TeleVPU: A High-Performance RISC-V Video Transcoding Card

Abstract

With the rapid growth of the video surveillance and streaming media markets, traditional video processing solutions are facing challenges such as inefficiency, high costs, and poor flexibility. To address these issues, China Telecom Research Institute has developed TeleVPU, a video transcoding card based on the RISC-V architecture, which employs a Multi-SoC RISC-V architecture, supports 40 channels 1080P@25fps parallel processing of real-time video streams, and equips with 20 TOPS AI computing power, achieving deep integration of video compression and AI analysis. The accompanying OpenVPU SDK and VPU Server engine provide efficient management interfaces and multi-card collaboration capabilities, significantly enhancing VPU cluster resource utilization. Verification tests have demonstrated that TeleVPU excels in various application scenarios, offering an efficient and cost-effective solution for large-scale video processing and laying the foundation for the construction of intelligent video processing infrastructure.

Introduction

Currently, the video surveillance market is experiencing robust growth. According to statistics, the number of installed surveillance cameras in China has exceeded 600 million, with an annual increase of over 100 million. Looking globally, the video streaming market holds even greater potential, with its market value expected to surpass $213 billion by 2028, representing a compound annual growth rate (CAGR) of 20%.

With the rapid development of video applications, the demand for cloud storage and computing resources by video codec technologies continues to rise. Traditional CPU-based computing methods are inefficient and struggle to handle the processing demands of massive video data. While GPUs offer some processing capabilities, their high costs, low resource utilization, and lack of flexibility make them unsuitable for meeting the needs of large-scale video services, which require efficiency, cost-effectiveness and scalability. In addition, the field of video processing currently faces some technical challenges. On one hand, traditional video transcoding and compression technologies struggle to eliminate content information that is semantically redundant within scenes or lacks user attention. On the other hand, deploying different video accelerators and AI inference cards to handle the tasks of video transcoding compression and AI recognition not only markedly increases deployment costs, but also brings challenges such as cross-platform data transmission and computing resources scheduling.

In this context, the VPU (Video Processing Unit) has emerged as a crucial solution. With its built-in hardware codecs, video stitching, post-processing acceleration, and other powerful features, the VPU significantly reduces server loads, decreases network bandwidth consumption, and enhances computational efficiency. Driven by market demand, ASIC-based VPUs have been commercially deployed, but VPUs based on the RISC-V architecture remain in a technological vacuum. Fundamentally, traditional ASIC-based VPUs rely on performance gains from wafer-level custom designs, which often come with high marginal costs and lengthy architecture iteration cycles. Additionally, their fixed functional modules struggle to meet the growing demands for diversity and elasticity in cloud computing scenarios.

RISC-V, with its extensibility and modular design, offers a broad space for innovation in video transcoding technology. At the same time, developing video transcoding solutions that can adapt to cloud computing scenarios addresses the limitations posed by traditional ASIC-based VPUs, becoming a critical breakthrough point for the industry. It is against this backdrop that we have independently developed TeleVPU, the industry's first video transcoding card based on the RISC-V architecture, specifically targeting high-concurrency video transcoding applications in video clouds and security surveillance.

TeleVPU adopts a multi-core RISC-V architecture, and a single card can support 40 channels 1080P@25fps parallel processing of real-time video streams, with a single channel video transcoding power consumption as low as 1W. It equips with 20 TOPS AI computing power, capable of simultaneously achieving video compression and AI analysis. Its modular design perfectly adapts to standard 2U/4U rack mounted servers, supports 8-card parallel deployment, and can achieve real-time transcoding capability for 320 channels video streams. At the commercial value level, TeleVPU can reduce video size by at least 90%, while ensuring subjective visual quality through AI recognition analysis. The large-scale deployment verification shows that TeleVPU help users save a significant amount of video storage costs. The technological breakthrough of TeleVPU not only provides high-density, high-performance, and cost-effective solutions for ultra large scale video cloud services, but also lays a key technological foundation for building intelligent video processing infrastructure for the AI era.

Architecture and Advantages

As the first video transcoding card based on the RISC-V instruction set, TeleVPU shows exceptional performance advantages through its architectural and capability innovations. In terms of hardware architecture design, TeleVPU employs a Multi-SoC RISC-V architecture, based on multiple domestic RISC-V IP cores and chips, integrating general-purpose RISC-V compute cores, NPUs, and codec accelerators. Its unique hardware design enables each chip in a single card to operate independently. When one chip experiences task blocking or abnormality, other chips can still work normally, ensuring disaster recovery and robustness of massive transcoding tasks. To better integrate into the server cluster operation and maintenance system, TeleVPU is equipped with independent board level BMC management. It is worth mentioning that by customizing RISC-V vector extensions, TeleVPU achieves autonomous control and flexibility at the instruction set level, ensuring technical independence and security from the ground up.

In terms of computational paradigm, TeleVPU adopts a Software-Defined VPU (SD-VPU) architectural approach, breaking through the limitations of fixed functional units in traditional VPU. It provides more powerful programmable capabilities and supports more flexible super-resolution virtualization. With a high memory capacity of up to 40G, TeleVPU's performance can be compared with that of GPUs/TPUs, enabling it to cache multiple AI models and a large number of video files simultaneously. Through its unique transcoding-storage-communication co-design architecture, TeleVPU not only supports parallel processing of multiple video streams but also allows dynamic configuration of various video coding strategies, including H.264/H.265. It achieves an efficient integration of video transcoding and compression with intelligent analysis on a single chip.

In terms of software architecture design, since SD-VPU involves the management of multiple chips, cards, and servers, it introduces a certain level of application complexity for video services. To address these issues, we developed the OpenVPU SDK and VPU Server engine as complements to TeleVPU. The OpenVPU SDK offers efficient management and robust interface support, covering key capabilities such as video transcoding, compression, stitching, AI bitrate control, and hyperparameter tuning. Through highly integrated interface design, it effectively abstracts the underlying complexities, allowing users to focus on core business logic. It facilitates agile development and rapid response to market changes. The VPU Server focuses on multi-card collaboration and unified management, efficiently handling complex scenarios such as single-card multi-chip configurations, multi-server multi-card setups, distributed servers, and multitasking scheduling. By employing intelligent allocation and scheduling mechanisms, VPU Server optimizes the utilization efficiency of VPU cluster resources, significantly enhancing overall performance, which provides strong support for large-scale video processing tasks. Together, these software tools ensure that TeleVPU can deliver high performance and flexibility, enabling users to manage complex video processing workflows with ease and efficiency.

Verification

TeleVPU demonstrates remarkable advantages in video transcoding and compression, particularly in typical application scenarios such as video cloud storage, security monitoring, and video conferencing. It fully meets the urgent demands for efficient video processing in data centers, government and enterprise units, educational and research institutions, and smart transportation sectors. Table 1 presents the code performance and key parameters of TeleVPU.

Table 1: Key parameters

Transcoding and compression performance	Resolution	Channels
	4K	10
	1080P	40
	720P	80
Codec	H.265、H.264	
Image Acceleration	Any resolution scaling, image stitching, sharpening, blurring, noise reduction, watermark/image/text overlay	
AI capability	20Tops@INT8	
Power	60W	

To validate TeleVPU's performance, we conducted large-scale deployment tests. For original videos, TeleVPU achieved over 90% peak video compression while maintaining basic clarity, significantly reducing storage costs in day and night monitoring scenarios, as shown in Figure 1.

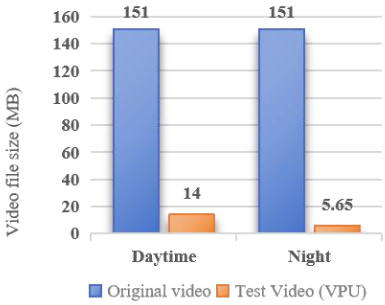

Figure 1: The result of video compression.

With its efficient transcoding and compression capabilities, TeleVPU notably reduces video storage costs across various complex scenarios while maintaining good clarity, which makes it an extremely cost-effective solution for large-scale video processing.