DaVinci: A Scalable Architecture for Neural Network Computing
AI Computation for All Scenarios

Memory Footprint (GB)

Scalability Requirement

Computation Requirement Range for Applications

- Smart Toy
- IP Camera
- Smart Phone
- IoT
- Drones
- Robotics
- Industrial Embedded AI
- Autonomous Driving
- Smart City
- Intelligent Surveillance
- Cloud AI Inference
- Model Training
- Meta-Learning Algorithm Discovery
- Auto ML
- General Artificial Intelligence

Computation (TOPS)
Huawei AI Processor for All Scenarios

Scalability: One Architecture to Fit All

Memory Footprint (GB)

Computation (TOPS)

D-Tiny
Wearable

D-Lite
Phone

D-Mini
Robot
Wireless

DaVinci
Edge-training

DaVinci Max
Cloud-training
Architecture Overview of DaVinci
Building Blocks and their Computation Intensity

1D Scalar Unit + 2D Vector Unit + 3D Matrix Unit

Full flexibility + Rich & efficient operations + High intensity

<table>
<thead>
<tr>
<th>N</th>
<th>N²</th>
<th>N³</th>
<th>GPU + Tensor core</th>
<th>AI core + SRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>64</td>
<td>512</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>16</strong></td>
<td><strong>256</strong></td>
<td><strong>4096</strong></td>
<td><strong>Area (normalized to 12 nm)</strong></td>
<td><strong>5.2mm^2</strong></td>
</tr>
<tr>
<td>32</td>
<td>1024</td>
<td>32768</td>
<td>Compute power</td>
<td>1.7Tops fp16</td>
</tr>
<tr>
<td>64</td>
<td>4096</td>
<td>262144</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**DaVinci Core**

- **Cube**: 4096\((16^3)\) FP16 MACs + 8192 INT8 MACs
- **Vector**: 2048bit INT8/FP16/FP32 vector with special functions (activation functions, NMS- Non Minimum Suppression, ROI, SORT)
- Explicit memory hierarchy design, managed by MTE
### Micro Architecture Configurations

<table>
<thead>
<tr>
<th>Core Version</th>
<th>Cube Ops/cycle</th>
<th>Vector Ops/Cycle</th>
<th>L0 Bus width</th>
<th>L1 Bus Width</th>
<th>L2 Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Davinci Max</td>
<td>8192</td>
<td>256</td>
<td></td>
<td>A:8192</td>
<td>910: 3TB/s ÷32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B:2048</td>
<td>610: 2TB/s ÷8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>310: 192GB/s+2</td>
</tr>
<tr>
<td>Davinci Lite</td>
<td>4096</td>
<td>128</td>
<td>Match Execution Units</td>
<td>A:8192</td>
<td>38.4GB/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Not bottleneck</td>
<td>B:2048</td>
<td></td>
</tr>
<tr>
<td>Davinci Tiny</td>
<td>512</td>
<td>32</td>
<td></td>
<td>A:2048</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B:512</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Set the</td>
<td>Minimize vector</td>
<td>Ensure this is not a bound</td>
<td>Scarce, limited by NoC, avoid bound where possible</td>
<td>performance baseline</td>
</tr>
</tbody>
</table>

- **Set the performance baseline**: Setting the baseline is crucial for optimization. Ensure performance is not bound by vector efficiency.
- **Minimize vector bound**: In this configuration, the vector unit is not bottlenecked. Ensure this is not a bound to avoid performance degradation.
- **Scarce, limited by NoC, avoid bound where possible**: The NoC (Network on Chip) is a limiting factor, so avoid setting the bound in this configuration to prevent performance issues.
Overview of the Software Stack

- **Level 3 Library (written by novice programmer)**
- **Level 3 Compiler (mathematical programming model)**
- **Level 2 Library (written by skilled programmer)**
- **Level 2 Compiler (parallel/kernel programming model)**
- **Level 1 Library (written by expert)**
- **Low Level 1 Compiler (Intrinsic C) (Architecture defined programming)**

- **Instruction Set Architecture**
- **GPU**
- **NPU**
- **TBE LIB**
- **TVM/XLA**
- **TBE**
- **Cuda/OpennCL**
- **TIK LIB**
- **TIK**
- **CCE Lib**
- **CCE C**
Putting All This Together

- User Program: AI model description
  - Tensorflow
  - Pytorch
  - Mindspore

- Operator Library, User defined operators
  - Graph Engine (L2: Architecture independent)
  - CANN (L1: Architecture dependent)
  - Task scheduler
  - Davinci Core
  - Davinci Core
  - Davinci Core

- User program AI model using familiar frameworks
- Extends operator library when necessary
- The tasks are executed in a single node, or over a network cluster
DaVinci AI Processors and Products
Mobile AP SoC: Kirin 990 contain D-lite version NPU

CPU 8 - Core, NPU 2+1 Core, GPU 16-Core, 2G/3G/4G/5G Modem, ISP 5.0, LPDDR 4X, UFS 3.0 / 2.1 HiFi Audio, 4K HDR Video, Security Engine

World’s 1st 5G SoC Powered by 7nm+ EUV
World’s 1st 5G NSA & SA Flagship SoC

World’s 1st 16-Core Mali-G76 GPU
World’s 1st Big-Tiny Core Architecture NPU
Mobile AP SoC: Kirin 990 contain D-lite version NPU
Ascend AI Processor: 310 and 910

**Ascend 310**
High Power Efficiency

- Ascend-Mini
  - Architecture: DaVinci
  - FP16: 8 TeraFLOPS
  - INT8: 16 TeraOPS
  - 16 Channel Video Decode – H.264/265
  - 1 Channel Video Encode – H.264/265
  - Power: 8W
  - Process: 12nm

**Ascend 910**
High Computing Density

- Ascend-Max
  - Architecture: DaVinci
  - FP16: 256 TeraFLOPS
  - INT8: 512 TeraOPS
  - 128 Channel Video Decode – H.264/265
  - 1 Channel Video Encode – H.264/265
  - Power: 300W
  - Process: 7+ nm EUV
Huawei AI Solutions For Inference

Atlas
AI Computing Platform

Atlas 200
AI Accelerator Module

Atlas 200 DK
AI Developer Kit

Atlas 300
AI Accelerator Card

Atlas 500
AI Edge Station

Now meet the market
**Ascend 310 Specification: D-mini version**

**SPECIFICATIONS**

<table>
<thead>
<tr>
<th>Description</th>
<th>Architecture</th>
<th>AI co-processor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>Up to 8T @FP16</td>
<td>Up to 16T@INT8</td>
</tr>
<tr>
<td>Codec</td>
<td>16 Channel Decoder – H.264/265 1080P30 1 Channel Encoder</td>
<td></td>
</tr>
<tr>
<td>Memory Controller</td>
<td>LPDDR4X</td>
<td></td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>2*64bit @3733MT/S</td>
<td></td>
</tr>
<tr>
<td>System Interface</td>
<td>PCIe3.0 /USB 3.0/GE</td>
<td></td>
</tr>
<tr>
<td>Package</td>
<td>15mm*15mm</td>
<td></td>
</tr>
<tr>
<td>Max Power</td>
<td>8Tops@4W, 16Tops@8W</td>
<td></td>
</tr>
<tr>
<td>Process</td>
<td>12nm FFC</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** This is typical configuration, high performance and low power sku can be offered based on your requirement.
Applications Built on Atlas 200 DK

Facial recognition

Vehicle detection

Unmanned vehicle

Handwritten text recognition

Robots

Image edge detection

AR shadow generation

HDR

Image marking

Facial recognition

Robotic arm

Classification network

Image processing

Age recognition

Cartoon image generation

Fundus retinal vascular segmentation

Semantic assisted high-precision 3D reconstruction

Facial expression transplant

Protein subcellular position prediction
Huawei AI Solutions For Training: Ascend 910 Server

- 8* Ascend 910
- Davinci Node
- X86 Node
AI Training SoC: Atlas 910 contains D-max version
## Ascend 910 Specification

<table>
<thead>
<tr>
<th>SPECIFICATIONS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>AI co-processor</td>
</tr>
<tr>
<td>Performance</td>
<td>Up to 256T @FP16</td>
</tr>
<tr>
<td>Performance</td>
<td>Up to 512T@INT8</td>
</tr>
<tr>
<td>Decoder Codec</td>
<td>128 Channel FHD Video Decoder – H.264/265</td>
</tr>
<tr>
<td>Memory Interface</td>
<td>32GB HBM Gen2</td>
</tr>
<tr>
<td>System Interface</td>
<td>PCIe3.0 x16 &amp; HCCS &amp; 100G RoCE</td>
</tr>
<tr>
<td>Max Power</td>
<td>Up to 350W</td>
</tr>
<tr>
<td>Process</td>
<td>7nm+</td>
</tr>
</tbody>
</table>

*Normalized to 16-bit*
Ascend 910 Cluster

- 2048 Node x 256TFlops = 512 Peta Flops

- 1024~2048 Node Cluster
Development Tool: Mind Studio
Mind Studio —— Atlas 200 DK Tool Chain

- Operator development
- Offline model tool
- Service orchestration tool
- App Development Tool
Service Orchestrator

Drag-and-drop mode: auto-generate codes
Model Manager

Wizard for OMG, configuring AI preprocessing (AI PP) inside of offline model.
AI PP involves image cropping, color space conversion (CSC), and mean subtraction and multiplication coefficient (pixel changing). All these functions are implemented by the AI core.
DVPP —— Brief Review

DVPP (Digital Video Pre-Processor): image/video encoding, decoding, resizing, cropping, format conversion.
VPC Common Operation: Crop+Resize:

Width Stride-aligned width
Real width

Height Stride-aligned height
Real height

Cropped area top left coordinates
Cropping area bottom right coordinates

Aligned output stride width

Aligned output stride height

Top left Coordinates
Bottom right coordinates

Overwritten Area
The OMG adapts to different network files and weight files in various frameworks and converts them into offline files in DaVinci format for the OME to use.

The OMG is independently executed offline on the host (Linux Ubuntu) and is provided to the IDE as an OMG executable program. The OMG can be called by command lines.

<table>
<thead>
<tr>
<th>Open-Source Framework</th>
<th>Input File</th>
<th>Storage Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffe</td>
<td>*.prototxt, network structure file</td>
<td>Protobuf format</td>
</tr>
<tr>
<td></td>
<td>*.caffemodel_weight file</td>
<td>Binary file</td>
</tr>
<tr>
<td>Caffe2</td>
<td>predict_net.pb, network definition file</td>
<td>Protobuf format</td>
</tr>
<tr>
<td></td>
<td>All input data of the init_net.pb network, including all weight data and input data description of the first operator</td>
<td>Protobuf format</td>
</tr>
<tr>
<td>Mxnet</td>
<td>xx-symbol.json, network structure file</td>
<td>JSON format</td>
</tr>
<tr>
<td></td>
<td>xx.params, weight file including the data, data type, and data format of the weight node</td>
<td>Binary file</td>
</tr>
<tr>
<td>Tensorsflow</td>
<td>*.pb: The network model and weight data are in the same file.</td>
<td>Protobuf format</td>
</tr>
</tbody>
</table>
Sample cmd:

```
./omg --model=./alexnet.prototxt --weight=./alexnet.caffemodel --framework=0 --output=./domi

./omg --model=${HOME}/ResNet18_deploy.prototxt --weight=${HOME}/ResNet18_model.caffemodel --framework=0 --output=${HOME}/outdata/ResNet18 --input_shape="data:1,3,224,224" --insert_op_conf=./aipp_con.cfg
```

Common Paras:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>--model</td>
<td>path to original model file</td>
</tr>
<tr>
<td>--weight</td>
<td>path to weight file (Caffe)</td>
</tr>
<tr>
<td>--framework</td>
<td>0: caffe 1: tensorflow 3: pycaffe</td>
</tr>
<tr>
<td>--output</td>
<td>output om file path</td>
</tr>
<tr>
<td>--plugin_path</td>
<td>customized Operator path</td>
</tr>
<tr>
<td>--insert_op_conf</td>
<td>aipp config file path</td>
</tr>
<tr>
<td>--input_shape</td>
<td>input data shape</td>
</tr>
<tr>
<td>--fp16_high_prec</td>
<td>generate FP16 model</td>
</tr>
<tr>
<td>--mode</td>
<td>0: generate davinci model 1: om model to json 3: only pre-check</td>
</tr>
</tbody>
</table>

For other parameters, please refer to Atlas200DK model conversion guide.
MindSpore: All-scenario AI computing framework

**MindSpore**

- Unified APIs for all scenarios
- Automatic differentiation
- Automatic parallelization
- Automatic tuning
- MindSpore intermediate representation (IR) for computation graphs
- On-device execution
- Pipeline parallelism
- Deep graph optimization
- Device-edge-cloud cooperative distributed architecture (for deployment, scheduling, communications, etc.)

**Processors:** Ascend, GPU, CPU

**Easy development:** AI Algorithm As Code

**Efficient execution:** Optimized for Ascend, GPU support

**Flexible deployment:** On-demand cooperation across all scenarios
ModelArts: Full-pipeline model production services

- Supports full pipeline – From data collection and model development to model training and deployment

- 4,000+ training tasks per day (total of 32,000 training hours)
  - Visual tasks: 85%; audio tasks: 10%; ML tasks: 5%

- 30,000+ developers
Software + hardware co-optimization: Stronger performance

**AI computing challenges**

- **Complex computing**
  - Scalar, vector, and tensor computing
  - Hybrid precision computing
  - Parallelism between data augmentation and minibatch computing
  - Parallelism between gradient aggregation and minibatch computing

- **Diverse computing power**
  - CPUs, GPUs, and Ascend processors
  - Diverse computing units: scalar, vector, and tensor

**MindSpore**

- **Framework optimization**
  - Pipeline parallelism
  - Cross-layer memory reuse

**Software + hardware co-optimization**

- On-device execution
- Deep graph optimization

- ResNet 50 V1.5
- ImageNet 2012
- Based on optimal batch sizes

<table>
<thead>
<tr>
<th>Mainstream training chip + TensorFlow</th>
<th>Ascend 910 + MindSpore</th>
</tr>
</thead>
<tbody>
<tr>
<td>965 (Images/Second)</td>
<td>1802 (Images/Second)</td>
</tr>
</tbody>
</table>
Ascend Developer Community: End-to-End Services for Developers

Ascend developer portal
https://ascend.huawei.com
Unified channel for open capability release and operation

One-stop support services
Technical documents, online resources, development tools, video tutorials, support services, interactive communities, and information release

Developer-centric experience
Provides rich, friendly, easy-to-use, and quick portal experience for developers
We also already started in Canada:

- SFU: 2020 Spring Term Computer Vision class for Professional Master program;
- MakeUofT Hackathon: Feb 15-16
Thank you.

Bring digital to every person, home and organization for a fully connected, intelligent world.

Copyright©2020 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

Huawei Confidential