8. Atlas AI Computing Platform

8.1 Software Architecture of Ascend Chips

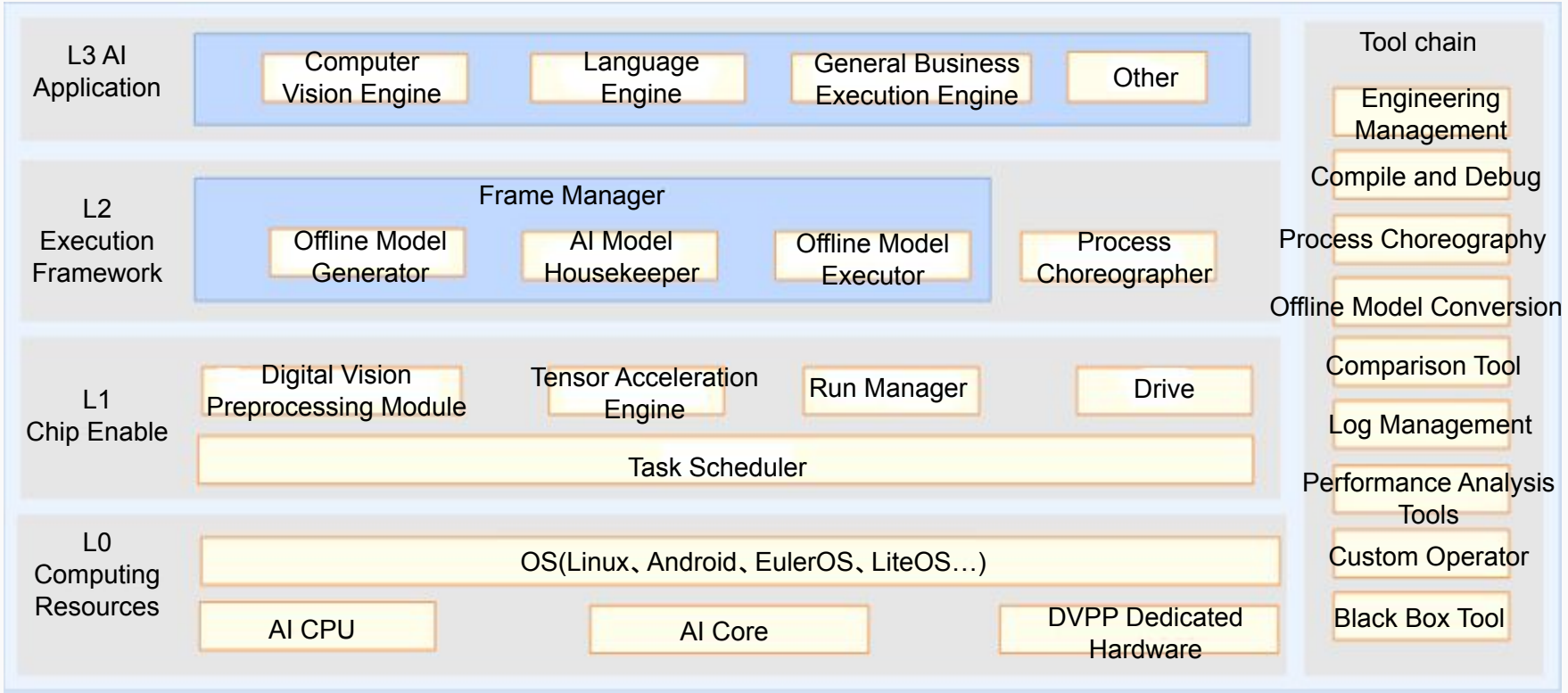
Ricardo Brauner

17/09/2020

1. Overview of AI Chips
2. Hardware Architecture of Ascend Chips
3. **Software Architecture of Ascend Chips**
   - Logic Architecture of Ascend 310
   - Neural Network Software Flow of Ascend 310
   - Data Flowchart of Ascend 310
5. Huawei Atlas AI Computing Platform

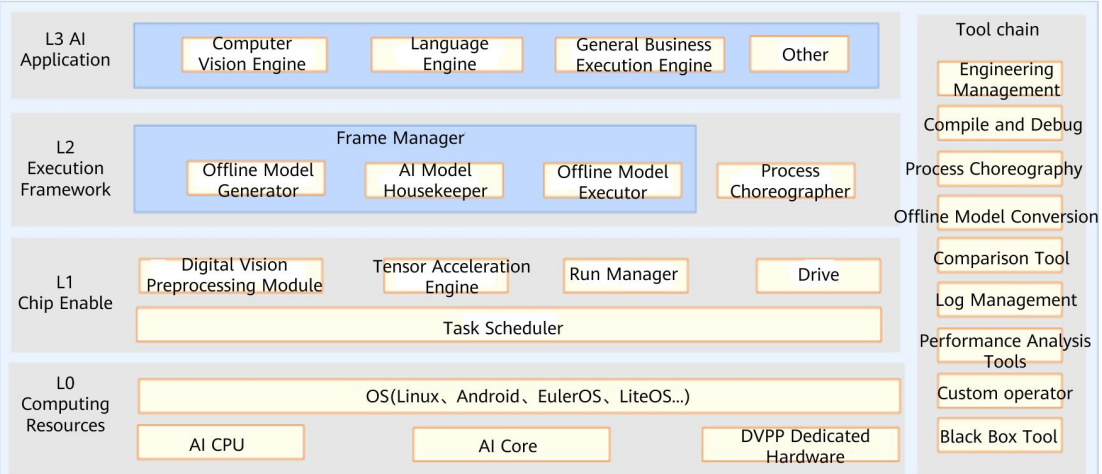
- This section describes the software architecture of Ascend chips, including the logic architecture and neural network software flow of Ascend AI processors.

Logic Architecture of Ascend AI Processor Software Stack (1)



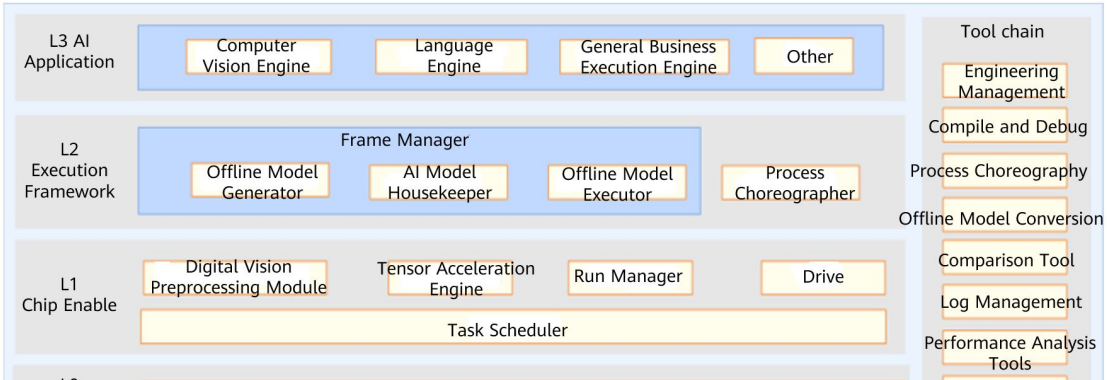
Logic Architecture of Ascend AI Processor Software Stack (2)

- **L3 application enabling layer: It is an application-level encapsulation layer that provides different processing algorithms for specific application fields. L3 provides various fields with computing and processing engines. It can directly use the framework scheduling capability provided by L2 to generate corresponding NNs and implement specific engine functions.**
  - Generic engine: provides the generic neural network inference capability.
  - Computer vision engine: encapsulates video or image processing algorithms.
  - Language and text engine: encapsulates basic processing algorithms for voice and text data.



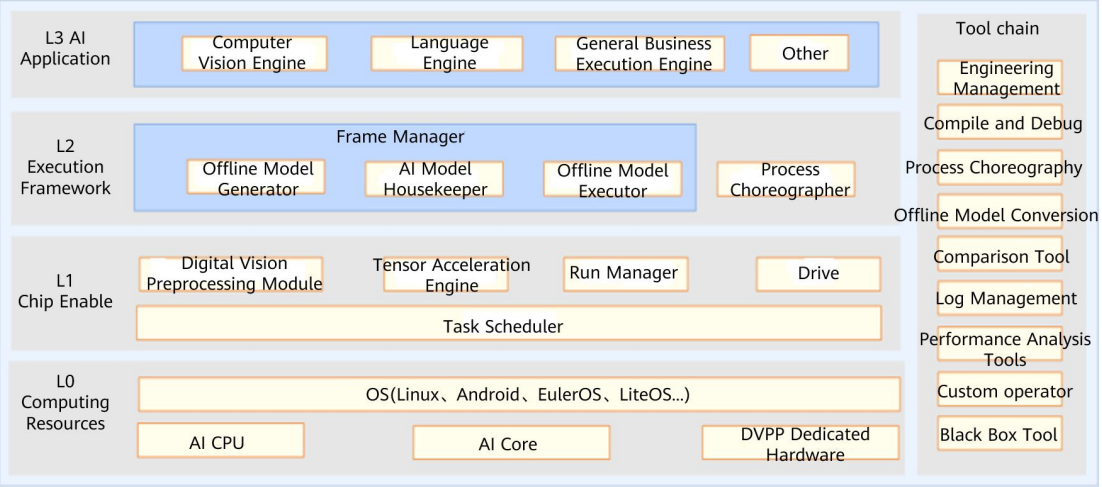
Logic Architecture of Ascend AI Processor Software Stack (3)

- **L2 execution framework layer:** encapsulates the framework calling capability and offline model generation capability. After the application algorithm is developed and encapsulated into an engine at L3, L2 calls the appropriate deep learning framework, such as Caffe or TensorFlow, based on the features of the algorithm to obtain the neural network of the corresponding function, and generates an offline model through the framework manager. After L2 converts the original neural network model into an offline model that can be executed on Ascend AI chips, the offline model executor (OME) transfers the offline model to Layer 1 for task allocation.



Logic Architecture of Ascend AI Processor Software Stack (3)

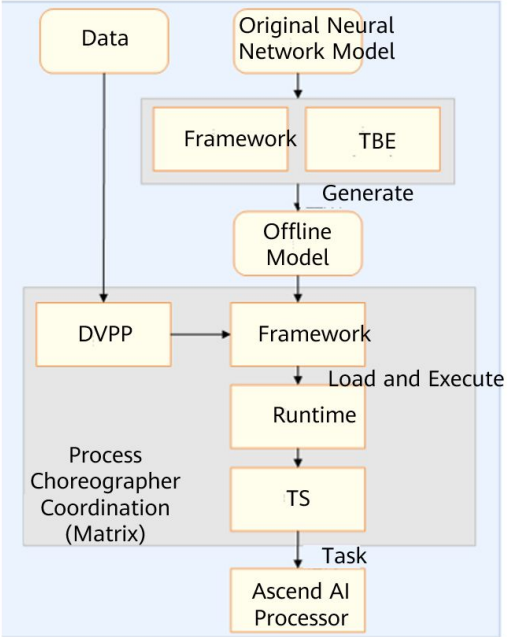
- **L1 chip enabling layer:** bridges the offline model to Ascend AI chips. L1 accelerates the offline model for different computing tasks via libraries. Nearest to the bottom-layer computing resources, L1 outputs operator-layer tasks to the hardware.
- **L0 computing resource layer:** provides computing resources and executes specific computing tasks. It is the hardware computing basis of the Ascend AI chip.



1. Overview of AI Chips
2. Hardware Architecture of Ascend Chips
3. **Software Architecture of Ascend Chips**
   - Logic Architecture of Ascend 310
   - Neural Network Software Flow of Ascend 310
   - Data Flowchart of Ascend 310
5. Huawei Atlas AI Computing Platform

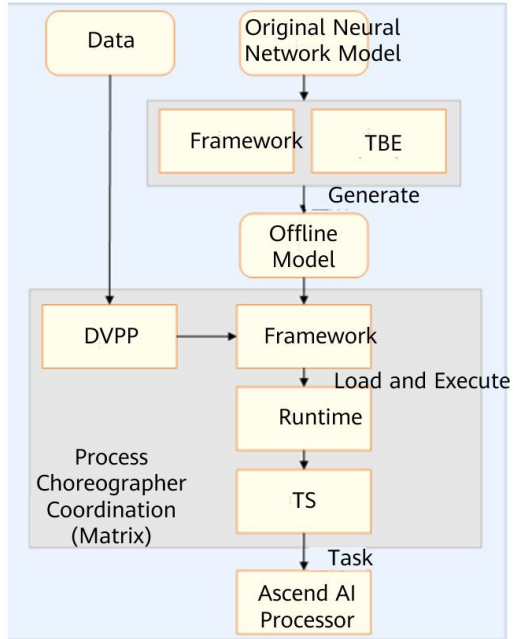
Neural Network Software Flow of Ascend AI Processors

- **The neural network software flow of Ascend AI processors is a bridge between the deep learning framework and Ascend AI chips. It realizes and executes a neural network application and integrates the following functional modules.**
- **Process orchestrator:** implements the neural network on Ascend AI chips, coordinates the whole process of effecting the neural network, and controls the loading and execution of offline models.
- **Digital vision pre-processing (DVPP) module:** performs data processing and cleaning before input to meet format requirements for computing.
- **Tensor boosting engine (TBE)**: functions as a neural network operator factory that provides powerful computing operators for neural network models.

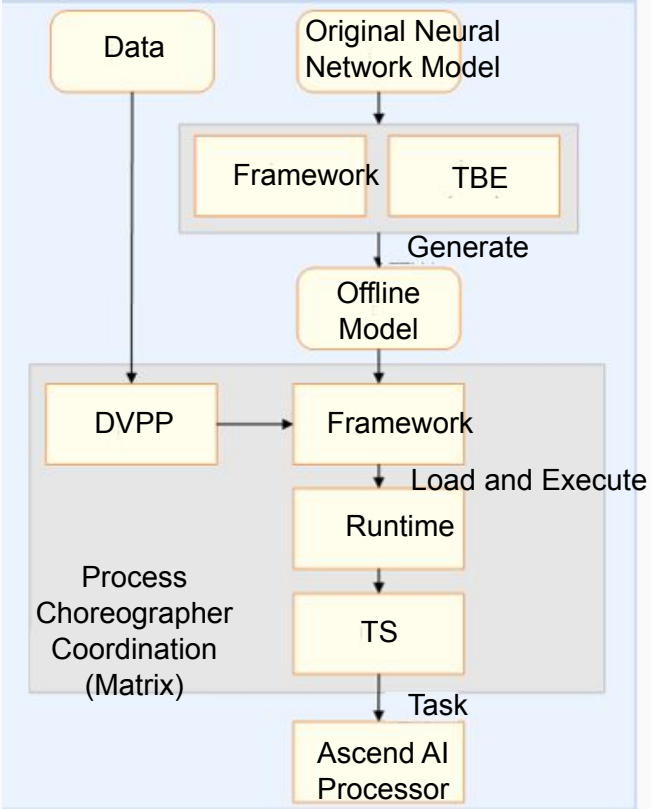


Neural Network Software Flow of Ascend AI Processors

- **Framework manager:** builds an original neural network model into a form supported by Ascend AI chips, and integrates the new model into Ascend AI chips to ensure efficient running of the neural network.
- **Runtime manager:** provides various resource management paths for task delivery and allocation of the neural network.
- **Task scheduler:** As a task driver for hardware execution, it provides specific target tasks for Ascend AI chips. The operation manager and task scheduler work together to form a dam system for neural network task flow to hardware resources, and monitor and distribute different types of execution tasks in real time.



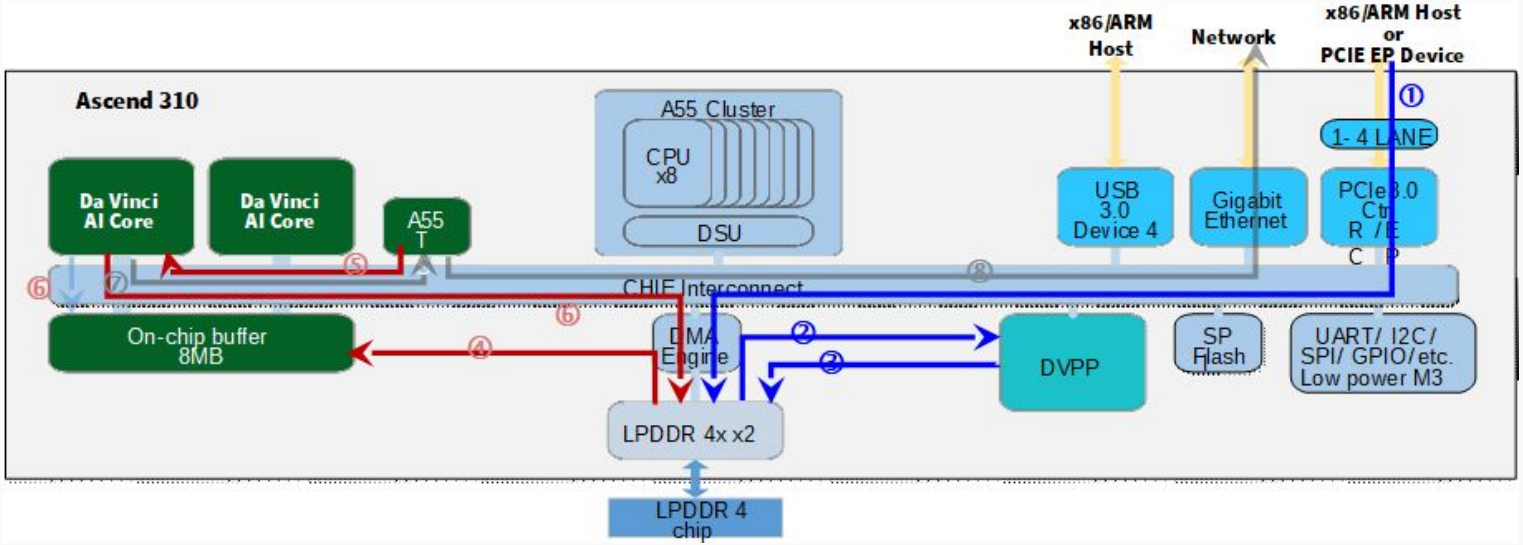
Neural Network Software Flow of Ascend AI Processors



1. Overview of AI Chips
2. Hardware Architecture of Ascend Chips
3. **Software Architecture of Ascend Chips**
    - Logic Architecture of Ascend 310
    - Neural Network Software Flow of Ascend 310
        - Data Flowchart of Ascend 310
5. Huawei Atlas AI Computing Platform

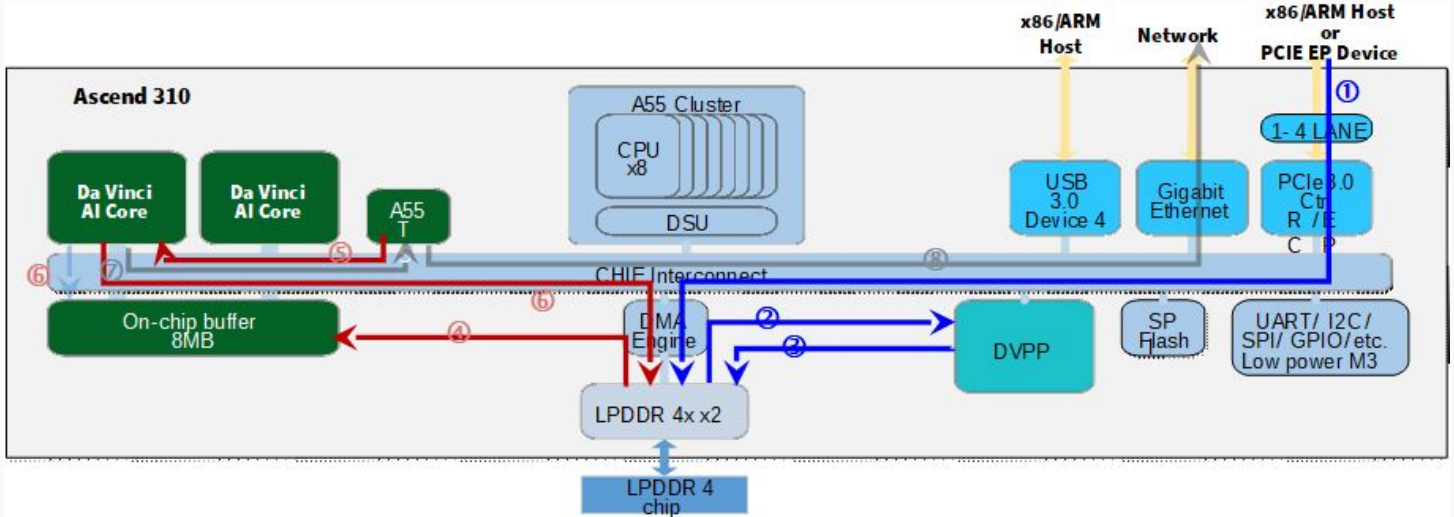
Data Flowchart of the Ascend AI Processor — Facial Recognition Inference Application (1)

- **Camera data collection and processing**
  - Compressed video streams are transmitted from the camera to the DDR memory through PCIe.
  - DVPP reads the compressed video streams into the cache.
  - After preprocessing, DVPP writes decompressed frames into the DDR memory.



Data Flowchart of the Ascend AI Processor — Facial Recognition Inference Application (2)

- **Data inference**
  - The task scheduler (TS) sends an instruction to the DMA engine to pre-load the AI resources from the DDR to the on-chip buffer.
  - The TS configures the AI core to execute tasks.
  - The AI core reads the feature map and weight, and writes the result to the DDR or on-chip buffer.



Data Flowchart of the Ascend AI Processor — Facial Recognition Inference Application (3)

- **Facial recognition result output**
  - After processing, the AI core sends the signals to the TS, which checks the result. If another task needs to be allocated, the operation in step ④ is performed.
  - When the last AI task is complete, the TS reports the result to the host.