Explore more resources
Altera® Design Hub

altera
An Intel Company

FPGA AI Suite

Version 2024.3 Release Notes

Updated for FPGA AI Suite: **2024.3**

Online Version
Send Feedback

772497
2024.12.05

Contents

1. FPGA AI Suite Version 2024.3 Release Notes

The *FPGA AI Suite Version 2024.3 Release Notes* provide late-breaking information about the FPGA AI Suite including new features, important bug fixes, and known issues.

About the FPGA AI Suite Documentation Library

Documentation for the FPGA AI Suite is split across a few publications. Use the following table to find the publication that contains the FPGA AI Suite information that you are looking for:

Table 1. FPGA AI Suite Documentation Library

Title and Description	
Release Notes Provides late-breaking information about the FPGA AI Suite including new features, important bug fixes, and known issues.	Link
Getting Started Guide Get up and running with the FPGA AI Suite by learning how to initialize your compiler environment and reviewing the various design examples and tutorials provided with the FPGA AI Suite	Link
IP Reference Manual Provides an overview of the FPGA AI Suite IP and the parameters you can set to customize it. This document also covers the FPGA AI Suite IP generation utility.	Link
Compiler Reference Manual Describes the use modes of the graph compiler (`dla_compiler`). It also provides details about the compiler command options and the format of compilation inputs and outputs.	Link
PCIe-based Design Example User Guide Describes the design and implementation for accelerating AI inference using the FPGA AI Suite, Intel® Distribution of OpenVINO™ toolkit, and a Terasic* DE10-Agilex Development Board.	Link
SoC-based Design Example User Guide Describes the design and implementation for accelerating AI inference using the FPGA AI Suite, Intel Distribution of OpenVINO toolkit, and an Arria® 10 SX SoC FPGA Development Kit (DK-SOC-10AS066S) or Agilex™ 7 FPGA I-Series Transceiver-SoC Development Kit.	Link

2. FPGA AI Suite New Features and Enhancements

FPGA AI Suite Version 2024.3 adds the following new features and enhancements:

- The FPGA AI Suite now supports Upsampling and Downsampling (nearest neighbor) with factors of 2 and 4.
- Running the `dla_compiler` command with the `--fanalyze-area` option now produces a `.ptc` file that you can open in the Quartus Power and Thermal Calculator (PTC), for estimating the power consumption of a particular architecture. (The PTC supports power estimation on Agilex 5, Agilex 7, and Stratix® 10 devices only.)
- The Model Analyzer now reports which aux modules can be safely disabled in an architecture when a graph does not require them. Improvements have also been made to the Model Analyzer `.dot` graphs to include more information from the text report files.
- The hardware input layout transform, which was previously released as an early access feature, is now available as a production feature. More information about this feature is provided in the *FPGA AI Suite IP Reference Manual*.

Multilane

- The new `num_lanes` architecture parameter provides parallelism across height for compatible layers.

  Using the `num_lanes` architecture parameter has the following effects:

  — Setting the `num_lanes` parameter scales the PE array in the FPGA AI Suite IP by the given number and provides additional parallelism at the cost of more DSPs and area.

  — The total stream buffer size scales with the `num_lanes` parameter. Because the feature surface of a graph is divided across multiple lanes, Altera recommends adjusting the `stream_buffer_depth` parameter listed in the `.arch` file by the inverse of the `num_lanes` parameter value. For example, a 4-lane architecture with 10k stream buffer depth indicates a 40k total stream buffer size.
- For FPGA AI Suite Version 2024.3, Altera recommends using multilane with the DDR-free inference option.

Design Examples

- The amount of RAM for the Agilex 7 SoC design example has been increased from 1GB to 8GB.
- The root file system size has increased for the SoC design examples as follows:
  - The Agilex 7 SoC design example root file system is now 4 GB.
  - The Arria 10 SoC design example root file system is now 2 GB.
- The following example designs are added in FPGA AI Suite Version 2024.3:
  - Hostless design example on Agilex 5 E-Series 065B Premium Development Kit. This example design relies on JTAG for host-FPGA communication.
  - PCIe-attach design example on Intel FPGA SmartNIC N6001-PL Platform (without an Ethernet controller).
  - PCIe-attach design example on Agilex 7 FPGA I-Series Development Kit (DK-DEV-AGI027RBES).

3. Changes in Software Behavior

- The PCIe-attach and SoC design examples are updated to use the Quartus® Prime Pro Edition software, version 24.3.
- The Yocto project version for the SoC design examples is updated from Nanbield (4.3.4) to Scarthgap (5.0.2).
- Python and Pip are added to the SoCdesign examples OS images.
- The Python benchmark now requires use of the `--input-precision=fp32` option to ensure that the input data is converted to the correct type. The Python benchmark produces an error if this option is not used and the input data type is unsupported.
- The `dla_build_example_design.py` command now requests that the `DISPLAY` environment variable be unset if the value is set to an invalid display. This avoids a situation where the Quartus Prime software may attempt to connect to the invalid display during bitstream compilation, resulting in an error.
- The file path location specified by the `--dump_dir` option of the `dla_compiler` command is now used as the destination location for most of the `.csv`, `.txt`, and `.dot` files created by the compiler. If you do not specify the `--dump_dir` option, these files are placed in the current working directory.
- The architecture parameters for layout transform have changed. The new parameters set maximum values for the for the transform operation while the previous parameter set exact values for the transform operation. Also, output dimensions are no longer specified for the layout transform. Architecture that have layout transformation enabled must be updated to use the new parameters. The old parameters are no longer supported. For details, refer to *FPGA AI Suite IP Reference Manual* .

4. Known Issues and Workarounds

FPGA AI Suite Version 2024.3 has the following known issues:

- When the input layout transform is enabled, the area estimates for some IP architectures can be too low.
- Building the Arria 10 SoC design example on a Red Hat Enterprise Linux system requires Red Hat Enterprise Linux 8.10.
- The area model reports the wrong DSP configuration with Agilex 5 architectures.

  The area model reports the DSP configuration being "Sum of 4 9x9" when it is a DSP tensor mode. This reporting results in a "PTC Import Warning" when you use a `.ptc` file generated from an Agilex 5 architecture.
- The `AGX7_Streaming_Ddrfree_Multilane` architecture file does not have a `pool aux` module. This lack of the `pool aux` module can produce `dla_compiler` errors on graphs that require pooling, for example, ResNet18 and ResNet50.
- If you use the hardware layout transform, then the `max_feature_height` parameter must be greater than or equal to the channel dimension of your graph. This limitation is due to a known issue with the parameter checks that causes the the input graph channel dimension to be compared against the `max_feature_height` instead of the `max_channels` value.

Streaming Flow Known Issues and Workarounds

- Only single-output graphs are supported.
- MobileNet-v2 is known to fail in the streaming flow when no external memory is used.

Multilane

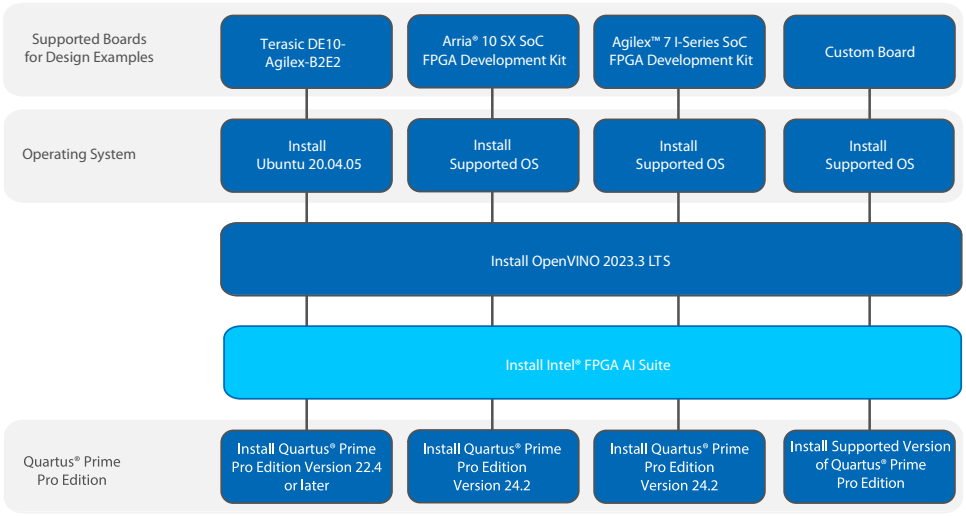
- The `num_lanes` parameter takes only the values 1, 2, or 4.
- When an architecture has a `num_lanes` parameter value greater than 1, the architecture has the following restrictions:
  - All `k_vector` and `c_vector` parameter values in the architecture must be the same.
  - The softmax auxiliary module is not supported.
- Multilane exhibits low performance when used with DDR when a graph has residual connections.

5. FPGA AI Suite Installation Flows

The FPGA AI Suite compiler, IP, and associated tools can be installed separately from the design examples. The design examples add additional hardware and software dependencies.

The following diagram describes an installation flow for the different design example hardware.

Figure 1. Installation Flows for FPGA AI Suite



For details about installing the FPGA AI Suite, refer to *FPGA AI Suite Getting Started Guide*

6. Supported Models

The FPGA AI Suite supports the following pretrained models from OpenVINO Model Zoo. Many other models not listed here are also supported.

Table 2. OpenVINO Model Zoo 2023.3 LTS

Model Zoo 2023.3 LTS path	Model	Framework
`public/mobilenet-v1-1.0-224`	MobileNet V1	Caffe
`public/mobilenet-v2`	MobileNet V2	Caffe
`public/mobilenet-v2-1.4-224`	MobileNet V2	TensorFlow
`public/mobilenet-v3-large-1.0-224-tf`	MobileNet V3	TensorFlow
`public/resnet-50-tf`	ResNet-50	TensorFlow
`intel/unet-camvid-onnx-0001`	UNet	PyTorch
`public/yolo-v3-tf`	YOLO v3	TensorFlow
`public/yolo-v3-tiny-tf`	TinyYOLO v3	TensorFlow
	Yolo v8 (all heads)	PyTorch
`public/squeezenet1.1`	SqueezeNet v1.1	Caffe
`public/i3d_rgb_tf`	Inflated 3D (I3D)	TensorFlow
	Multilayer Perceptrons (MLPs)	

Not all supplied IP configuration `.arch` files support all models. For a detailed list that shows performance/support for these models on each of the supplied IP configuration (`.arch`) files, refer to "Model Performance" in *FPGA AI Suite IP Reference Manual*.

A. FPGA AI Suite Release Notes Archives

For the latest and previous versions of the release notes, refer to FPGA AI Suite Release Notes. If an FPGA AI Suite software version is not listed, the release notes for the previous software version applies.

B. FPGA AI Suite Version 2024.3 Release Notes Revision History

Document Version	FPGA AI Suite Version	Changes
2024.12.05	2024.3	• Added additional known issue.
2024.11.27	2024.3	• Added additional known issues. • Revised wording for some entries.
2024.11.25	2024.3	• Initial release.