Design of Image Compression Accelerator for Edge Computing

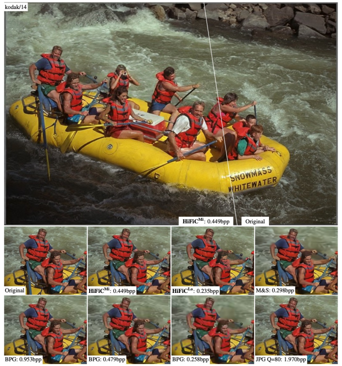
Zhuo Chen
Southwest Jiaotong University, Sichuan

OpenHW2022

AMD
XILINX

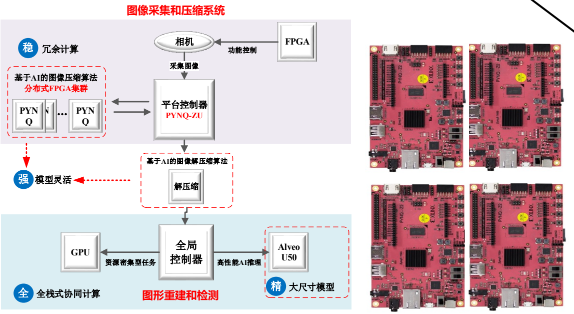
INTRODUCTION

Image compression is to store and transmit image data by removing redundant information under the premise of ensuring image quality and using a low bitrate as much as possible. The image is compressed **by transforming, quantizing, encoding** operations, and then **reconstructed** into a new image. **Deep learning** has been used for image compression since the 1980s, and has been extended to technologies such as MLPs, SNN, CNN, GAN etc. **Distributed FPGA cluster computing architecture** is divided into independent subsystems, each of which can run independently in a distributed structure and communicate through RPC. When some edge devices have failures that are difficult to repair, a board can be **automatically masked** to continue compression tasks.

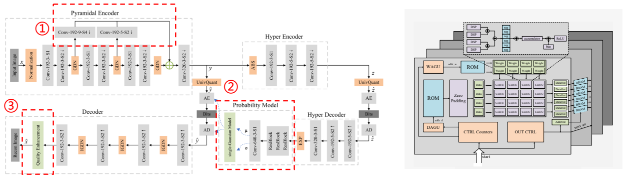


An image compression example

Distributed FPGA cluster computing architecture



The accelerator is developed using SpinalHDL language, and the algorithm is deployed on hardware using distributed FPGA. **The compressed FPGA cluster** mainly completes the compression of images, and the **PYNQ-ZU master controller** is responsible for data distribution and task scheduling, and the image acquisition system. The rebuilt system uses a PC, the GPU can complete resource-intensive tasks, and the **U50** can complete tasks with relatively high real-time performance. This work realizes the **circuit level, module level and board level redundancy design,** significantly improving the stability of the system. **AI algorithm** for image compression promote the performance of the compression, with corresponding **hardware accelerator** (ComACC).



AI image compression model (left)and corresponding hardware accelerator(right)

CREATIVE DESING

On board test by PYNQ-Z2 & Alveo U50

RESULT

The algorithm proposed in this project is superior to the traditional compression algorithm and **existing literature reports in both objective and subjective evaluation indicators**. The **error distribution** of PSNR and MS-SIM implemented by hardware is less than 0.5% and 1.5%. Centralizing multiple computing devices, each completing the same task, alleviates concurrency pressure and single point of failover achieving **high scalability, high performance, low cost, and high availability**. The distributed image compression architecture **compresses the file size by 15 times** and is more **real-time task**.

Performance of image compression (upper) and reconstruction (lower)

Compression Platform	Distributed Image Compression System	Image Collecting System
Data	31.9 k * 8	NA
Resolution	1024*1024	512*512
Delay	400 ms	120 FPS
Compression Ratio	15倍	NA

Reconstruct Platform	GPU reconstruction	GPU inspection	U50 recorgnition
Data	41 k	16.3 M	61.524 M
Resolution	1024*1024	1024*1024	512*512
Delay	10 ms	250 ms	30 ms