QUICK START GUIDE

Intel® Data Center Blocks for Enterprise AI Inferencing

Introduction

Intel® Data Center Blocks for Enterprise AI Inferencing deliver purpose-built server solutions that provide low-latency, high-throughput inference utilizing the power and features of the CPU, not a separate accelerator card. These fully configured systems jumpstart deployment of efficient AI inferencing algorithms on a server solution composed of validated Intel server building blocks.

The Intel® distribution of Open Visual Inference and Neural Network Optimization toolkit (Intel® Distribution of OpenVINO™ toolkit) is recommended to complete your AI Inferencing solution. The Intel® Distribution of OpenVINO™ toolkit is a developer suite that accelerates high-performance deep learning (DL) inference deployments for computer-vision workloads. The toolkit takes models trained in different frameworks and optimizes them for different Intel hardware options in order to provide high flexibility for deployment. The toolkit also quantizes DL models, a process in which the toolkit transforms models from using large, high-precision 32-bit floating point numbers, which are typically used for training, to using 8-bit integers. Swapping out floating-point numbers for integers leads to significantly faster AI inference with almost identical accuracy.[1] The toolkit can convert and execute models built in a variety of frameworks, including TensorFlow*, Apache MXNet*, and any framework supported by the Open Neural Network Exchange* (ONNX*) ecosystem. Intel® Data Center Blocks for Enterprise AI Inferencing also provide large amounts of memory to enable inferencing against larger—and more—models simultaneously.

Hardware Components

Intel® Data Center Blocks for Enterprise AI Inferencing combine Intel compute, storage, and networking hardware within a performance-optimized server system which allows for quick deployment of a fast AI inferencing solution.

Intel® Server S2600WF Product Family

1U and 2U rack mount system options based on the Intel® Server Board S2600WF product family delivers high compute performance, large memory capacity, and outstanding power efficiency

Intel® Xeon® Scalable Processor Family

2nd Generation Intel® Xeon® Scalable processors provide the solution with an excellent performance-to-cost ratio and built-in technologies that enhance performance and efficiency for inferencing on AI models:

- **Intel® Advanced Vector Extensions 512 (Intel® AVX-512),** which provides 512-bit instructions that can accelerate performance for demanding workloads and usages like AI inferencing
- **Intel® DL Boost,** which accelerates AI inference by doing in one instruction on the processor what previously took multiple instructions

Intel® SSD Data Center Family

Storage latency and size can be bottlenecks for AI inference. For this reason, the solution uses Intel® SSD DC P4610 drives for data storage. Based on Intel® 3D NAND technology, these enterprise data center SSDs provide a 3.2x lower annualized failure rate (AFR) than hard-disk drives (HDDs).[2]
For data caching, the solution uses the Intel® Optane™ SSD DC P4800X. These SSDs are based on Intel® Optane™ technology and provide extremely low read latency to accelerate inferencing. Intel® Optane™ SSD DC P4800X drives also economically accommodate larger cache sizes, which permits simultaneously caching more and larger DL models to improve inferencing performance.

Intel® Ethernet Connections and Intel® Ethernet Adapters Options

Intel® Data Center Blocks for Enterprise AI Inferencing offer connectivity options that deliver outstanding network throughput and low latency, and are performance-validated to meet high-quality thresholds for data resiliency and service reliability with broad interoperability.

Technology Selections for AI Inferencing

In addition to the Intel hardware foundation used for Intel® Data Center Blocks for Enterprise AI Inferencing, other available Intel technologies deliver further performance and reliability gains:

Intel Distribution of OpenVINO™ toolkit:

A command-line tool based on Python* that imports trained models from popular DL frameworks such as Caffe*, TensorFlow, and MXNet, in addition to any framework supported by ONNX. The OpenVINO™ toolkit quickly deploys applications and solutions that emulate human vision. Based on Convolutional Neural Networks (CNN), the toolkit extends computer vision (CV) workloads across Intel® hardware, maximizing performance. The OpenVINO™ toolkit includes the Deep Learning Deployment Toolkit (DLDT).

Intel® Distribution for Python:

The Intel® Distribution for Python is a binary distribution of the Python interpreter and commonly used packages for computation and data intensive domains, such as scientific and engineering computing, big data, and data science. It accelerates AI-related Python libraries such as NumPy*, SciPy*, and scikit-learn* with integrated Intel® Performance Libraries such as Intel MKL for faster AI inferencing.

Intel® Math Kernel Library (Intel® MKL):

This library optimizes code for future generations of Intel processors with minimal effort. It is compatible with a broad array of compilers, languages, operating systems, and linking and threading models.

Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN):

An open source, performance-enhancing library for accelerating DL frameworks on Intel hardware.

AI frameworks optimized on Intel architecture:

- **TensorFlow:** This Python-based DL framework is designed for ease of use and extensibility on modern deep neural networks and has been optimized for use on Intel® Xeon® processors.
- **Apache MXNet:** This open source, DL framework includes built-in support for Intel MKL and optimizations for Intel AVX-512 instructions.

AI Inferencing Software Technology Installation

Intel® Data Center Blocks for Enterprise AI Inferencing make installation of all prescribed software technologies simple. Intel provides registered users with secure access to a set of deployment scripts integrated within a common deployment package that validates the hardware configuration, installs the OpenVINO™ toolkit and its dependencies, and prepares the system for computer vision inference workloads.

Prerequisites

Before downloading and running the software deployment package, there are several system prerequisites that must be performed and/or confirmed prior to software installation. The software deployment script may fail should any of the following prerequisites not be met.

System Firmware Update:

Your Intel server system may not have the latest system firmware installed. To ensure the latest security patches are installed, and for best system performance and reliability, Intel highly recommends that the system firmware, including: BIOS, BMC Firmware, ME Firmware, and FRU & SDR data be updated to the latest revisions available. The latest available system update package (SUP) for your Intel Server System can be downloaded from the following Intel Web Site:

https://downloadcenter.intel.com/

Under Server Products, select the update package for the following Intel server product family:

Intel® Server Board S2600WF Product Family

Follow the system update instructions included with the update package. Once the firmware stack is updated, it can be verified by accessing the <F2> BIOS Setup utility during system POST (Power on Self Test).

System BIOS Settings:

For best performance, it is highly recommended that the <F2> BIOS Setup utility options identified in the following table be configured as specified.

BIOS Setup Utility Menu Path	BIOS Option	Setting	Setting Notes
Advanced / Processor Configuration	Intel Hyper-Threading	**Disable**	**Required**
Advanced / Processor Configuration	Memory Latency Checker (MLC) Streamer	Enable (default)	Recommended
Advanced / Processor Configuration	Memory Latency Checker (MLC) Spacial Prefetcher	Enable (default)	Recommended
Advanced / Processor Configuration	Data Cache Unit (DCU) Data Prefetcher	Enable (default)	Recommended
Advanced / Processor Configuration	Data Cache Unit (DCU) Instruction Prefetcher	Enable (default)	Recommended
Advanced / Processor Configuration	Last Level Cache (LLC) Prefetch	Disable (default)	Recommended
Advanced / Power & Performance/ CPU P State Control	Intel Turbo-Boost	Enable (default)	**Required**
Advanced / Power & Performance	Hardware P State	Enable (default)	Recommended
Advanced / Power & Performance	CPU C State Control	Enable (default)	Recommended
Advanced / Power & Performance	CPU Power and Performance Policy	Balanced Performance (default)	Recommended
Advanced / Power & Performance	Workload Configuration	Balanced (default)	Recommended
Advanced / Power Performance / Uncore Power Management	Uncore Frequency Scaling	Enable (default)	Recommended
Advanced / PCI Configuration	Intel Volume Management Device (Intel VMD)	Enable (default)	Recommended

Supported Operating Systems:

The prescribed software technologies identified for use on the Intel® Data Center Blocks for Enterprise AI Inferencing were validated with the following Linux Operating System:

CentOS* Linux release 7.6.1810

The operating system must be installed onto the system prior to running the software deployment package.

Other Installation Requirements:

1. The ability to install packages from CentOS and EPEL repositories
2. A proxy must be configured (if required) for Centos yum, by adding a proxy directive to **/etc/yum.conf**:

   Example:

   [main]

   proxy=proxy.example.com:1234

Replace the proxy information in the example above with settings appropriate to the local operating environment.

3. A proxy must be configured (if required) by adding the following variables to **/etc/environment**

   Example:

   http_proxy=proxy-url.com:123

   https_proxy=proxy-url.com:124

Replace the proxy information in the example above with settings appropriate to the local operating environment.

Account Setup and Registration

The software deployment package is available by accessing the following Intel® Xeon® Inference Solution web site:

https://registrationcenter.intel.com/en/forms/?productid=3379

A valid user account is necessary to access the site. To register and setup a user account, enter and submit the requested information. Intel will send a registration confirmation email to the specified email address. Retain the email and the provided product serial number, which may be needed to download and/or extract the files from the software deployment package.

Software Deployment Package Installation

1. Once registration has been completed, login to the web site, download the following software deployment file, and copy it to the Intel® Data Center Blocks for Enterprise AI Inferencing system:

   dsgXIRKit1_0.tar.gz

   example:

   scp dsgXIRKit1_0.tar.gz <host IP address>

2. Login to the server.

   example:

   ssh <host IP address>

3. From the command prompt, extract the files from the software deployment file using one of the following command lines:

[ ]# tar xvfz dsgXIRKit1_0.tar.gz

(Note: An End User License Agreement (EULA) will be displayed and will require acceptance before continuing the file extraction process)

Alternate installation option for automation purposes

[ ]# tar xvfz dsgXIRKit1_0.tar.gz --accept-eula

(Note: No End User License Agreement (EULA) will be displayed. The EULA should be reviewed post installation.)

A software installation directory structure will be created and all installation files will be copied into it.

4. At the command prompt, change directories:

[ ]# cd dsgXIRKit1_0

5. At the command prompt, run the software installation script file using the following command line. Ensure environment variables for http_proxy and https_proxy (if required) are set appropriately before attempting to run the installation script file.

[ ]# ./installer.bash install

The installation script will verify all hardware, BIOS, and management settings to ensure a proper operating environment. The software installer will perform all the tasks necessary to configure OpenVINO™ and Intel python. Error messages with instructions will be displayed should any failures occur during the installation process.

The script will display the following message upon a successful installation

OpenVINO™ Installation Successful

6. It is highly recommended that a sample be run to confirm the successful installation of the software. At the command prompt, enter the following **two** command lines.

[ ]# cd /opt/intel/openvino/deployment_tools/demo/

[ ]# CC=/opt/rh/devtoolset-8/root/bin/gcc CXX=/opt/rh/devtoolset-8/root/bin/g++ ./demo_squeezenet_download_convert_run.sh

(Note: the command above is entered as a single line at the command prompt)

The expected output will end as follows:

```
classid probability label
------- ----------- -----
817     0.8363345   sports car, sport car
511     0.0946484   convertible
479     0.0419131   car wheel
751     0.0091071   racer, race car, racing car
436     0.0068161   beach wagon, station wagon, wagon, estate car, beach waggon, station waggon, waggon
656     0.0037564   minivan
586     0.0025741   half track
717     0.0016069   pickup, pickup truck
864     0.0012027   tow truck, tow car, wrecker
581     0.0005882   grille, radiator grille
total inference time: 3.4052890
Average running time of one iteration: 3.4052890 ms
Throughput: 293.6608295 FPS
[ INFO ] Execution successful
###################################################
Demo completed successfully.
```

Learn More

Additional product information can be found at the following Intel web sites:

Intel® Distribution of OpenVINO™ Toolkit:

https://software.intel.com/en-us/openvino-toolkit
http://docs.openvinotoolkit.org/2019_R1.1/_docs_IE_DG_Introduction.html

Intel® Data Center Blocks:

https://www.intel.com/content/www/us/en/products/servers/data-center-blocks.html

Intel® Xeon® Scalable Processors:

https://www.intel.com/content/www/us/en/products/processors/xeon/scalable.html

Intel® SSD Data Center Family:

https://www.intel.com/content/www/us/en/products/memory-storage/solid-state-drives/data-center-ssds.html

Intel® Ethernet Technology:

https://www.intel.com/content/www/us/en/architecture-and-technology/ethernet.html

Intel® Deep Learning Boost:

https://www.intel.ai/intel-deep-learning-boost/#gs.adc5tx

Intel Framework Optimizations:

https://www.intel.ai/framework-optimizations/#gs.adelrb

Intel® Server S2600WF Product Family:

https://www.intel.com/content/www/us/en/support/products/89018/server-products/server-boards/intel-server-board-s2600wf-family.html

https://www.intel.com/content/www/us/en/support/articles/000023750/server-products/server-boards.html?productId=89017&localeCode=us_en

BIOS Setup - Intel® Server Board BIOS Setup User Guide:

https://www.intel.com/content/dam/support/us/en/documents/server-products/Intel_Xeon_Processor_Scalable_Family_BIOS_User_Guide.pdf

[1] Intel. “Lower Numerical Precision Deep Learning Inference and Training.” October 2018. https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training.

[2] Based on initial product AFR of 0.66 percent vs. industry AFR average (2.11%). Source: Backblaze. “Hard Drive Stats for Q1 2017.” May 2017. backblaze.com/blog/hard-drive-failure-rates-q1-2017/.

[3] The Intel® Ethernet 700 Series includes extensively tested network adapters, accessories (optics and cables), hardware, and software, in addition to broad operating system support. A full list of the product portfolio’s solutions is available at intel.com/ethernet. Hardware and software is thoroughly validated across Intel® Xeon® Scalable processors and the networking ecosystem. The products are optimized for Intel® architecture and a broad operating system ecosystem: Windows*, Linux* kernel, FreeBSD*, Red Hat* Enterprise Linux (RHEL*), SUSE*, Ubuntu*, Oracle Solaris*, and VMware ESXi*. Supported connections and media types for the Intel Ethernet 700 Series are: direct-attach copper and fiber SR/LR (QSFP+, SFP+, SFP28, XLPPI/CR4, 25G-CA/25G-SR/25G- LR), twisted-pair copper (1000BASE-T/10GBASE-T), backplane (XLAUI/XAUI/SFI/KR/KR4/KX/SGMII). Note that Intel is the only vendor offering the QSFP+ media type. The Intel Ethernet 700 Series supported speeds include 10GbE, 25GbE, 40GbE.

Performance results are based on testing as of the date set forth in the configurations and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit intel.com/benchmarks.

Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Optimization Notice: Intel’s compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice Revision #20110804

K76836-001
Rev 2