3.3 Kunlun: A 14nm High-Performance AI Processor for Diversified Workloads

Jian Ouyang, Xueliang Du, Yin Ma, Jiaqiang Liu

Baidu, Beijing, China

2021 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-7281-9549-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSCC42613.2021.9366056

In order to be able to handle a wide range of AI applications, such as for speech, image, language and autonomous driving, it is necessary that an AI accelerator be flexible enough to handle diversified workloads. Baidu Kunlun, an AI chip designed in-house by Baidu, achieves this capability with high programmability, flexibility and performance. Baidu Kunlun was inspired by the XPU architecture [1]. The chip is implemented in Samsung 14nm process technology. Its peak performance is 230TOPS@INT8 at 900MHz and up to 281TOPS@INT8 at 1.1GHz boost frequency. The memory bandwidth is 512GB/s and the peak power is 160W. Baidu Kunlun achieves good performance across various types of workloads. With 900MHz base frequency, the latencies of BERT, ResNet50, YOLOv3 are 1.7×, 1.2× and 2× less than an Nvidia T4 GPU, respectively, with optimizations from TensorRT. Recently, Baidu Kunlun has been deployed in data centers in Baidu to serve many applications. It achieves 1.5-to-3× better performance for several models within the search engine vs. the Nvidia T4.

Figure 3.3.1 shows the top-level building blocks of Baidu Kunlun. The chip has two compute units, each with a dedicated 8GB HBM and 16MB on-chip shared SRAM memory. The on-chip shared memory is scratchpad memory – its presence simplifies the hardware implementation and it provides more flexibility for programmers to specify data layout and data management. There is a high bandwidth NoC between two units with up to 256GB/s, which equals to the bandwidth of one HBM. This enables one unit to access HBM and the on-chip shared memory connected within the other unit. XPU-SDNN, the software-defined neural network engine, and the XPU-cluster are the two computation components. XPU-SDNN is highly optimized for operations on large tensors, such as matrix multiply, convolution, deconvolution, activations and element-wise (see Fig. 3.3.2). This component targets at high TOPS performance and has a power efficient MAC array structure. At the same time, the XPU-cluster is designed for flexibility. It is a manycore design, which provides flexibility similar to general-purpose processors. Both XPU-SDNN and XPU-cluster share the same instruction set, which makes programming easily. A scheduler acts as a system management unit to dispatch computational kernels to corresponding compute engines. With such a hybrid architecture, Baidu Kunlun chip achieves high performance, high power efficiency, as well as high flexibility [2].

Different AI workloads require different numerical precisions, and as such, Baidu Kunlun provides four data types: int8, int16, int32 and float32. With this flexibility, Baidu Kunlun works well for low-precision image processing, as well as with extremely high precision requirements of CTR systems. In summary, there are two key contributions: the hybrid structure provides both high efficiency and high flexibility at the same time, and the unified programming model, which hides the heterogeneous hardware and makes programming easily.

Figure 3.3.3 shows the XPU-cluster architecture. Each XPU-cluster is composed of scalar units and vector units and these operate in parallel. The scalar unit can support basic instructions with an ALU, as well as Special Function Unit (SFU) instructions like log, exp, sqrt, div and pow. The vector unit has 256b parallel data width. Each scalar unit operates with explicit parallelism. It supports flexible load/store operations both for the scalar ALU and for SIMD. The scalar unit is flexible for general-purpose computing; meanwhile, the SIMD unit is powerful for data parallel computing. Therefore, Kunlun can support not only deep learning algorithms, but also traditional machine learning algorithms efficiently.

Considering the typical spatial locality and temporal locality in AI algorithms, there are three memory levels for the XPU-cluster: register file, local memory and shared memory. XPU-cluster adopts scratchpad memory to store temporary data, which can provide programmability and flexibility to programmers. Figure 3.3.4 shows the software stack for Baidu Kunlun. The stack comprises a comprehensive toolchain, grouped into two major components, XTCL and XTDK. XTDK contains a C/C++ compiler and XTCL is an AOT/JIT tensor compiler that has the capability to drive the XPU C++ compiler inside XTDK. All frameworks supported can feed subgraphs or export pretrained model files into XTCL.

The XPU C/C++ compiler supports a data-parallel programming model, which utilizes prefix keywords to declare the location of functions or variables in hardware. XPU C/C++ further allows pointer operations and inline assembly to control hardware directly. XDNN is a fully optimized operator library with advanced math computations, such as BLAS. Users can directly call APIs defined in XDNN to perform tasks. Both inferencing and training are fully supported.

Figure 3.3.5(a) shows a micro benchmark demonstrating peak performance of matrix multiplication. The hardware board (Fig. 3.3.6), called K200, contains a Baidu Kunlun chip operating at 900MHz. It is a full-length full-height PCIE board with a PCIE4.0×8 interface. All the benchmarks in this paper are evaluated on the K200. It achieves 204TOPS in int8 and 49TOPS in int16, which is 88% and 85% of designed peak performance, respectively, and outperforms an NVIDIA T4 with 2.76× and 1.13× improvement in int8 and int16, respectively.

Figure 3.3.5(b) shows the performance of two representative public models. In YOLOv3, convolution dominates the computation. While in BERT, matrix multiplication dominates the computation. Although the computational formats differ, Kunlun is able to support both models with high throughput due to the flexible architecture. Figure 3.3.5(d) shows the advantage of exploiting Kunlun in Baidu's search engine service. All three models are developed after Kunlun tape-out. Among them, Model1 and Model3 are NLP models. Model2 is a vision model. We observe 1.7-to-2.8× improvement in throughput, which further demonstrates Kunlun's high flexibility and performance to support diverse workloads.

In addition to Baidu's own service, Kunlun has also been deployed by external clients. Figure 3.3.5(e) shows the deployment of Kunlun in a smart industry system, which uses proprietary MaskRCNN models to detect defects in products. By using Kunlun, the throughput of an accelerator increases 1.8×. The physical layout is shown in Fig. 3.3.6. Four XPU-SDNNs with L-shaped layouts are placed in the four corners of each compute engine, with the NoC which connects all the compute engines and memories located in the center for congestion. Four XPU-Clusters are crammed into the space between the NoC and XPU-SDNNs for better area utilization. The high-speed HBM PHYs are placed in the middle of the left side. With the careful layout of compute engines and on-chip SRAMs, we achieved area efficiency with high performance.

High-end AI chips using 2.5D-IC technology with high bandwidth memory (HBM) technology have emerged recently [3]. The Kunlun chip is fabricated in the Samsung 14lpp process and packaged with two HBM2 memory dies, achieving up to 512GB/s bandwidth. The Kunlun chip's die size is about 500mm^2 and all compute engines, including the NoC, work at the same 900MHz frequency. High-density metal capacitance also mitigates IR drop/rise effects. In addition, various combinations of decoupling capacitors in both the package and PCB design are applied to minimize the voltage drop to optimize the PI performance. Through tightly optimized for the real applications listed above, the Kunlun chip can be used as an AI training and inference accelerator, offering better power efficiency than a GPU. Figure 3.3.5(c) shows the Kunlun power efficiency for each QPS (query per second) is better than a T4 GPU. The K200 hardware board is easily deployed into a commodity server. A 4U server can accommodate up to 8 K200 cards. Moreover, data can be shared via a PCIE switch on the motherboard.

References:

[1] J. Ouyang et al. " XPU – A Programmable FPGA Accelerator for Diverse Workloads," *IEEE Hot Chips Symp.*, 2017.

[2] J. Ouyang et al. "Baidu Kunlun: An AI processor for diversified workloads," *IEEE Hot Chips Symp.*, 2020

[3] D.U. Lee et al., "A 1.2 V 8 Gb 8-Channel 128 GB/s High-Bandwidth Memory (HBM) Stacked DRAM With Effective I/O Test Circuits", *IEEE JSSC*, vol.50, no. 1, pp. 191-203, 2015.

[4] YOLOv3 model, https://github.com/PaddlePaddle/PaddleDetection/blob/release/0.2/docs/featured_model/YOLOv3_ENHANCEMENT.md

[5] BERT model, https://github.com/PaddlePaddle/models/tree/release/1.6/PaddleNLP/PaddleLARK/BERT

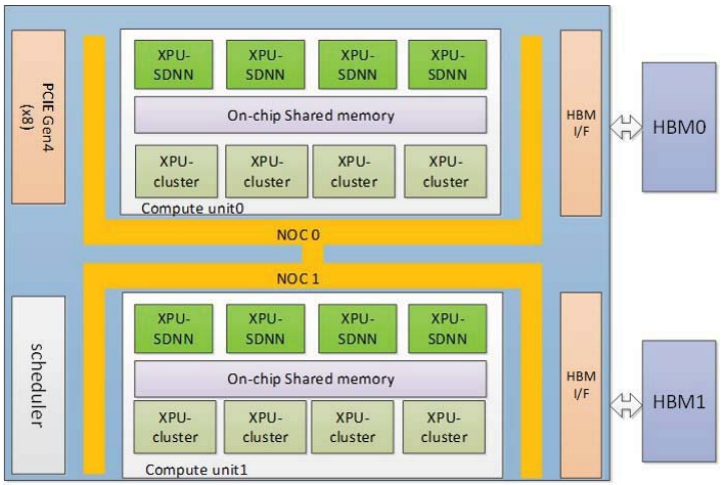


Figure 3.3.1: Baidu Kunlun architecture.

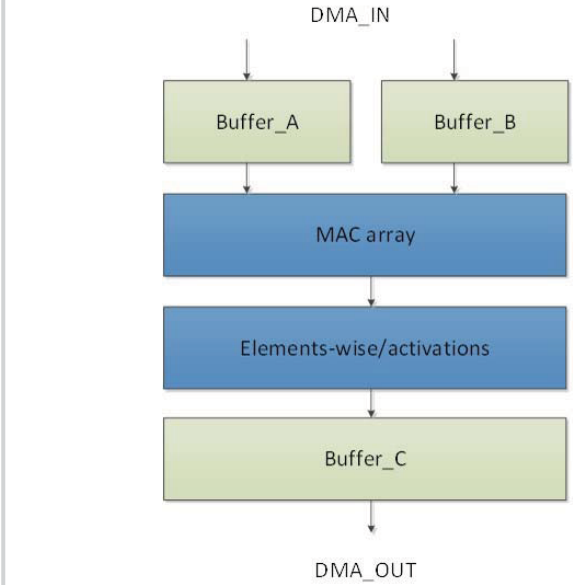


Figure 3.3.2: Building blocks of the XPU-SDNN.

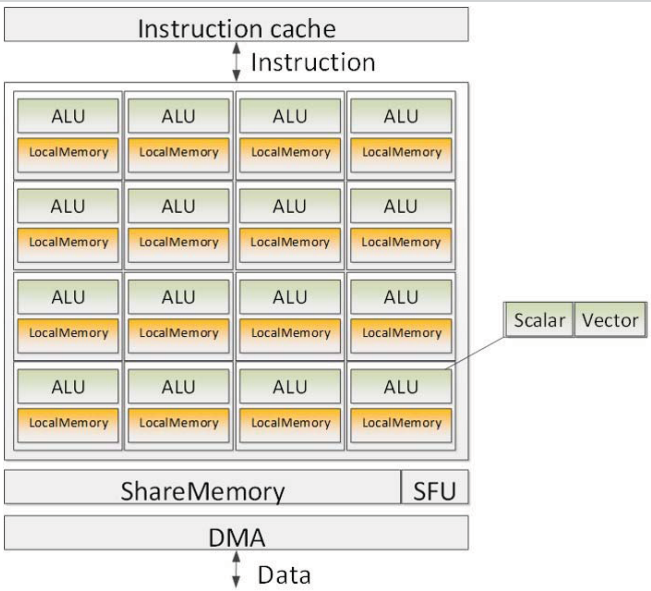


Figure 3.3.3: Building blocks of the XPU-cluster.

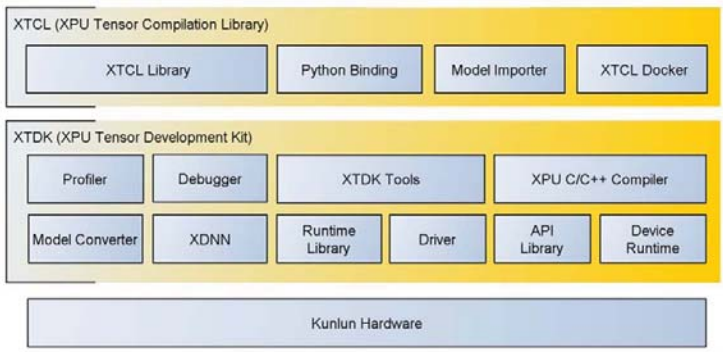
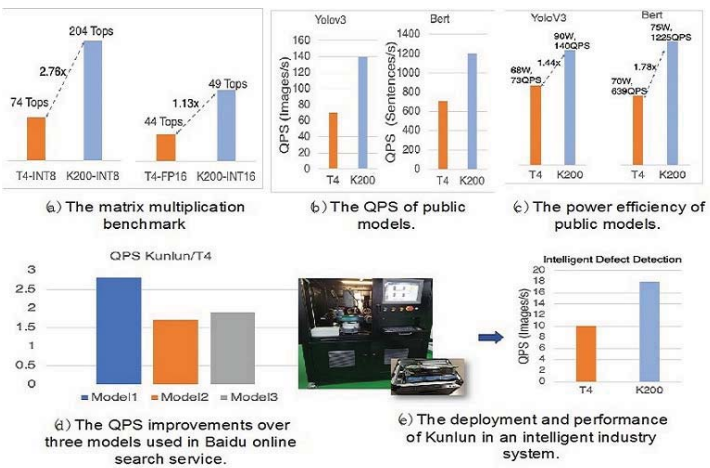


Figure 3.3.4: Software stack of Baidu Kunlun.



(a) The matrix multiplication benchmark

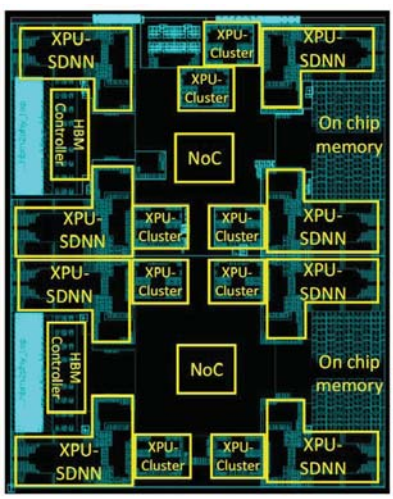
(b) The QPS of public models.

(c) The power efficiency of public models.

(d) The QPS improvements over three models used in Baidu online search service.

(e) The deployment and performance of Kunlun in an intelligent industry system.

Figure 3.3.5: Performance of Kunlun. In (b)-(e), the K200 uses INT16 and the T4 uses FP16. In (b), YoloV3 [4] image size is 608×608. BERT [5] is the BERT-base model.



Process	14lpp
Metal layer	13
Vt	RVT/LVT/SLVT
Track	9T
Supply Voltage	0.85-0.95
Area (mm^2)	504
Speed (GHz)	0.9-1.1

Figure 3.3.6: Physical layout and characteristics.