IEEE Xplore Full-Text PDF: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&ar...

2022 IEEE/ACM 7th Symposium on Edge Computing (SEC)

Gemini: a Real-time Video Analytics System with Dual Computing Resource Control

Rui Lu*, Chuang Hu†, Dan Wang*, Jin Zhang§
*The Hong Kong Polytechnic University {csrlu, csdwang}@comp.polyu.edu.hk
†Wuhan University hchuchuang@gmail.com *Corresponding Author*
§Southern University of Science and Technology zhangj4@sustech.edu.cn

***Abstract*—Edge-side real-time video analytics systems recognize spatial or temporal events (e.g., vehicle counting) in a video stream. To meet the delay requirement, existing systems in smart edge cameras conduct video preprocessing to filter out unnecessary frames and model inference using appropriately selected neural network (NN) models. Video preprocessing is instruction-intensive computing (IIC) and executed by the CPU of the edge camera, and model inference is data-intensive computing (DIC) and executed by the GPU of the edge camera.**

In this paper, we show that the analytics accuracy of existing systems can largely vary in fields. The root cause is that video analytics applications have different *contents*, which result in *dynamic* IIC and DIC workloads. Unfortunately, intelligent cameras in fields have *fixed* CPU and GPU resources and cannot effectively adapt to workload dynamics. We develop Gemini, a new real-time video analytics system enhanced by a dual-image FPGA. The newly developed dual-image FPGAs can be pre-configured with two FPGA images with a key advantage of negligible image switching time. We thus pre-configure one CPU image and one GPU image and elastically multiplex the dual CPU-GPU resources in the *time* dimension. The Gemini system design requires both hardware and software revisions. We overcame a challenge that the application development on different dual-image FPGAs is hardware-dependent. We develop a new abstraction of hardware functions to make the Gemini system hardware-agnostic. It is also a challenge to adapt to the dynamic workloads and optimize video analytics accuracy. We develop a bandit learning approach to capture content dynamics and conduct dual computing resource control. We implement Gemini and show that Gemini can improve the analytics accuracy to 90.35%. We further evaluate Gemini by a case study where we use Gemini to support an intrusion detection application, and Gemini shows consistent high analytics accuracy.

I. Introduction

Video analytics systems nowadays support many applications such as video surveillance, vehicle counting, traffic control, self-driving, and others. These systems feed video frames into a pre-trained neural network (NN) model (e.g., vehicle counting) and conduct model inference. In this paper, we study *edge-side real-time video analytics systems*, where videos are generated in edge-side smart cameras and the video analytics are conducted in the edge for real-time response and/or privacy protection. There are orthogonal research directions where video analytics is conducted on pre-stored videos [1], or the real-time videos are sent to the cloud for cloud or edge-cloud analytics [2] [3]. The hardware used for edge-side video analytics systems are smart cameras such as AWS DeepLens [4], with a CPU and a GPU. Typical edge-side video analytics systems include Microsoft Rocket [5], Amazon Rekognition [6], Canon Milestone [7], and others.

A real-time video analytics system needs to achieve high video analytics accuracy while satisfying delay requirements. To face limited edge-side resources, existing systems have an execution pipeline to preprocess video frames to filter out unnecessary frames or to extract only the Region of Interest (ROI) in a frame. When sending the preprocessed frame to model inference, existing systems will select appropriate NN models that best balance the model inference accuracy and delay. In this execution pipeline, the computing workloads of video preprocessing, which involve a large number of searching, sorting, matching operations, are *instruction-intensive computing (IIC)* and are executed in the CPU of the edge camera. The computing workloads of model inference, which involve simple operations but on a large amount of data, are *data-intensive computing (DIC)*, and the DIC workloads are executed in the GPU of the edge camera.

When using real-time edge-side video analytics systems in fields, we observe that the analytics accuracy of the systems can greatly vary. We take Microsoft Rocket (vehicle counting) as an example (details in §II-B). The analytics accuracy at dawn time is 85.7%, and it drops to 65.2% at rush hours. We observe that at dawn time with fewer vehicles, 82% of frames can be filtered and only 18% of frames are fed to model inference. When it comes to rush hours, only 27% of frames can be filtered and 86% of frames are fed to model inference. The video applications have *contents*, e.g., Dawn Time and Rush Hours, and different contents can result in *dynamic* IIC and DIC workloads that most of these content changes are in minute-level. Unfortunately, the CPU/GPU resources are *fixed* in field cameras. This limits the potential to adapt to the workloads and optimize the video analytics accuracy, as we often see that one of the CPU/GPU has reached its maximum capacity, yet the resource utilization of the other is still low.

In this paper, we propose Gemini, a new real-time video analytics system enhanced by a dual-image FPGA. An image in FPGA is a bit file to configure every Logic Unit to the target functions. The newly developed dual-image FPGA, e.g., Intel Max10, Xilinx Artix-7, etc., can pre-store two or more images in the FPGA image flash and switch them with negligible switching time. The dual-image FPGA has a key advantage over current FPGAs on the reconfiguration time, which can take minutes. Moreover, we can alternatively choose the image

2022 IEEE/ACM 7th Symposium on Edge Computing (SEC) | 978-1-6654-8611-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SEC54971.2022.00020

978-1-6654-8611-8/22/$31.00 ©2022 IEEE
DOI 10.1109/SEC54971.2022.00020

Authorized licensed use limited to: University of South-Eastern Norway. Downloaded on February 06,2023 at 10:37:51 UTC from IEEE Xplore. Restrictions apply.