Huawei Atlas 300

Technical White Paper

Issue	02
Date	2019-07-29

HUAWEI

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://e.huawei.com

About This Document

Purpose

This document describes the Atlas 300 AI accelerator card (Atlas 300 for short) in detail, including its appearance, performance parameters, and configuration application.

Intended Audience

This document is intended for:

- Huawei presales engineers
- Channel partner presales engineers
- Enterprise presales engineers

Symbol Conventions

The symbols that may be found in this document are defined as follows.

Symbol	Description
	Indicates an imminently hazardous situation which, if not avoided, will result in death or serious injury.
	Indicates a potentially hazardous situation which, if not avoided, may result in minor or moderate injury.
	Indicates a potentially hazardous situation which, if not avoided, may result in minor or moderate injury.
	Indicates a potentially hazardous situation which, if not avoided, could result in equipment damage, data loss, performance deterioration, or unanticipated results. NOTICE is used to address practices not related to personal injury.

Symbol	Description
NOTE	Calls attention to important information, best practices, and tips. NOTE is used to address information not related to personal injury, equipment damage, and environment deterioration.

Change History

Issue	Date	Description
01	2019-07-29	This issue is the second official release.
01	2019-05-10	The issue is the first official release.

Contents

1 Product Introduction

1.1 Overview

The Atlas 300 AI accelerator card is a standard PCIe HHHL card using four HiSilicon Ascend 310 processors. Working with main devices (such as x86 and ARM servers), it implements fast and efficient inference, image identification, and processing.

NOTE

The Ascend 310 is a high-performance and low-power AI chip designed for image recognition, video processing, inference computing, and machine learning. The chip has two built-in AI cores, supporting the 128-bit LPDDR4X DRAM. Each core can provide Ascend 310 with up to 16 TOPS (INT8) computing capability.

1.2 Appearance

Figure 1-1 and **Figure 1-2**shows the appearance of the Atlas 300.

Figure 1-1 Atlas 300 with a half-height ejector lever

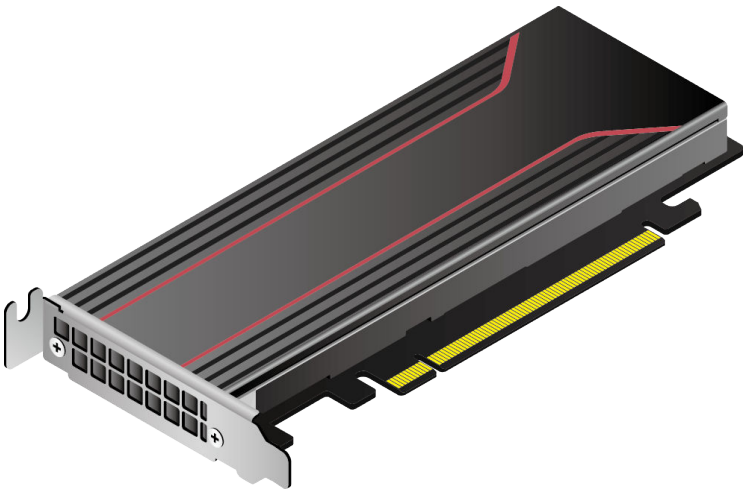
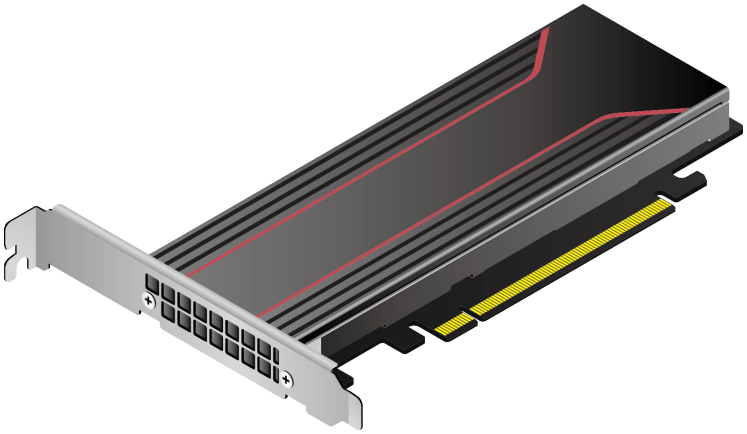


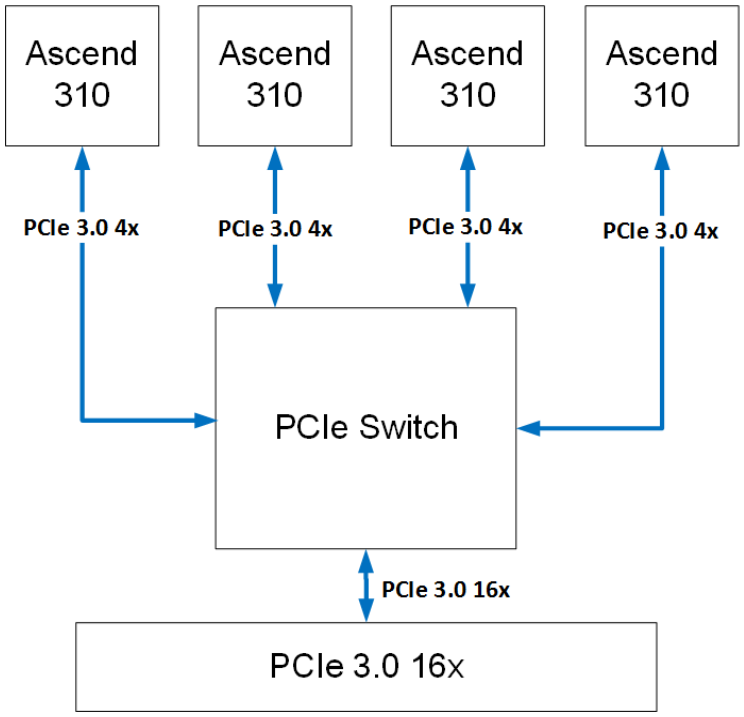
Figure 1-2 Atlas 300 with a full-height ejector lever



1.3 System Architecture

Figure 1-3 shows the system architecture of the Atlas 300.

Figure 1-3 System architecture of the Atlas 300



2 Product Features

2.1 Performance

- Uses four high-performance and low-power Ascend 310 processors. Each can provide Ascend 310 with up to 16 TOPS (INT8) computing capability.
- Supports H.264 and H.265 video encoding and decoding, meeting various video processing requirements.

2.2 Maintainability

- Supports in-band online upgrade, facilitating users' routine maintenance.
- Supports in-band and out-of-band obtainment of device status information such as temperature, voltage, and power consumption, simplifying management with a graphical user interface.
- Provides complete command line management function. Users can use commands to perform routine device management.
- Supports in-band and out-of-band asset management, and provides such information as production date and serial numbers to facilitate asset management.

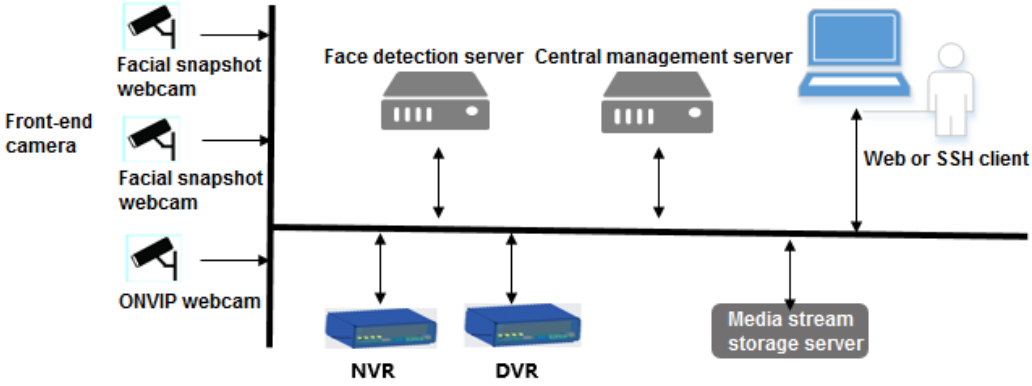
2.3 Typical Application Scenario

The Atlas 300 provides a facial recognition system. It uses the face detection algorithm, face tracking algorithm, face quality scoring algorithm, and high-speed facial recognition algorithm to implement real-time face snapshot modeling, real-time blacklist alignment alerting, and facial image retrieval.

Figure 2-1 shows the architecture of the facial recognition system. The main components include the front-end HD webcams or face snapshot cameras, media stream storage server (optional), intelligent facial analysis server, face alignment search server, central management

server, and client management software. The Atlas 300 is deployed on the intelligent facial analysis server to implement frequency decoding and preprocessing, face detection inference, face alignment (correction), and facial feature extraction inference.

Figure 2-1 Typical facial recognition system architecture



3 Product Specifications

3.1 Basic Specifications

Table 3-1 Specifications of the Atlas 300

Feature	Specifications
Model	Low Profile standard card, supporting full-height and half-height ejector levers
Memory	LPDDR4X 32 GB, 3200 Mbps, total bandwidth 204.8 GByte/s, supporting ECC
AI Computing capability	64 TOPS/INT8, 32TFLOPS/FP16

Feature	Specifications
Encoding and decoding capability	● H.264 hardware decoding, 64-channel 1080p 30 FPS (8-channel 3840 x 2160 60 FPS), YUV420 ● H.265 hardware decoding, 64-channel 1080p 30 FPS (8-channel 3840 x 2160 60 FPS), YUV420 ● H.264 hardware encoding, 4-channel 1080p 30 FPS, YUV420 ● H.265 hardware encoding, 4-channel 1080p 30 FPS, YUV420 ● JPEG decoding capability at 4 x 1080p 256 FPS and encoding capability at 4 x 1080p 64 FPS, up to 8192 x 8192 resolution ● PNG decoding capability at 4 x 1080p 48 FPS, up to 4096 x 4096 resolution
PCIe interface	● x16 lanes, compatible with x8/x4/x2/x1PC ● PCIe Gen3.0, compatible with 2.0/1.0
Power consumption	67 W
Dimensions	169.5 mm x 68.9 mm (6.67 in. x 2.71 in.)
Weight (g)	319 g
Mean time to failure (MTTF)	69.6 years

Table 3-2 Basic specifications of the Atlas 300 software

Feature	Specification
Operating System (OS)	● CentOS 7.4 Kernel Version: 3.10 ● Ubuntu 16.04.03 Kernel Version: 4.4.0-93.116+
Deep learning framework	TensorFlow, Caffe
Compiler	CCE/CCE compiler Tool

3.2 Environmental Conditions

Table 3-3 lists the requirements for the hardware application environment of the Atlas 300.

Table 3-3 Atlas 300 hardware application environmental conditions

Item	Specification
Operating temperature	0°C to 55°C (32°F to 131°F)
Storage temperature	-40°C to +75°C (-40°F to +167°F)
Relative humidity (RH, non-condensing)	5% to 90%
Storage humidity (RH, non-condensing)	5% to 95%
Maximum altitude	3000 m (9842.52 ft.). For altitudes above 900 m (2952.72 ft.), the highest operating temperature decreases by 1°C (1.8°F) for every increase of 300 m (984.24 ft.) in altitude.

3.3 Clock Requirements

The Atlas 300 complies with the standard PCIe card protocol (*PCI Express® Card Electromechanical Specification Revision 3.0*). The card only needs to provide the standard PCIe 3.0 (backward compatible with 2.0 and 1.0) differential clocks. The signal quality meets the PCIe specifications.

3.4 Hot Swap

The Atlas 300 supports orderly hot swap, but does not support surprise hot swap.

3.5 Power Management

The Atlas 300 complies with the standard PCIe standard card protocol (*PCI Express® Card Electromechanical Specification Revision 3.0*). The maximum power consumption of the card is 67 W. The corresponding PCIe slots are required to provide 5.5 A@12 V and 0.5 A@3.3 V standard power supply capabilities.

3.6 Heat Dissipation Specifications

3.6.1 Heat Dissipation Requirements

The Atlas 300 is used in an active heat dissipation environment with fans. It supports bidirectional air intake and air exhaust. The air volume must meet the heat dissipation requirements.

Table 3-4 Heat dissipation requirements of the Atlas 300

Mean Temperature at the Air Intake Vent	Minimum Required Air Volume (CFM)	Voltage Drop/Inch H_2O
55°C (131°F)	6.7	0.53
50°C (122°F)	5.3	0.37
45°C (113°F)	4.4	0.28
40°C (104°F)	3.7	0.21
35°C (95°F)	3.3	0.18
30°C (86°F)	3.0	0.16
Any scenario	3.0	0.16

NOTE

- The minimum required air volume refers to the air volume through the heat sink of the Atlas 300.
- The ambient temperature at the heat sink inlet refers to the mean temperature at the air intake vent of the module.
- The required air volume here is just a recommended value. The air volume and incoming flow temperature provided by different systems for the Atlas 300 may vary, and are determined according to the actual test results.
- When the Atlas 300 is powered on, the module needs air volume for heat dissipation. The minimum required air volume is 3.0 CFM.

3.6.2 Heat Dissipation Specifications

The Atlas 300 supports an air intake vent temperature from 0°C to 55°C (32°F to 131°F). Internal temperature monitoring points are available. Both in-band and out-of-band modes can be used to monitor the Ascend 310 and storage chip in real time to ensure that the temperature of the card is lower than the specification value when the card is in the working status.

Table 3-5 Temperature specifications of key components

Temperature	Ascend 310 Temperature	Storage Chip Temperature
Power-off temperature	106°C (222.8°F)	100°C (212°F)
Underclocking temperature	101°C (213.8°F)	90°C (194°F)
Long-term operating temperature	100°C (212°F)	85°C (185°F)

3.6.3 Over Temperature Protection

The Atlas 300 supports the detection of the junction temperature of key components, such as the Ascend 310 and memory chip, through out-of-band and in-band channels, and also supports the detection of the entire board temperature. The log of the Atlas 300 records information such as the maximum temperature, the count of over temperature, and the total over temperature time.

The Atlas 300's main component Ascend 310 and its storage chip support a maximum air intake vent temperature of 55°C (131°F). Sufficient air volume for heat dissipation and proper temperature control policies are required to ensure reliable operations. Therefore, the Atlas 300 uses the following alarming mechanism:

- Severity 1: major alarm. The major alarm threshold of the Ascend 310 chip is 101°C (213.8°F), and the major alarm threshold of the storage chip is 90°C (194°F). When the chip temperature or operating temperature reaches this value, the firmware will restrict the device performance.
- Severity 2: critical alarm. The critical alarm threshold of the Ascend 310 chip is 106°C (222.8°F), and the critical alarm threshold of the storage chip is 100°C (212°F). When the chip temperature or operating temperature reaches the value of this parameter, the Atlas 300 powers off.

4 Management

The Atlas 300 provides various maintenance and management functions, including in-band management command sets running in the operating system (OS) and out-of-band management functions provided by the BMC.

NOTE

If the AI chip is not loaded with a driver, out-of-band management cannot accurately identify whether the AI chip is faulty and does not report any alarm when the AI chip is faulty. In-band management provides only the function of querying the health status of the AI chip. If an upper-layer service requires that an alarm be reported in real time when the AI chip is faulty, the upper-layer service needs to call the related interfaces in the DCMI API module and perform related processing.

4.1 In-band Management

In-band management provides the following functions:

- Online upgrade: The firmware is upgraded to facilitate device maintenance.
- Asset management: Information, such as production date and serial numbers, is provided to facilitate asset management.

  For details about asset management operations, see the *Atlas 300 Command Reference*.
- Log query: O&M personnel can use the log to analyze problems.

  For details about how to query logs, see section "Log Tool" in the *Ascend 310 MindSpore Studio Auxiliary Tools*.

4.2 Out-of-band Management

The Atlas 300 provides the SMBUS interface to support the out-of-band management of servers. The BMC manages assets and monitors the Atlas 300 temperature, voltage, real-time power consumption, and chip status. In addition, the BMC can take over Atlas 300 alarms.

- For details about the out-of-band management function of the Atlas 300, see the BMC user guide.
- For details about the alarms of the Atlas 300, see the BMC alarm reference.

5 Certifications

Table 5-1 Certifications

No.	Country/Region	Certification	Standard
1	Europe	CE	Safety: ● IEC 60950-1:2005(2nd Edition)+A1:2009 +A2:2013 ● EN 60950-1:2006+A11:2009+A1:2010+A12:2011+A2:2013 EMC: ● EN 55032:2012/AC:2013 ● CISPR 32:2012 ● EN 55032:2015 ● CISPR 32:2015 ● EN 55024:2010 ● CISPR 24:2010 ● EN 55024:2010+A1:2015 ● CISPR 24:2010+A1:2015 ● ETSI EN 300 386 V1.6.1:2012 ● ETSI EN 300 386 V2.1.1:2016 ● EN61000-3-2:2014 ● EN61000-3-3:2013 ● EN61000-6-2:2005

No.	Country/Region	Certification	Standard
			● EN61000-6-4:2006+ A1:2010
2	Europe	RoHS	EN 50581: 2012
3	Japan	VCCI	VCCI 32-1

6 Warranty

For details, see the **Maintenance & Warranty**.

A Acronyms and Abbreviations

A	
AI	Artificial Intelligence
B	
BMC	Baseboard Management Controller
C	
CFM	Cubic Feet Per Minute
E	
ECC	Error Checking and Correction
O	
OS	Operating System
P	
PCIe	Peripheral Component Interconnect Express
S	
SMbus	System Management Bus

