 ms, median =
30.8851 ms, percentile(90%) = 30.8965 ms, percentile(95%) = 30.9001 ms,
percentile(99%) = 30.9065 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: 10.0693 s

Total GPU Compute Time: 10.0688 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 24.101 qps
```

```
Latency: min = 41.3747 ms, max = 41.522 ms, mean = 41.4902 ms, median = 41.4966
ms, percentile(90%) = 41.5088 ms, percentile(95%) = 41.5129 ms, percentile(99%) =
41.519 ms

Enqueue Time: min = 0.0273438 ms, max = 0.385742 ms, mean = 0.078688 ms, median =
0.0616455 ms, percentile(90%) = 0.120605 ms, percentile(95%) = 0.13916 ms,
percentile(99%) = 0.320312 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 = 41.3747 ms, max = 41.522 ms, mean = 41.4902 ms, median =
41.4966 ms, percentile(90%) = 41.5088 ms, percentile(95%) = 41.5129 ms,
percentile(99%) = 41.519 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: 10.0826 s

Total GPU Compute Time: 10.0821 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 48.8238 qps

Latency: min = 20.2842 ms, max = 23.3073 ms, mean = 20.4801 ms, median = 20.3057
ms, percentile(90%) = 20.3633 ms, percentile(95%) = 21.546 ms, percentile(99%) =
23.296 ms

Enqueue Time: min = 0.0214844 ms, max = 0.352539 ms, mean = 0.0685095 ms, median =
0.059082 ms, percentile(90%) = 0.109009 ms, percentile(95%) = 0.121887 ms,
percentile(99%) = 0.214355 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 = 20.2842 ms, max = 23.3073 ms, mean = 20.4801 ms, median =
20.3057 ms, percentile(90%) = 20.3633 ms, percentile(95%) = 21.546 ms,
percentile(99%) = 23.296 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: 10.0361 s

Total GPU Compute Time: 10.0352 s
```