AIXPRT Community Preview user guide (OpenVINO on Windows)

Introduction

The AIXPRT OpenVINO on Windows package has the capability of running on the following platforms:

- CPUs
- Intel processor graphics

The workloads are implemented using the publicly available libraries and SDKs for each platform.

Precompiled Intel OpenVINO (CPU and GPU) on Windows

To simplify the installation process for some testers, we offer an AIXPRT download package with a precompiled version of the Intel Distribution of OpenVINO toolkit for Windows. This package only runs OpenVINO, and does not contain the TensorFlow and TensorFlow-TensorRT frameworks. It contains the resnet50_v1 and ssd-mobilenet workloads and runs single and multi-batch size scenarios.

System requirements

Operating System

- Windows 10

CPU

- 6th to 8th generation Intel Core and Intel Xeon processors
- Intel Pentium processor N4200/5, N3350/5, N3450/5 with Intel HD Graphics

Installation and system configuration

1. Download and unzip the installation package.
2. Open a Command Prompt in Windows.
3. Navigate to AIXPRT_CP2_OpenVINO\install, and run the following command, which will install all dependencies and prepare the benchmark to run:

```
setup_AIXPRT.bat
```

4. During the installation process, please review any prompts and allow the installation of necessary dependencies.

Running the benchmark

1. Navigate to the AIXPRT harness directory:

```
cd AIXPRT/Harness
```

2. Edit the benchmark configuration file as necessary. Please see Appendix A for instructions. By default, AIXPRT, tests ResNet50 and SSD-MobileNet workloads on CPU in FP32 and INT8 precisions and up to a batch size of 32.
3. Run the benchmark using the following script:

```
python3 index.py
```

4. If targeting the GPU, edit AIXPRT/Config/{filename.json} to set “hardware” to GPU.

Results

When the test is complete, the benchmark saves the results to AIXPRT/Results in JSON format, and also generates CSV files with the name {ConfigName}_RESULTS_SUMMARY.csv. To summarize the results and convert them to a more readable CSV format, please run the following command:

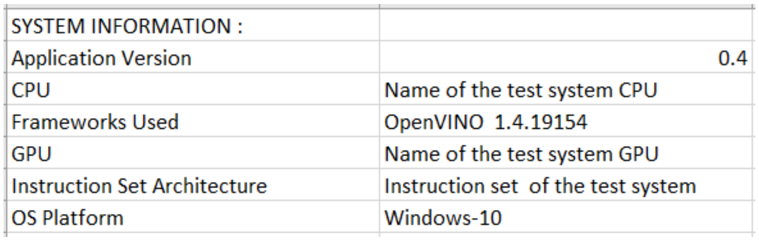
```
python3 resultsParser.py
```

To submit results, please follow the instructions in ResultSubmission.md or at https://www.principledtechnologies.com/benchmarkxprt/aixprt/2019/submit-results.php.

Sample results summary file

Each results summary file has three sections: SYSTEM INFORMATION, RESULTS SUMMARY, and DETAILED RESULTS.

1. SYSTEM INFORMATION
   This section provides basic information about the system under test.

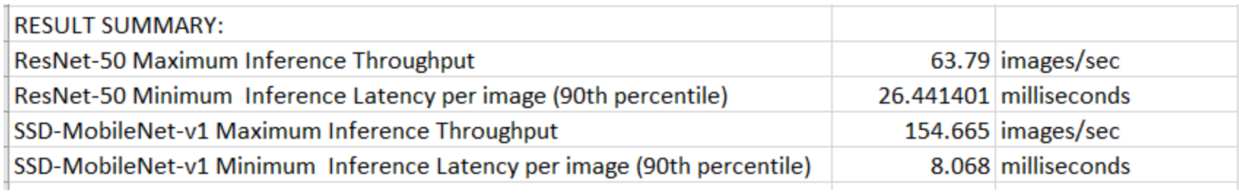


SYSTEM INFORMATION :	
Application Version	0.4
CPU	Name of the test system CPU
Frameworks Used	OpenVINO 1.4.19154
GPU	Name of the test system GPU
Instruction Set Architecture	Instruction set of the test system
OS Platform	Windows-10

2. RESULTS SUMMARY
   AIXPRT measures inference latency and throughput for image recognition (ResNet-50) and object detection (SSD-MobileNet) tasks. Batching tasks allows AI applications to achieve higher levels of throughput, but higher throughput may come at the expense of increased latency per task. In real-time or near real-time use cases like performing image recognition on individual photos being captured by a camera, lower latency is important to enable better user experience. In other cases, like performing image recognition on a large library of photos, higher throughput through batching images or concurrent instances may allow faster completion of the overall workload.

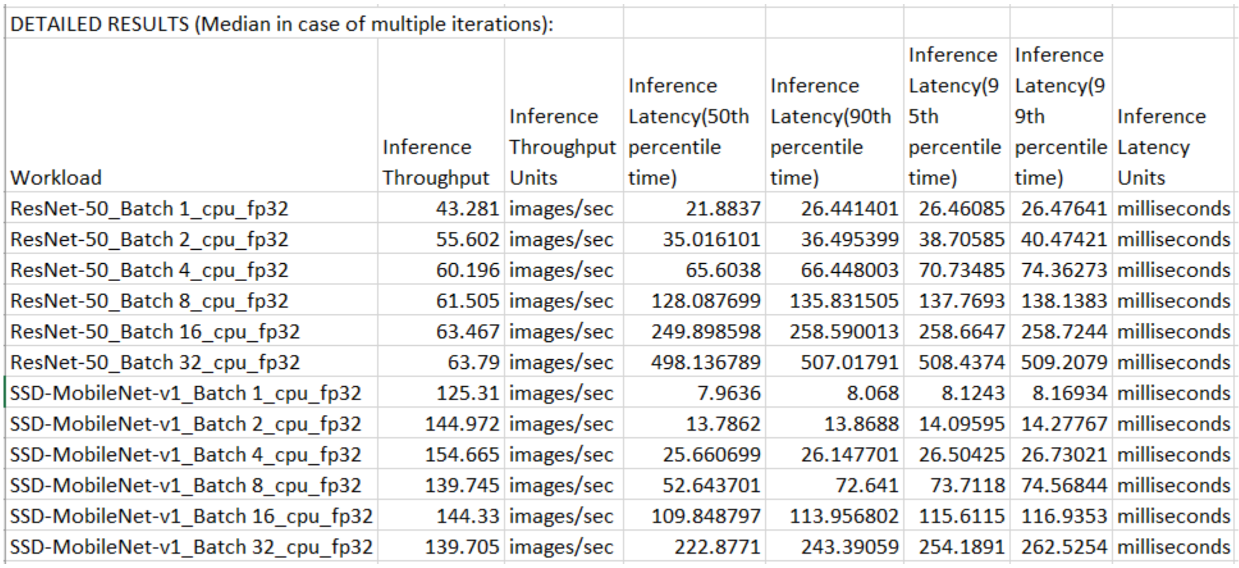
   The achieve optimal latency and/or throughput levels, AI applications often tune batch sizes and/or concurrent instances according to a system’s hardware capabilities, such as the number of available processor cores and threads. To represent a spectrum of common tunings, AIXPRT tests AI tasks in different batch sizes (1 – 32 is the default in this package) that are relevant to the target test system. AIXPRT then reports the maximum throughput and minimum latency for image recognition (ResNet-50) and object detection (SSD-MobileNet v1) usages.

   The AIXPRT results summary (example below) makes it easier to quickly identify relevant comparisons between systems.



RESULT SUMMARY:		
ResNet-50 Maximum Inference Throughput	63.79	images/sec
ResNet-50 Minimum Inference Latency per image (90th percentile)	26.441401	milliseconds
SSD-MobileNet-v1 Maximum Inference Throughput	154.665	images/sec
SSD-MobileNet-v1 Minimum Inference Latency per image (90th percentile)	8.068	milliseconds

3. DETAILED RESULTS
   This section shows the throughput and latency results for each AI task configuration tested by the benchmark. AIXPRT runs each AI task (e.g. ResNet-50, Batch1, on CPU) multiple times and reports the average inference throughput and corresponding latency percentiles.



DETAILED RESULTS (Median in case of multiple iterations):							
Workload	Inference Throughput	Inference Throughput Units	Inference Latency(50th percentile time)	Inference Latency(90th percentile time)	Inference Latency(95th percentile time)	Inference Latency(99th percentile time)	Inference Latency Units
ResNet-50_Batch 1_cpu_fp32	43.281	images/sec	21.8837	26.441401	26.46085	26.47641	milliseconds
ResNet-50_Batch 2_cpu_fp32	55.602	images/sec	35.016101	36.495399	38.70585	40.47421	milliseconds
ResNet-50_Batch 4_cpu_fp32	60.196	images/sec	65.6038	66.448003	70.73485	74.36273	milliseconds
ResNet-50_Batch 8_cpu_fp32	61.505	images/sec	128.087699	135.831505	137.7693	138.1383	milliseconds
ResNet-50_Batch 16_cpu_fp32	63.467	images/sec	249.898598	258.590013	258.6647	258.7244	milliseconds
ResNet-50_Batch 32_cpu_fp32	63.79	images/sec	498.136789	507.01791	508.4374	509.2079	milliseconds
SSD-MobileNet-v1_Batch 1_cpu_fp32	125.31	images/sec	7.9636	8.068	8.1243	8.16934	milliseconds
SSD-MobileNet-v1_Batch 2_cpu_fp32	144.972	images/sec	13.7862	13.8688	14.09595	14.27767	milliseconds
SSD-MobileNet-v1_Batch 4_cpu_fp32	154.665	images/sec	25.660699	26.147701	26.50425	26.73021	milliseconds
SSD-MobileNet-v1_Batch 8_cpu_fp32	139.745	images/sec	52.643701	72.641	73.7118	74.56844	milliseconds
SSD-MobileNet-v1_Batch 16_cpu_fp32	144.33	images/sec	109.848797	113.956802	115.6115	116.9353	milliseconds
SSD-MobileNet-v1_Batch 32_cpu_fp32	139.705	images/sec	222.8771	243.39059	254.1891	262.5254	milliseconds

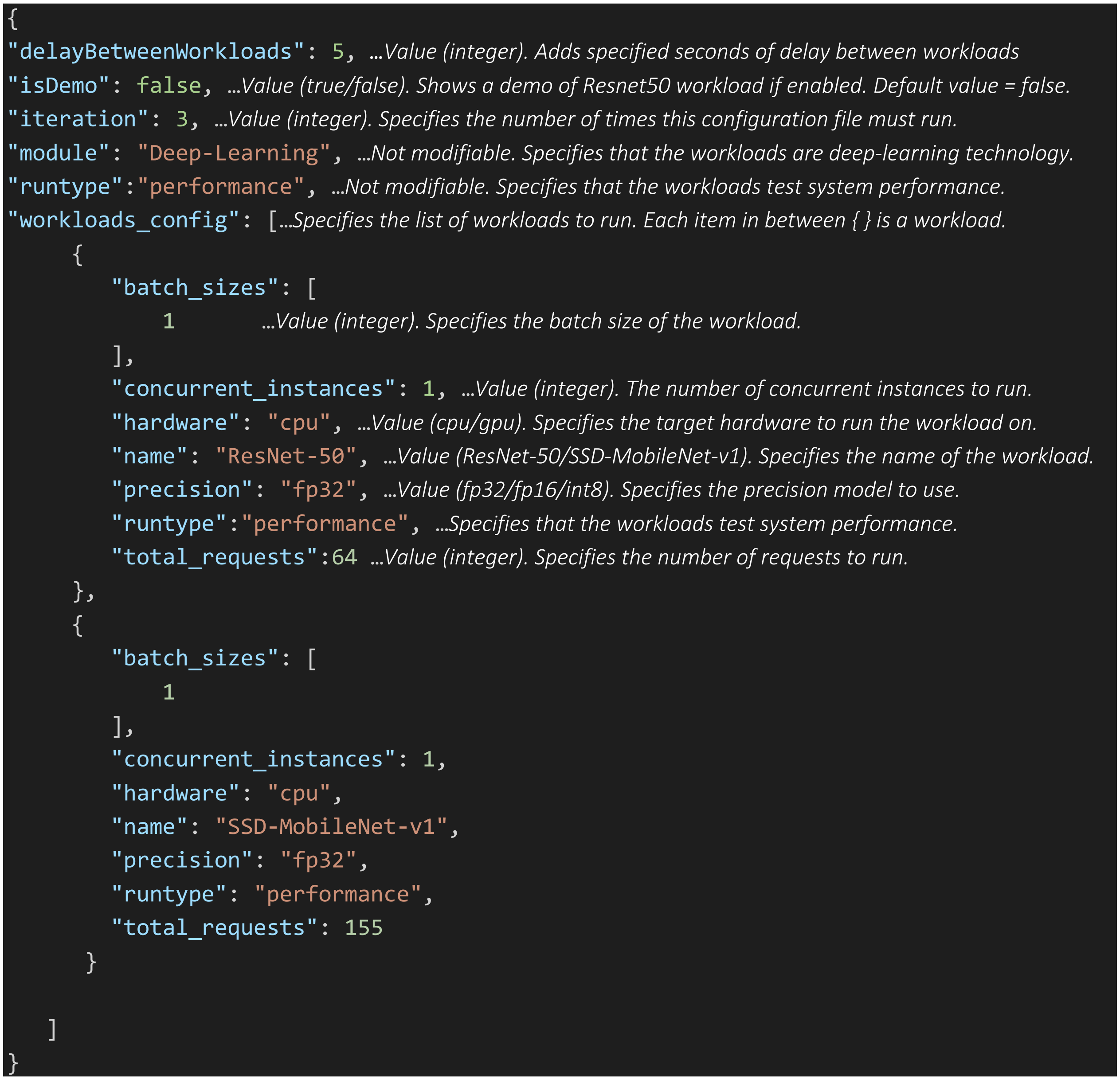
Support

If you need technical support or have any questions, please send a message to BenchmarkXPRTsupport@principledtechnologies.com.

Appendix A: How to configure an AIXPRT test run

AIXPRT configuration files, located in AIXPRT/Configs, allow the user to customize a benchmark run according to their use case. In this package, the default configuration is set to test AIXPRT workloads in FP32 and INT8 precisions, up to a batch size of 32. The sample file below displays an explanation of each adjustable parameter in the config file.

NOTE: If users notice significant variation between runs, please adjust the “total_requests” values to be equal to the Max Inference Throughput value of that network. This will ensure that the workload runs for longer than one second, thereby minimizing variation.



```
{
"delayBetweenWorkloads": 5, …Value (integer). Adds specified seconds of delay between workloads
"isDemo": false, …Value (true/false). Shows a demo of Resnet50 workload if enabled. Default value = false.
"iteration": 3, …Value (integer). Specifies the number of times this configuration file must run.
"module": "Deep-Learning", …Not modifiable. Specifies that the workloads are deep-learning technology.
"runtype":"performance", …Not modifiable. Specifies that the workloads test system performance.
"workloads_config": […Specifies the list of workloads to run. Each item in between { } is a workload.
     {
        "batch_sizes": [
            1         …Value (integer). Specifies the batch size of the workload.
        ],
        "concurrent_instances": 1, …Value (integer). The number of concurrent instances to run.
        "hardware": "cpu", …Value (cpu/gpu). Specifies the target hardware to run the workload on.
        "name": "ResNet-50", …Value (ResNet-50/SSD-MobileNet-v1). Specifies the name of the workload.
        "precision": "fp32", …Value (fp32/fp16/int8). Specifies the precision model to use.
        "runtype":"performance", …Specifies that the workloads test system performance.
        "total_requests":64 …Value (integer). Specifies the number of requests to run.
     },
     {
        "batch_sizes": [
            1
        ],
        "concurrent_instances": 1,
        "hardware": "cpu",
        "name": "SSD-MobileNet-v1",
        "precision": "fp32",
        "runtype": "performance",
        "total_requests": 155
      }

   ]
}
```