Baidu Kunlun
An AI processor for diversified workloads

Jian Ouyang, 1 (ouyangjian@baidu.com)
Mijung Noh2, Yong Wang1, Wei Qi1, Yin Ma1, Canghai Gu1, SoonGon Kim2,
Ki-il Hong2, Wang-Keun Bae2, Zhibiao Zhao1, Jing Wang1, Peng Wu1,
Xiaozhang Gong1, Jiaxin Shi1, Hefei Zhu1, Xueliang Du1

1Baidu, Inc. 2Foundry Business, Samsung Electronics
The diversified AI applications

Speech
Recognition, generation..

Vision
Classification, detection, Segmentation..

NLP
QnA, recommend..
The diversified AI scenarios

- Cloud Data Center
- HPC
- Smart Industry
- Smart City
Design AI chip products from industry perspectives

• Target at mainstream market

• Try to explore market volume as much as possible

• Need to support AI applications and scenarios as many as possible
But, the challenge

• Large variety of computing and memory accessing patterns
  – Up to thousand operators in mainstream frameworks
  – Mix of tensor, vector and scalar operations
  – With sequential and random memory access

• Rapid change in algorithm and applications

• Developers have high threshold to new hardware
Baidu Kunlun’s product vision

- Large variety of computing and memory accessing patterns
- Rapid change in algorithm and applications
- The high threshold of developers to new hardware

- Generic
- Flexibility
- Usability and programmability
- High performance
The history of Baidu Kunlun

- 2010: Kickoff SDA Project
- 2014: Hotchips 2014 SDA
- 2016: Hotchips 2016 SDA-II
- 2017: Hotchips 2017 XPU
- 2019: Baidu Kunlun Tapeout
- 2020: Deployment

- Move from FPGA to ASIC
- Evolve from full customization to full programmability

- SDA: software-define Accelerator
- XPU: the X processor unit for diversified workloads
- Baidu Kunlun: the name of Baidu first AI chip, Kunlun is the famous mountain in China
The overview of Baidu Kunlun

- Samsung Foundry 14nm, 2.5D PKG
- 2 x HBM, 512GB/s
- PCIE 4.0 x 8
- 150W, 256Tops
The overview of Baidu Kunlun board

<table>
<thead>
<tr>
<th>Model</th>
<th>Baidu Kunlun K200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>XPU</td>
</tr>
<tr>
<td>Precision</td>
<td>INT4/8</td>
</tr>
<tr>
<td></td>
<td>FP32</td>
</tr>
<tr>
<td></td>
<td>INT/FP16</td>
</tr>
<tr>
<td>Computing capability</td>
<td>INT8: 256TOPS</td>
</tr>
<tr>
<td></td>
<td>INT/FP16: 64TOPS</td>
</tr>
<tr>
<td></td>
<td>INT/FP32: 16TOPS</td>
</tr>
<tr>
<td>HBM Memory Size</td>
<td>16GB</td>
</tr>
<tr>
<td>HBM Bandwidth</td>
<td>512GB/s</td>
</tr>
<tr>
<td>Host IF</td>
<td>PCIE Gen4.0 * 8</td>
</tr>
<tr>
<td>Processing</td>
<td>14nm</td>
</tr>
<tr>
<td>Thermal Cooling</td>
<td>Passive</td>
</tr>
<tr>
<td>Package</td>
<td>2.5D</td>
</tr>
<tr>
<td>TDP</td>
<td>150W</td>
</tr>
</tbody>
</table>
The overview of Baidu Kunlun architecture

- **XPU v1**, FPGA based: Hotchips 2017
- Customized logic for tensor and vector
- Tiny cores for scalar

- **XPU v2**
- With the same design methodology
- More powerful than FPGA version

SDNN - software-defined Neural Network engine
The overview of Baidu Kunlun architecture

- Two units, each unit has
  - 8GB HBM, 256GB/s
  - 16MB on-chip memory
  - 4 XPU-SDNN and 4 XPU-Cluster

- XPU-SDNN
  - Software-defined Neural Network engine
  - Aims at tensor and vector

- XPU-Cluster
  - Aims at scalar and vector
  - With SIMD Instructions
  - 16 tiny core in one cluster
The overview of Baidu Kunlun software stack

- Support multiple frameworks with graph compiler
  - Paddle Paddle, Tensorflow, Pytorch

- Support new operators by user-written kernels
  - XPU C/C++ programming language

- Deep learning library
  - APIs for common operators used in deep learning network
Inference performance – micro benchmark

M,N,K = 4096
Inference performance – YoloV3

- YoloV3 darknet53, 608
- Baidu Kunlun: int16; T4 : TensorRT-FP16. Both accuracy are the same as FP32
- The accuracy of tensorRT-int8 is 5% ~8% less than FP32. so we use FP16/int16 as benchmark
Inference performance – BERT

- **Bert_Base_Uncased**: 12 layer, heads_num = 12, hidden_size = 768, sequence length = 128
- **GPU**: TensorRT-FP16; Kunlun: Int16

QPS: queries per second

**Batch=12**

**Batch=16**

QPS: queries per second
Inference performance – real models in search engine

Notes: model1 and model3 are NLP models. Model2 is vision model
Inference performance – customized MaskRCNN

- Input size: 920x1120

- K200 was used in a customized machine for smart industry
- Running a series of models including MaskRCNN
Conclusion

• Baidu Kunlun is an AI processor for diversified workloads
  − 256Tops int8 and 64Tops int16/fp16
  − 512GB/s memory bandwidth
  − Samsung Foundry 14nm processing, TDP 150W

• Proven in real applications
  − Large collection of models: NLP, vision, speech and etc.
  − Wide ranging scenarios from data center to big edge

• It is available now!
  − Can be accessed via Baidu Cloud