8. Atlas AI Computing Platform

8.2 Hardware Architecture of Ascend Chips

Cristiano Bacelar de Oliveira

Setembro/2020

Index

Hardware Architecture of Ascend Chips

- Logic Architecture of Ascend AI Processors
- Da Vinci Architecture

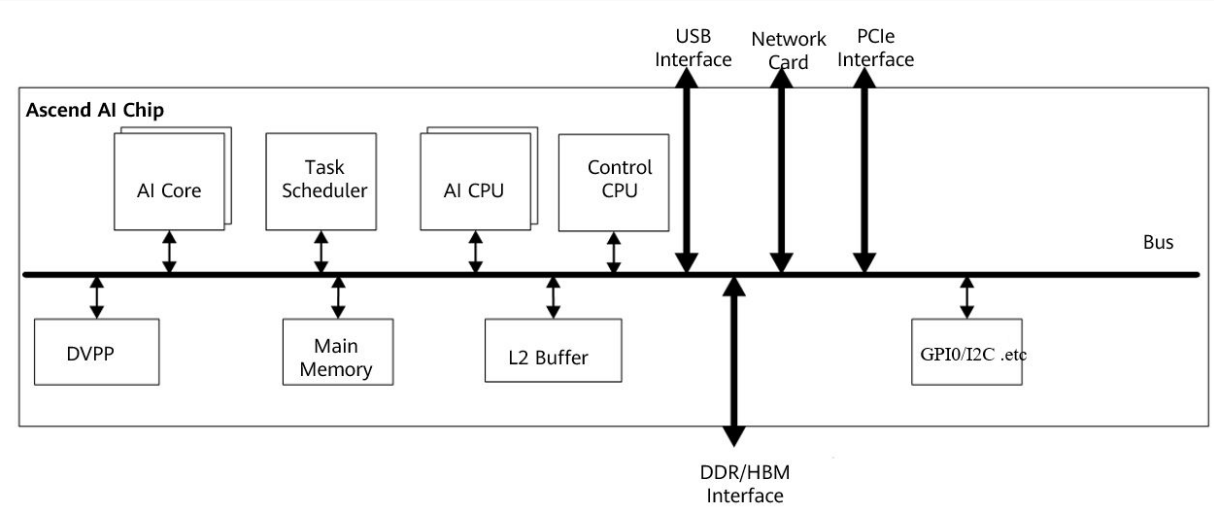
Disclaimer

The following content is heavily based on HCIA-AI Course material by Huawei Technologies Co., Ltd., authored by Shu Xiaodong. Distribution is not allowed.

Logic Architecture of Ascend AI Processors

Logic Architecture of Ascend AI Processors

- **Ascend AI processor consist of:**
  - Control CPU
  - AI computing engine, including AI core and AI CPU
  - Multi-layer system-on-chip (SoC) caches or buffers
  - Digital vision pre-processing (DVPP) module



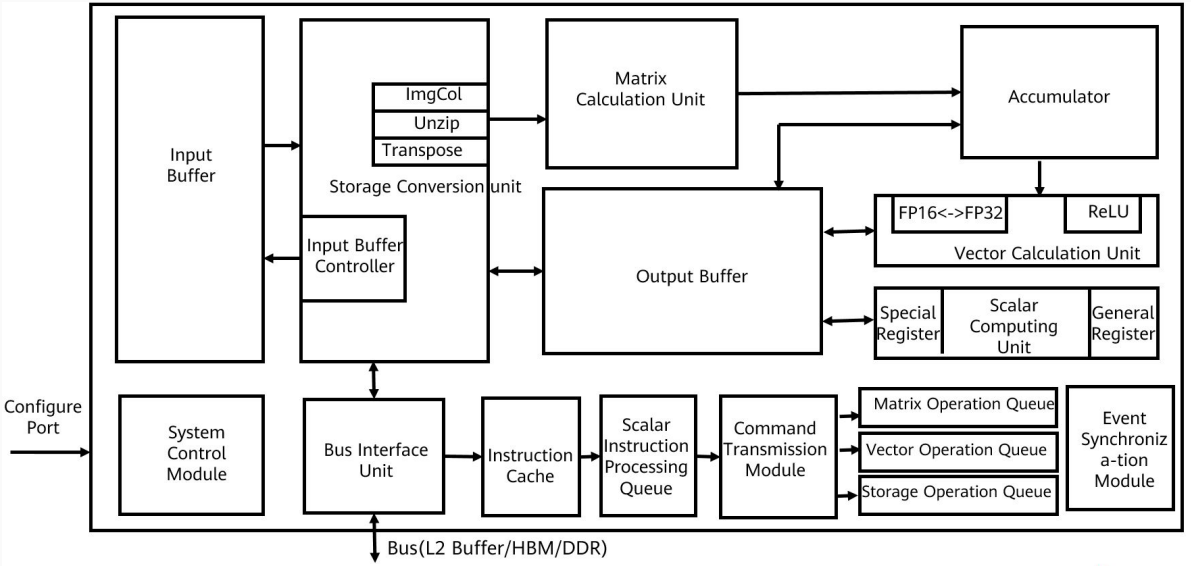
Da Vinci Architecture

Ascend AI Computing Engine - Da Vinci Architecture

- One of the four major architectures of Ascend AI processors is the AI computing engine, which consists of the **AI core (Da Vinci architecture)** and AI CPU. The Da Vinci architecture developed to improve the AI computing power serves as the core of the Ascend AI computing engine and AI processor.

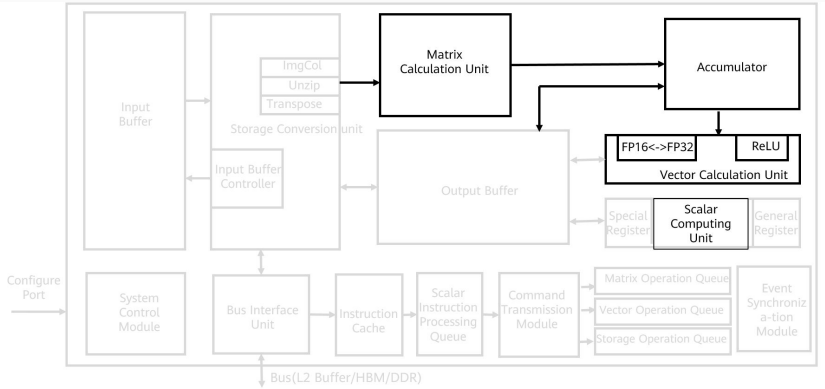
Da Vinci Architecture (AI Core)

- **Main components of the Da Vinci architecture:**
  - Computing unit: It consists of the cube unit, vector unit, and scalar unit.
  - Storage system: It consists of the on-chip storage unit of the AI core and data channels.
  - Control unit provides instruction control for the entire computing process. It is equivalent to the command center of the AI core and is responsible for the running of the entire AI core.



Da Vinci Architecture (AI Core) — Computing Unit

- **Three types of basic computing units: cube, vector, and scalar units, which correspond to matrix, vector and scalar computing modes respectively.**
  - Cube computing unit: The matrix computing unit and accumulator are used to perform matrix-related operations. Completes a matrix (4096) of 16x16 multiplied by 16x16 for FP16, or a matrix (8192) of 16x32 multiplied by 32x16 for the INT8 input in a shot.
  - Vector computing unit: Implements computing between vectors and scalars or between vectors. This function covers various basic computing types and many customized computing types, including computing of data types such as FP16, FP32, INT32, and INT8.
  - Scalar computing unit: Equivalent to a micro CPU, the scalar unit controls the running of the entire AI core. It implements loop control and branch judgment for the entire program, and provides the computing of data addresses and related parameters for cubes or vectors as well as basic arithmetic operations.



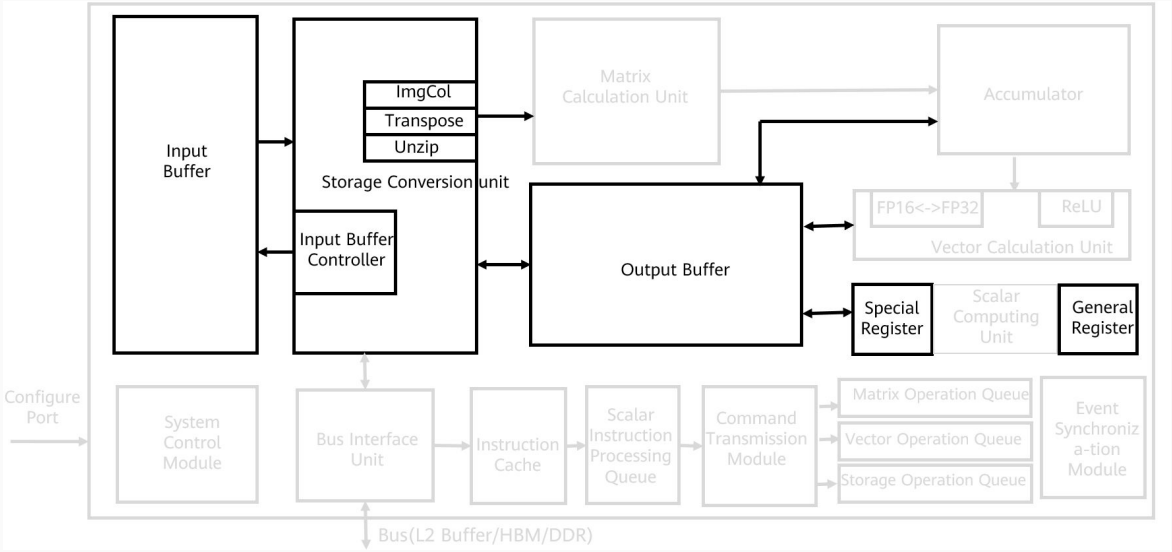
Da Vinci Architecture (AI Core) — Storage System (1)

- **The storage system of the AI core is composed of the storage unit and corresponding data channel.**
- **The storage unit consists of the storage control unit, buffer, and registers:**
  - Storage control unit: The cache at a lower level than the AI core can be directly accessed through the bus interface. The memory can also be directly accessed through the DDR or HBM. A storage conversion unit is set as a transmission controller of the internal data channel of the AI core to implement read/write management of internal data of the AI core between different buffers. It also completes a series of format conversion operations, such as zero padding, Img2Col, transposing, and decompression.
  - Input buffer: The buffer temporarily stores the data that needs to be frequently used so the data does not need to be read from the AI core through the bus interface each time. This mode reduces the frequency of data access on the bus and the risk of bus congestion, thereby reducing power consumption and improving performance.
  - Output buffer: The buffer stores the intermediate results of computing at each layer in the neural network, so that the data can be easily obtained for next-layer computing. Reading data through the bus involves low bandwidth and long latency, whereas using the output buffer greatly improves the computing efficiency.
  - Register: Various registers in the AI core are mainly used by the scalar unit.

HUAWEI

Da Vinci Architecture (AI Core) - Storage System (2)

- **Data channel: path for data flowing in the AI core during execution of computing tasks**
  - A data channel of the Da Vinci architecture is characterized by multiple-input single-output. Considering various types and a large quantity of input data in the computing process on the neural network, parallel inputs can improve data inflow efficiency. On the contrary, only an output feature matrix is generated after multiple types of input data are processed. The data channel with a single output of data reduces the use of chip hardware resources.

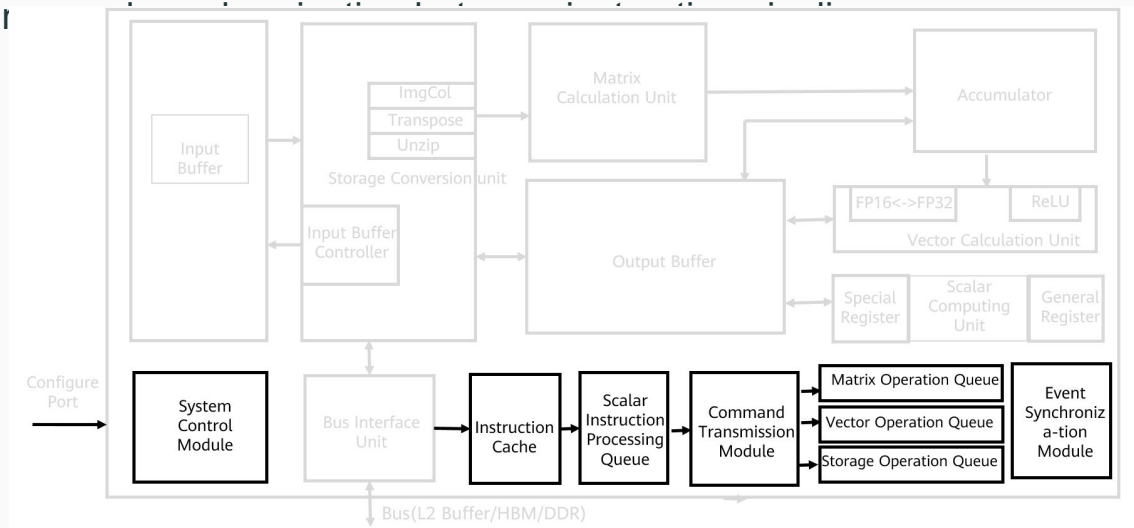


Da Vinci Architecture (AI Core) — Control Unit (1)

- **The control unit consists of the system control module, instruction cache, scalar instruction processing queue, instruction transmitting module, matrix operation queue, vector operation queue, storage conversion queue, and event synchronization module.**
  - System control module: Controls the execution process of a task block (minimum task computing granularity for the AI core). After the task block is executed, the system control module processes the interruption and reports the status. If an error occurs during the execution, the error status is reported to the task scheduler.
  - Instruction cache: Prefetches subsequent instructions in advance during instruction execution and reads multiple instructions into the cache at a time, improving instruction execution efficiency.
  - Scalar instruction procession queue: After being decoded, the instructions are imported into a scalar queue to implement address decoding and operation control. The instructions include matrix computing instructions, vector calculation instructions, and storage conversion instructions.
  - Instruction transmitting module: Reads the configured instruction addresses and decoded parameters in the scalar instruction queue, and sends them to the corresponding instruction execution queue according to the instruction type. The scalar instructions reside in the scalar instruction processing queue for subsequent execution.

Da Vinci Architecture (AI Core) — Control Unit (2)

- Instruction execution queue: Includes a matrix operation queue, vector operation queue, and storage conversion queue. Different instructions enter corresponding operation queues, and instructions in the queues are executed according to the entry sequence.
- Event synchronization module: Controls the execution status of each instruction pipeline in real time, and analyzes dependence relationships between different pipelines to resolve problems of data dependen



Thank You!

Next: 8.3 - Software Architecture of Ascend