Sophon SS1
White Paper

V1.0

Version	Update Content	Release Date
V1.0	-	2017/10/25

CONTENT

1. Overview

Sophon SS1 is a new artificial intelligence (AI) server, which is integrated with Bitmain's latest deep learning accelerator card Sophon SC1/SC1+ and image recognition solution.

Sophon SS1 server is the first inference platform, which is designed to meet the huge needs of deep learning inference. It provides the most convenient hardware infrastructure to help the customers build personalized, intelligent service based on image recognition technology and to bring customers the fastest AI enabling experience.

2. Sophon SS1

Sophon SS1 is a deep learning system which provides a complete set of deep learning solutions for video and image recognition technology. The core components are one or two Sophon SC1/SC1 + deep learning accelerator cards, which are connected to the application system through the PCIE interface. The application system is built on X86 platform, for start-up, storage management and deep learning SDK coordination. SS1's entire system is highly concentrated into a 4-rack-unit (4U) chassis, which integrates power, cooling, network, multi-system interconnection and file system, enabling customers to achieve rapid secondary development or system integration.

2.1 Sophon SS1 specification

CPU	Intel E3 1275V6, 3.8GHz, turbo to 4.2 GHz,15MB Cache, 4 core 8 thread
Number of SC1/SC1+	1 or 2
Number of Sophon NPU	64 per SC1 (128 per SC1+)
Maximum Power	800W, ATX, AC 90~264V
System Memory	2x8G DDR4 for standard, 16G、32G、64G for option
System Storage	Maximum support 6 SATA3.0 hard disk, support RAID0,1, 5, 10
Network	2 x GLAN(by Intel® i210+ Intel® i219)
Number of USB3.0	4
Number of HDMI	1
Operation System	Ubuntu 16.04（64bit） FreeBSD x86 Linux，32-bit or 64-bit
System Weight	10Kg
System Size	380 (L) x 425(W) x 177(H)　mm
Package Size	555(L) x515(W) x260(H) mm
Operating Temperature	0°C ～ 45°C
Working Humidity	10～90%

2.2 Sophon SS1 figure

2.3 Sophon SS1 system technology

Sophon SC1 is the latest DL acceleration card from Bitmain, and it is also the first accelerator card equipped with TPU in China， which is designed for deep learning applications and design.

There is one BM1680 on Sophon SC1, which contains 64 NPU, each NPU contains the following configuration:

64 single precision (FP32) EU operation core
512KB SRAM (program visible)

The single-precision peak performance of Sophon SC1 is 2TFLOP / s, and Sophon SC1+ which with two BM1680 reach the single-precision peak performance to 4TFLOP/s.

Sophon SC1 card equipped with 16GB of DDR4 memory, bandwidth up to 83.2GB/s. With the provided SDK from Bitmain, you can further simplify the buffer management of DL workload to achieve better load balancing.

Each Sophon SC1 card has a PCIE3.0 X8 interface, can access large-capacity system DRAM memory, to achieve two-way data exchange and streaming between the card and the system.

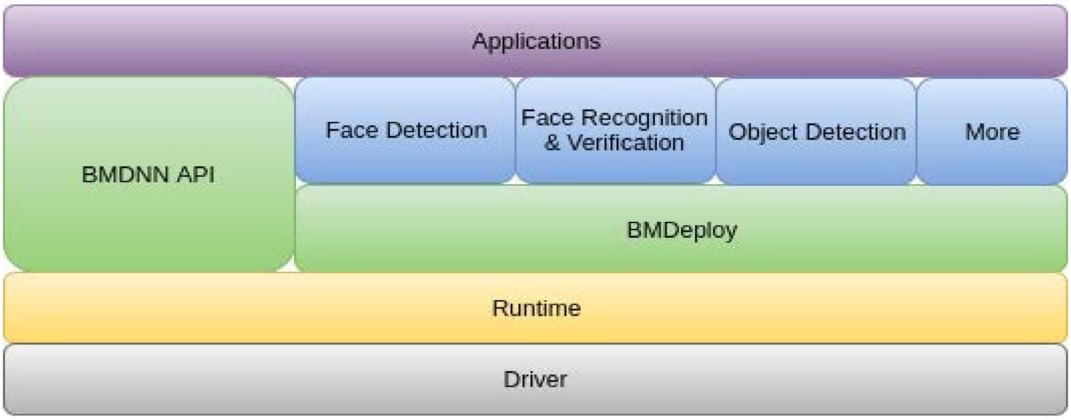
Sophon SC1 card max power consumption is 85W, Sophon SC1+ card max power consumption is 150W, the actual power consumption according to the workload was dynamic. For the convenience of deployment, the board were carried out Active cooling design, to ensure that it can work stably under 0 ℃ ~ 50 ℃ ambient temperature.

3. Sophon SS1 Software

3.1 Software system framework

Bitmain provides a range of software tools from the underlying software to the application layer for the convenience of customers using the SS1 server. The system software design concept is to provide one-stop software services, to enable practitioners to deploy deep learning frameworks and applications on SS1 with minimal setup effort, so that they can achieve the rapid implementation of application services.

Runtime Software Stack



The software architecture includes hardware drivers, Runtime software environments, tools for network model acceleration (BMDNN API) and tools for network deployment (BM Deploy). The most distinctive feature is that there are some function modules for image recognition in this architecture, which could be directly used by the upper applications, including the face detection module, face recognition module, human body detection module and motor/non-motor/people detection and classification module. Bitmain will continue to develop more functional modules. These tools and network models are developed based on Sophon hardware acceleration platform, so you can get the best performance on the SS1 server.

3.2 Component features

3.2.1 Runtime Software Stack

3.2.1.1 BMDNN API

The BMDNN API is an application program interface based on the SC1/SC1+ accelerator card. The SC1/SC1+ accelerator card hardware is abstracted into a common computing interface, so that the upper application software can quickly and easily invoke various computing resources of SC1/SC1+, and the BMDNN API has also been greatly optimized for some of the commonly used calculations to further improve the computational efficiency. The BMDNN API covers all of the available features of SC1/SC1+ and is the most comprehensive way to call.

3.2.1.2 BM Deploy

BM Deploy is responsible for parsing the trained DNN model and scheduling and deploying the inference task. BM Deploy supports two modes: one is BMDNN API mode to achieve DNN network forward, through calling BMDNN API; another is BMNET Compiler mode to achieve DNN network Inference, run BMNET Compiler to generate BM Machine Code.

BMNET Compiler can be used to optimize the DNN network in both Graph level and Backend level, and finally output BM Machine Code, which will be used by the BMDNN Runtime deployment.

3.2.1.3 Module of Face Detection

Face Detection module is used to detect the face in the image and video: Detecting and locating faces in a picture, and returning high-precision face box coordinates. This function module can achieve millisecond level of face detection, can be applied to a variety of real environment.

Usually, face detection is the first step in the analysis and processing of face data, all detected faces can be further analyzed and processed to obtain more face information.

3.2.1.4 Module of Face Recognition

Face recognition module is used to find the known faces by comparing the face features between the captured faces and the known faces. This module can capture lots of face features accurately in a picture which is taken in the actual environment and compare them with the features of known faces rapidly, so that we could identify a known face very accurately and quickly though this technology.

3.2.1.5 Module of Pedestrian Detection

The main function of the human body detection module is to detect the human body in the image or video. Through this module, we can detect and locate the human body in the picture, return to high-precision human rectangular box coordinates. Human body detection is the first step in the analysis and processing of the human body. All the detection of the human body can be further analysis of human properties, access to more complete human body information.

3.2.1.6 Module of Vehicle, non-Vehicle, and Pedestrian Detection and Classification

The motor/non-motor/people detection and classification is used to distinguish motor/non-motor/people in a picture. This feature can be widely used in intelligent traffic or smart city, helping to monitor and manage in a variety of traffic environments, such as crossroads, crosswalks and so on.