KUAISHOU

快手基于 Alveo 优化大规模网络直播和短视频自动语音识别服务

赛灵思 Alveo™ 加速卡助力快手实现超低时延实时自动语音识别 (ASR) 服务

公司简介

快手成立于 2011 年 3 月，总部位于北京，是全球用户利用短视频或者直播形式记录和分享日常生活的领先内容社区和社交平台，每天产生上千万条原创新鲜视频

行业：互联网
总部：中国北京
成立时间：2011
https://www.kuaishou.com

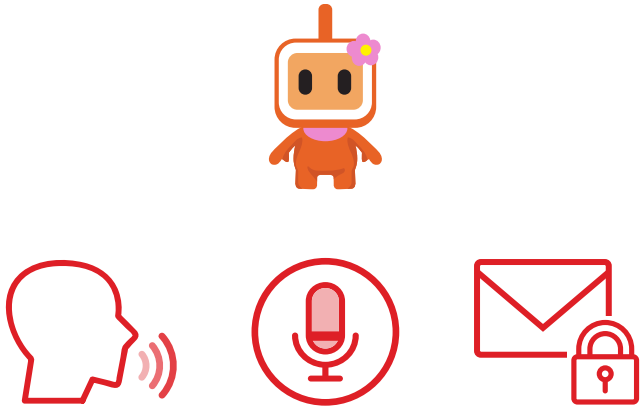

图1：快手 ASR 应用场景

项目概述

自动语音识别 (ASR) 是电子商务、短视频、直播等众多应用的核心技术之一。ASR 在快手有许多的应用场景，是快手 APP、直播、风控、游戏等众多业务的核心功能，尤其是在直播和短视频应用领域。ASR 自动语音识别已经成为快手各种创新服务的核心技术之一。比如广受播主们喜爱的直播间小快机器人（语音助手），快手 APP 语音搜索、直播间语音输入法、语音魔法表情、一甜相机实时字幕，以及快影自动字幕服务等等，为全球数亿用户带来了前所未有的各种创新体验。

作为全球最受欢迎的直播和短视频应用平台之一，快手 2021 年 Q3 财报显示，全球平均有高达 3.2 亿日活用户通过各种终端在快手平台记录和分享他们的生活或体验。面对如此庞大的规模用户和应用场景，快手希望能够优化其 ASR 服务，满足不断增长的客户需求，并为他们提供更好的用户体验。

用户体验最重要的性能指标，就是时延和并发路数。致力于“围绕快手核心业务打造技术护城河”的快手异构计算中心，借助赛灵思 Alveo™ 及相关工具套件，大幅提升了 ASR 整体服务性能和用户满意度，成为国内大规模网络直播和短视频应用场景的 ASR 典范。

项目挑战

在快手之前基于 CPU 框架的处理流程中，特征提取等前处理模块运行时间占比约为 5%～10%， TDNN+LSTM 声学模型运行时间占比约为 60%～80%，而包含语言模型的解码器部分运行时间占比约为 15%～30%。快手异构计算中心，希望找到一个更合适的异构底层器件，将最耗时的 TDNN+LSTM 声学模型转移到这个器件上并进行优化。

快手异构计算中心团队认为，以 TDNN+LSTM 为主结构的流式声学模型优化的关键痛点有三个，那就是是时延 (Latency)、实时率 (RTF，Real Time Factor) 和并发数 (Concurrency)，具体而言需要解决如下问题：

- 缩短时延，为用户提供实时的流式语音识别 ASR 体验
- 提高并发数，保障海量流式数据并发处理的带宽需求
- 提供灵活性及易用性，满足现有多业务模型的特点。如可以同时运行多个模型可以多模型实时任意切换，且能满足未来模型的升级换代
- 降低单位算力成本，实现更低总拥有成本
- 满足 AI 算法的高精度需求

此外，快手对 GPU 进行了评估，发现其硬件使用率比较低，不能满足 RTF 需求，SRAM 容量也无法满足 TDNN+LSTM 模型高并发性的需求。至于主流的 ASIC，除了以上介绍的硬件使用率问题外，还存在不支持 Kaldi 框架，定点实际只有 12bit 等问题，很难满足 ASR 优化在精度上的需求。

综上所述，快手技术团队认为，满足上述需求的理想的异构器件平台，应当是一个可以全定制的专用平台，可以通过软硬件协同设计确保精度符合各种不同业务的标准。

解决方案

快手异构计算中心经过评估后，决定选用赛灵思的 Alveo U50LV 加速器卡来优化 ASR 服务。

Alveo U50 数据中心加速器卡基于赛灵思高性能 UltraScale+ 架构，采用了高效的 75 瓦小型封装，而且配备了 100 Gbps 网络 I/O 和高带宽内存。这些特性为快手的 ASR 解决方案提供了关键的低功耗、高带宽、大 SRAM 内存和小尺寸优势。而 Alveo U50LV (Low Voltage) 则是 U50 系列的低电压版本，和标准电压版本相比，功耗更低，散热要求更少。

“我们认为理想的 ASR 加速解决方案，是可以支持高带宽、大 SRAM 和定点推断的硬件平台，”快手异构计算中心总监刘凌志博士表示：“赛灵思的 Alveo FPGA U50LV 完全符合我们的要求。”

异构器件	吞吐	延时	成本	功耗	灵活性
FPGA	Medium/High	Low	Medium	Low	High
GPU	High	High	High	High	High
ASIC	Very High	Low	Low/Medium	Very Low	Low

图 2：快手各种器件选型比较

结合公司自研的定点通用推理框架和定点 C 模型，快手基于 Alveo U50LV 及赛灵思相关 Vitis HLS 高层次综合及 Vitis Design Flow，从算法、系统、软件和硬件等多个关键层面对 ASR 系统进行了多方位的创新， 应用了多项最先进的优化技术：

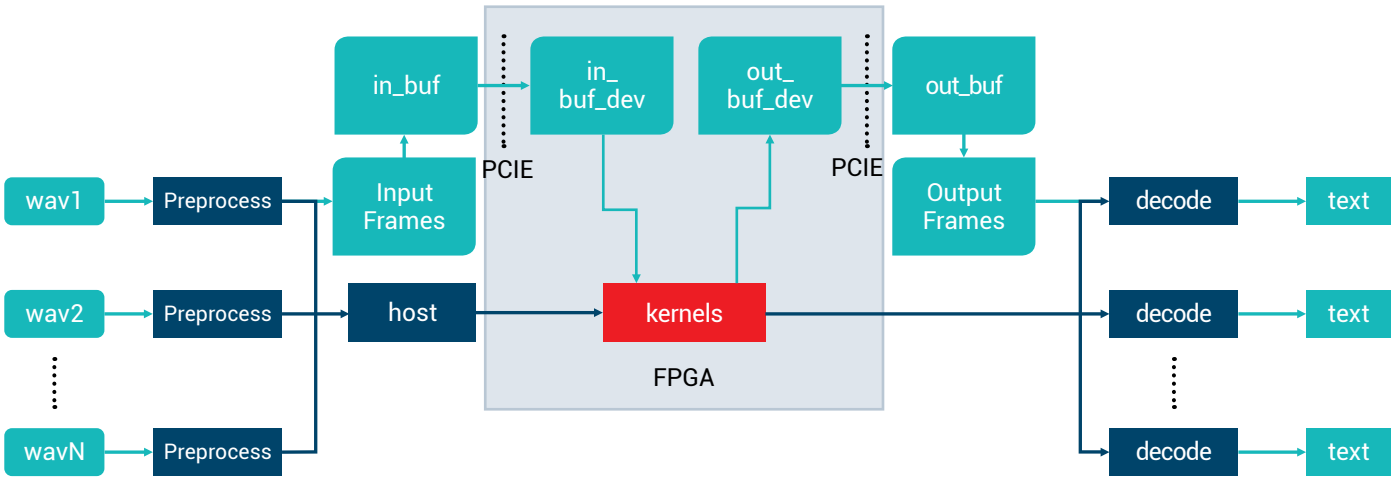

图 3： ASR系统整体架构

算法层面：采用图融合、图优化、图同构、图分割，以及不重训的高精度量化技术，在保证精度的同时，有效的压缩了模型，使之更有利于 FPGA 计算效率的发挥

系统层面：自研通用推理框架及适合 FPGA 的通用 Host 调度框架，支持多模型，模型可扩展，自动化部署，具有很强的易用性

软件层面：设计 Batch 机制，基于 OpenCL 的任务调度及负载均衡策略，实现了任务级的数据传输、kernel 计算等高效并行处理

硬件层面：自定义基于超长指令字的指令集架构，设计编译器，并基于 Vitis™ HLS 高层次综合优化技术，快速完成了高效的 FPGA 底层设计。和直接采用硬件描述语言（如 Verilog HDL）相比，HLS 技术在更高的抽象层次上使用 C/C++ 的语法描述硬件行为，不仅达到了和 Verilog 接近的效果，而且加速了各种优化技术的实现，将开发时间从 3 个月缩短到 6 周

优化后的系统整体架构如图三所示：调度加速引擎的代码框架 (Host) 接收输入的语音数据，经过前处理、神经网络推断和后处理过程，生成识别后的文本。其中黄色部分的神经网络推断过程是卸载到 Alveo 加速卡上来完成的。

设计成效

借助赛灵思 Alveo 加速卡及相关设计工具，快手最终实现了针对 TDNN+LSTM 声学模型的全定点推理硬件加速方案，全面优化了 ASR 服务，实现了：

1. 大幅减轻了 CPU 的工作负载，并发路数提升 7.5 倍
2. 大幅降低了端到端时延，平均缩短达 37.67%
3. 大幅缩减了系统总成本，降至 0.29 （相当于总成本锐减 71% ）
4. 大幅缩短了开发周期。通过采用 OpenCL 实现了与现有业务无缝集成，并借助 Vitis Design Flow 将设计周期从 3 个月减少到 6 周

这是 FPGA 在国内大规模直播及短视频自动语音识别场景落地的首个成功案例，展示了快手各种创新应用背后技术团队强大的实力。2021 年中以来，优化的 ASR 服务已经在快手直播及短视频应用平台广泛部署，目前有数亿用户正在享受其所带来的前所未有的语音识别体验。

更多信息:

点击链接了解赛灵思 Alveo U50 加速卡

点击链接了解快手 Kuaishou

公司总部
Xilinx, Inc.
2100 Logic Drive
San Jose, CA 95124
USA
电话：408-559-7778
www.xilinx.com

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

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

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

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

在美国印刷。SF12/21