AMD XILINX

CASE STUDY

KUAISHOU

Kuaishou Provides ASR Services for Large-Scale, Live-Stream Broadcast and Video Apps with Xilinx

Xilinx Alveo™ U50LV Acceleration Cards Enable Low Latency and High Throughput Advantage for Kuaishou's ASR Service

AT A GLANCE:

Kuaishou was established in March 2011 and is headquartered in Beijing China. It is the leading social platform for live-stream broadcast and short-form video content chosen by global users to record and share their daily lives.

Industry: Internet
Headquarters: Beijing, China
Established: 2011
Website: https://www.kuaishou.com

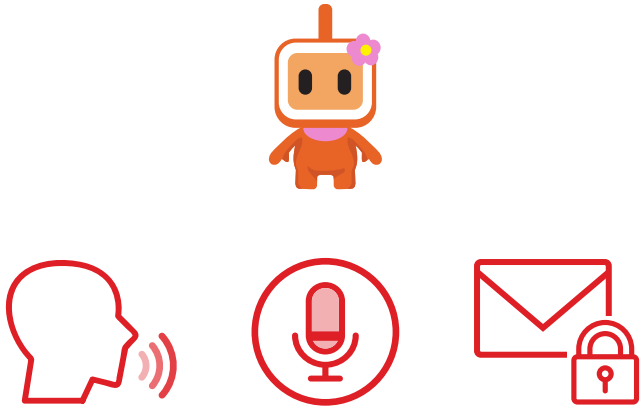

Figure 1. Kuaishou ASR Applications Scenarios

OVERVIEW:

Automatic Speech Recognition (ASR) is a core technology that is used in multiple fields such as e-commerce, short-form video, live-stream broadcast, and more. ASR services is one of the core technologies that has enabled Kuaishou to offer such innovations as the popular Xiaokuai robot voice assistant for the live-broadcast room. It enables voice input, voice search, magic voice expression, real-time and automatic subtitle services, and more to hundreds of millions of users around the world.

With one of the most popular live-steaming and short video apps in the world, Kuaishou's online user base has grown very quickly. The platform averages 320 million daily active users) who are recording and sharing their lives or doing business online. With this high volume of activity, Kuaishou wanted to optimize its current ASR service that not only met the needs of a growing number of users, but also provided the best user experience on Kuaishou's platform by driving continuous innovation, reducing latency, and delivering higher throughout.

The most important performance indicators driving the user experience are latency and the number of concurrent channels in operation. Leveraging Xilinx's Alveo™ acceleration cards, the Kuaishou Heterogeneous Computing Center significantly improved the overall ASR service performance and user satisfaction.

CHALLENGE:

With the previous architecture, the running time of pre-processing modules, such as feature extraction, accounted for about 5% to 10% of the total processing time, the TDNN+LSTM acoustic model (including language models) accounted for about 60% to 80%, and the decoder function used about 15% to 30%. The Kuaishou Heterogeneous Computing Center was actively looking for a more suitable heterogeneous device where it could offload the most time-consuming TDNN+LSTM acoustic model operations in order to optimize its ASR services and achieve better performance and user satisfaction.

Kuaishou's Heterogeneous Computing Center saw three key pain points in the optimization of the streaming acoustic model based on TDNN+LSTM architecture. These included shortening the latency, improving the real-time rate (RTF, Real Time Factor), and increasing the number of concurrent operations (Concurrency). To achieve these, it would need to:

- Reduce latency and provide users with a real-time streaming voice recognition ASR experience;
- Ensure the bandwidth requirements for concurrent processing of massive streaming data;
- Provide flexibility to meet the characteristics of existing multi-service models including running multiple models at the same time, switching multiple models in real time, and ability to meet future model upgrades;
- Reduce unit cost of computing power to achieve lower TCO;
- Meet the high-precision requirements of AI algorithms

Kuaishou evaluated GPUs and discovered that their hardware utilization rate is very low and they could not meet RTF requirements. In addition, the capacity limitations of their on-board SRAM is also unable to meet the high concurrency requirements of a TDNN + LSTM model. As for ASICs, in addition to the hardware utilization problems described above, mainstream ASICs have even more limitations, such as not being able to support the Kaldi architecture, and a fixed point that has challenges meeting the accuracy requirements of ASR optimization.

Based on the aforementioned issues, the Kuaishou team decided that the ideal heterogeneous device platform should be a dedicated platform that could be fully customized and designed to ensure accuracy in compliance with various business standards through software and hardware collaborative design.

SOLUTION:

The company finally selected Xilinx's Alveo U50LV low-voltage accelerator card to optimize its ASR service.

Alveo™ U50 data center accelerator cards are built on Xilinx's high performance UltraScale+ architecture and packaged in an efficient 75-watt, small form factor. They are armed with 100 Gbps networking I/O, and high-bandwidth memory. These features provide Kuaishou's ASR solutions with critical low power, high bandwidth, large SRAM memory, and small-form-factor advantages.

"We believed that the ideal solution for ASR acceleration would be a hardware platform that supported high-bandwidth, large SRAM, and fixed-point inference," Dr.Lingzhi Liu, head of Kuaishou Heterogeneous Computing Center, said. "Xilinx's Alveo FPGA U50LV perfectly matched our requirements."

Heterogeneous Devices	Throughput	Latency	TCO	Power	Flexibility
FPGA	Medium/High	Low	Medium	Low	High
GPU	High	High	High	High	High
ASIC	Very High	Low	Low/Medium	Very Low	Low

Figure 2. Processing Platform Comparison

Combining the company's self-developed fixed-point, general-purpose framework and fixed-point C model, based on Xilinx Alveo U50LV and Vitis™ HLS high-level synthesis and Vitis design flow, Kuaishou optimized its ASR services from multiple key levels including algorithms, systems, software, and hardware with a number of state-of-the-art optimization technologies.

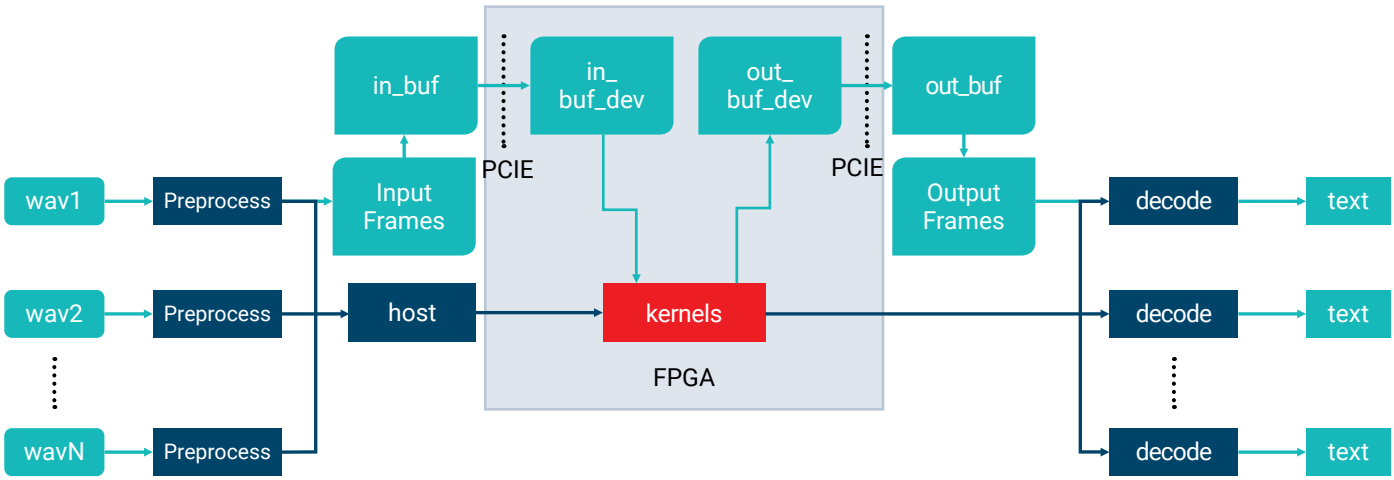

Figure 3. Kuaishou's ASR System Architecture

Here are some of the improvements:

Algorithm level: Leveraged graph fusion, optimization, isomorphism and segmentation technology as well as high-precision quantization technology to effectively compress the model without any compromise of the accuracy and to make the model more effective at leveraging the advantages of FPGA computing;

System level: Developed a general-purpose framework and general host scheduling framework suitable for FPGAs, that supports multiple models, model scalability, automatic deployment, and strong ease-of-use;

Software level: Designed the Batch Mechanism and realized efficient parallel processing such as task-level data transmission and kernel computing based on OpenCL-based task scheduling and load-balancing strategy;

Hardware level: Customized the instruction set architecture based on ultra-long instruction words, designed the compiler, and completed the basic FPGA design efficiently leveraging Xilinx Vitis ™ HLS high-level synthesis optimization technology. Compared with the direct use of hardware description language (such as Verilog HDL), HLS technology uses C / C + + to describe hardware behavior at a higher level of abstraction, which not only achieves the result close to Verilog, but also accelerates the implementation of various optimization technologies and shortens the development time from three months to six weeks.

From the architecture of the optimized system (Figure 3), we can see the code framework (Host) of the scheduling acceleration engine receives the input voice data, and generates the recognized text after pre-processing, neural network inference, and post-processing. The neural network inference part is offloaded to the Alveo card to process.

RESULT:

Leveraging Xilinx's Alveo acceleration card and relative design tools, Kuaishou finally realized a full fixed-point inference hardware acceleration solution for the TDNN+LSTM acoustic model, fully optimized the company's ASR service, and achieved:

1. A significant reduction in the CPU workload and an increase in the business processing capacity of a single server by 7.5 times;
2. Significantly shortened end-to-end video transformation latency up to 37.67%, bringing a greater user experience to live broadcast and short video applications;
3. A 71% (0.29) reduction in total system cost;
4. A significantly shortened development cycle by leveraging OpenCL to achieve seamless integration with existing businesses and the Vitis Design Flow to reduce the design cycle from three months to six weeks.

This is the first successful deployment of an FPGA-based ASR application in China's hyperscale, large-scale live broadcast and short video applications. It demonstrates the strong strength of the technical team behind the various innovative applications of Kuaishou.

Kuaishou's new ASR technology has been widely deployed since mid-2021 and is currently serving hundreds of millions of users through the Kuaishou application.

ADDITIONAL RESOURCES:
Learn More about Xilinx's Alveo U50 acceleration cards
Learn More About Kuaishou

Corporate Headquarters
Xilinx, Inc.
2100 Logic Drive
San Jose, CA 95124
USA
Tel: 408-559-7778
www.xilinx.com

Xilinx Europe
One Logic Drive
Citywest Business Campus
Saggart, County Dublin
Ireland
Tel: +353-1-464-0311
www.xilinx.com

Japan
Xilinx K.K.
Art Village Osaki Central Tower 4F
1-2-2 Osaki, Shinagawa-ku
Tokyo 141-0032 Japan
Tel: +81-3-6744-7777
japan.xilinx.com

Asia Pacific Pte. Ltd.
Xilinx, Asia Pacific
5 Changi Business Park
Singapore 486040
Tel: +65-6407-3000
www.xilinx.com

India
Meenakshi Tech Park
Block A, B, C, 8th & 13th floors,
Meenakshi Tech Park, Survey No. 39
Gachibowli(V), Seri Lingampally (M),
Hyderabad -500 084
Tel: +91-40-6721-4747
www.xilinx.com

Printed in the U.S.A. SF12/21