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+INT8

Batch Size 1

```
=== Performance summary ===
Throughput: 358.416 qps

Latency: min = 2.77094 ms, max = 2.94812 ms, mean = 2.78832 ms, median = 2.78955 ms, percentile(90%) = 2.79297 ms, percentile(95%) = 2.7937 ms, percentile(99%) = 2.7959 ms

Enqueue Time: min = 0.0185547 ms, max = 0.450195 ms, mean = 0.0424371 ms, median = 0.03125 ms, percentile(90%) = 0.0727539 ms, percentile(95%) = 0.0847168 ms, percentile(99%) = 0.146484 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 = 2.77094 ms, max = 2.94812 ms, mean = 2.78832 ms, median =
2.78955 ms, percentile(90%) = 2.79297 ms, percentile(95%) = 2.7937 ms,
percentile(99%) = 2.7959 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: 10.0051 s

Total GPU Compute Time: 9.99891 s
```

BatchSize 4

```
=== Performance summary ===
Throughput: 105.776 qps

Latency: min = 9.43817 ms, max = 9.48633 ms, mean = 9.45221 ms, median = 9.4502
ms, percentile(90%) = 9.46777 ms, percentile(95%) = 9.47559 ms, percentile(99%) =
9.48047 ms

Enqueue Time: min = 0.019043 ms, max = 0.359863 ms, mean = 0.0455117 ms, median =
0.0390625 ms, percentile(90%) = 0.0771484 ms, percentile(95%) = 0.0927734 ms,
percentile(99%) = 0.148438 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 = 9.43817 ms, max = 9.48633 ms, mean = 9.45221 ms, median =
9.4502 ms, percentile(90%) = 9.46777 ms, percentile(95%) = 9.47559 ms,
percentile(99%) = 9.48047 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: 10.0117 s

Total GPU Compute Time: 10.0099 s
```

BatchSize 8

```
=== Performance summary ===
Throughput: 51.734 qps

Latency: min = 19.243 ms, max = 19.3516 ms, mean = 19.3279 ms, median = 19.3291
ms, percentile(90%) = 19.3379 ms, percentile(95%) = 19.3403 ms, percentile(99%) =
19.3477 ms

Enqueue Time: min = 0.0224609 ms, max = 0.878418 ms, mean = 0.0757419 ms, median =
0.0610352 ms, percentile(90%) = 0.118164 ms, percentile(95%) = 0.13623 ms,
percentile(99%) = 0.242676 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 = 19.243 ms, max = 19.3516 ms, mean = 19.3279 ms, median =
19.3291 ms, percentile(90%) = 19.3379 ms, percentile(95%) = 19.3403 ms,
percentile(99%) = 19.3477 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: 10.0321 s

Total GPU Compute Time: 10.0312 s
```

BatchSize 12

```
=== Performance summary ===
Throughput: 34.1574 qps

Latency: min = 29.2311 ms, max = 29.3091 ms, mean = 29.2746 ms, median = 29.2751
ms, percentile(90%) = 29.2866 ms, percentile(95%) = 29.291 ms, percentile(99%) =
29.2966 ms

Enqueue Time: min = 0.0263672 ms, max = 0.88623 ms, mean = 0.0745563 ms, median =
0.059082 ms, percentile(90%) = 0.11377 ms, percentile(95%) = 0.122803 ms,
percentile(99%) = 0.209473 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 = 29.2311 ms, max = 29.3091 ms, mean = 29.2746 ms, median =
29.2751 ms, percentile(90%) = 29.2866 ms, percentile(95%) = 29.291 ms,
percentile(99%) = 29.2966 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: 10.071 s

Total GPU Compute Time: 10.0704 s
```

BatchSize 16

```
=== Performance summary ===
Throughput: 25.1008 qps
```

```
Latency: min = 39.3687 ms, max = 46.1865 ms, mean = 39.8375 ms, median = 39.4756
ms, percentile(90%) = 39.5117 ms, percentile(95%) = 42.2615 ms, percentile(99%) =
46.1763 ms

Enqueue Time: min = 0.0263672 ms, max = 0.223145 ms, mean = 0.0738013 ms, median =
0.059082 ms, percentile(90%) = 0.120239 ms, percentile(95%) = 0.130859 ms,
percentile(99%) = 0.171875 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 = 39.3687 ms, max = 46.1865 ms, mean = 39.8375 ms, median =
39.4756 ms, percentile(90%) = 39.5117 ms, percentile(95%) = 42.2615 ms,
percentile(99%) = 46.1763 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: 10.0794 s

Total GPU Compute Time: 10.0789 s
```

BatchSize Resilience

```
=== Performance summary ===
Throughput: 51.7499 qps

Latency: min = 19.2522 ms, max = 19.3545 ms, mean = 19.322 ms, median = 19.3271
ms, percentile(90%) = 19.3379 ms, percentile(95%) = 19.3408 ms, percentile(99%) =
19.3477 ms

Enqueue Time: min = 0.0239258 ms, max = 0.717773 ms, mean = 0.0703685 ms, median =
0.0570374 ms, percentile(90%) = 0.110352 ms, percentile(95%) = 0.126465 ms,
percentile(99%) = 0.255371 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 = 19.2522 ms, max = 19.3545 ms, mean = 19.322 ms, median =
19.3271 ms, percentile(90%) = 19.3379 ms, percentile(95%) = 19.3408 ms,
percentile(99%) = 19.3477 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: 10.0483 s

Total GPU Compute Time: 10.0475 s
```