TOSHIBA

DAC 2019

Hybrid ADAS Development Environment for Architecture-specific Performance Estimation

Akira Takeda, Yuji Ishikawa, Tatsuya Mori, Takeshi Kodaka, Yuji Okuda, and Takashi Yoshikawa
Toshiba Electronic Devices & Storage Corporation
2019.06

Motivation: early accuracy / performance estimation in real-time

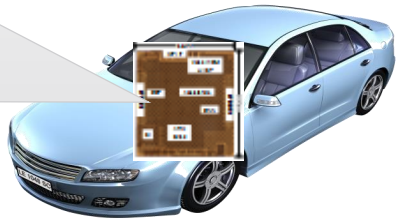
(1) Prototype: running in real time on real vehicle

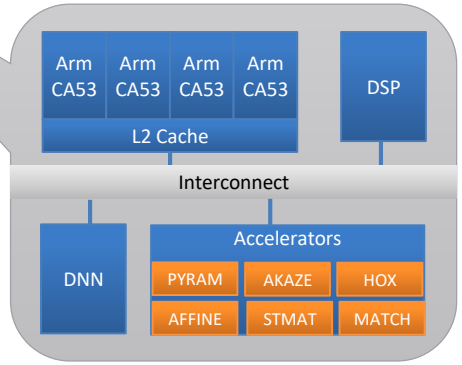
(2) Production: high efficient, high quality

Reference App
(Autoware, Apollo, etc.)
ROS & OpenCV
PC

Porting

Embedded App
Small OS & CVLib
Arm® + HWAs





Unpredictable

Limited precision (e.g. FP16)
Different performance curve

Network size change,
re-training, etc.

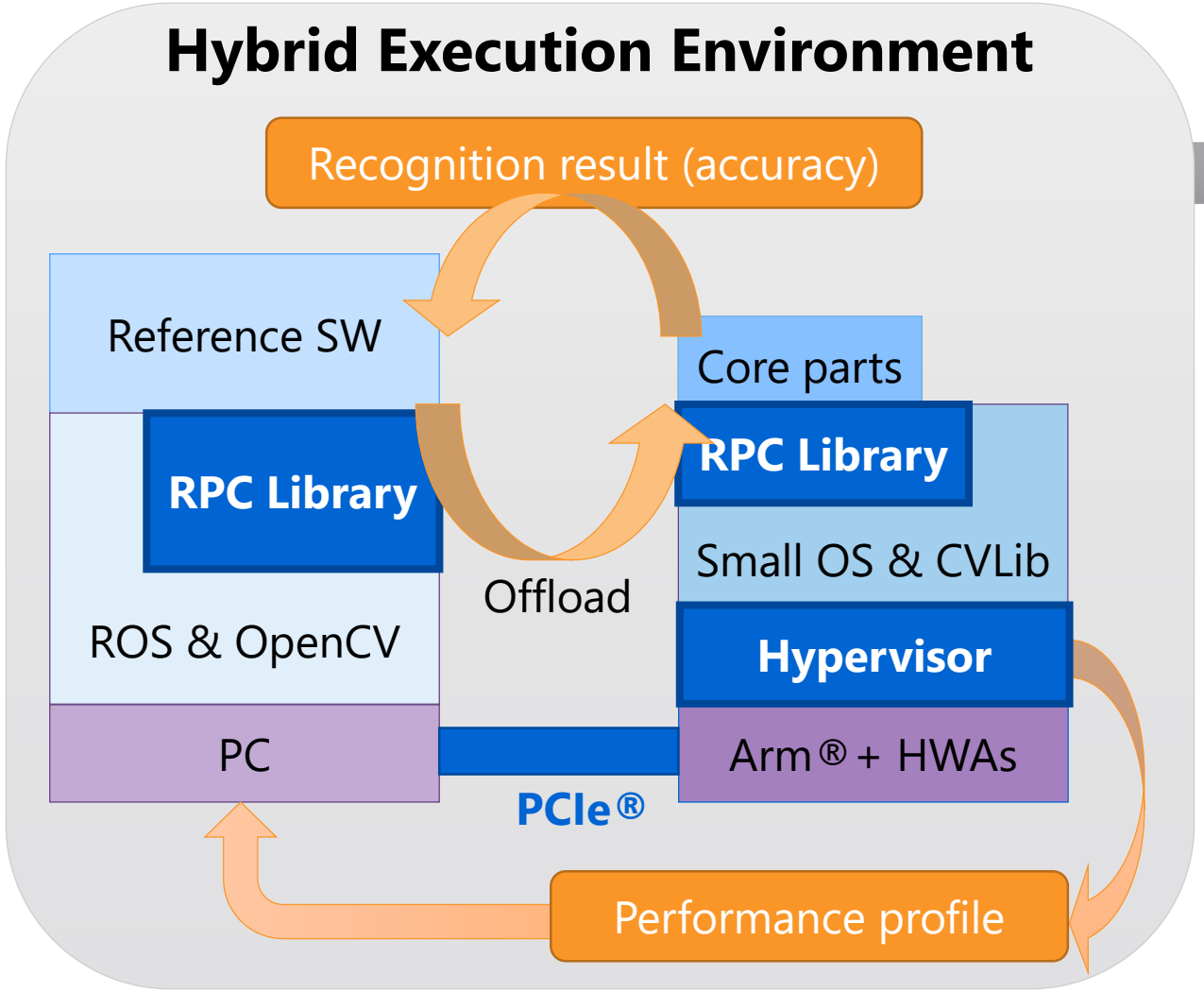
(3) Unexpected
Accuracy / Performance

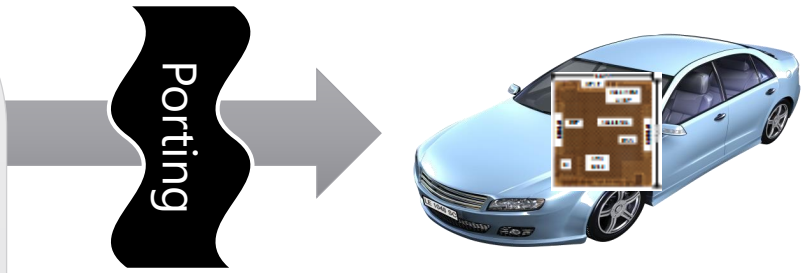
Rich PC-based environment with capability of:

- **early estimation of accuracy and performance** on target SoC
  - **without porting** of overall application
- **real-time execution** (on real vehicle is possible)

Our Solution

"Rich PC" x "Profilable SoC" Hybrid achieves early estimation





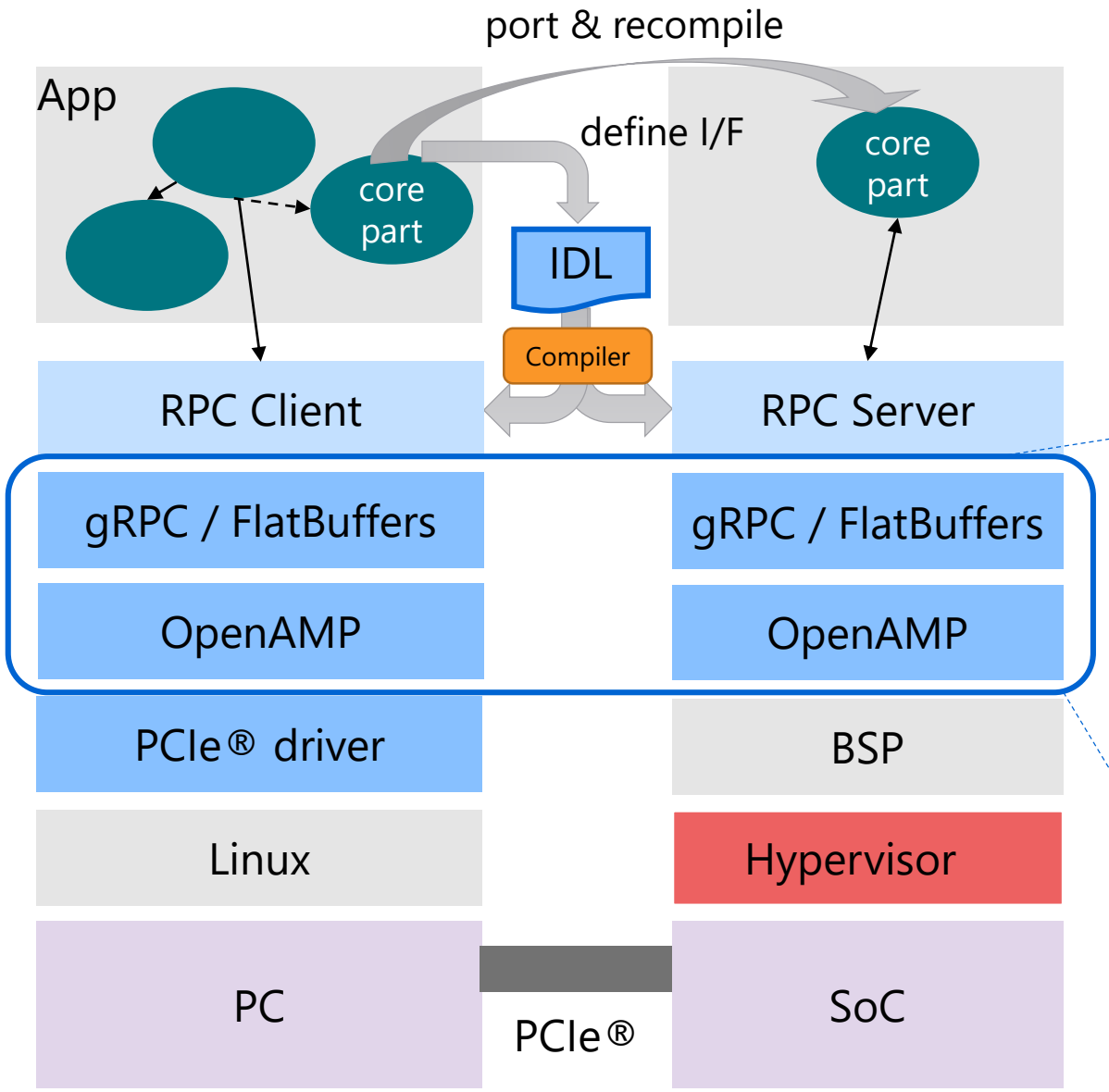
RPC Library

- **Fast** enough for real-time execution
- **Offload only core parts** of application
- Quick definition of core part I/Fs

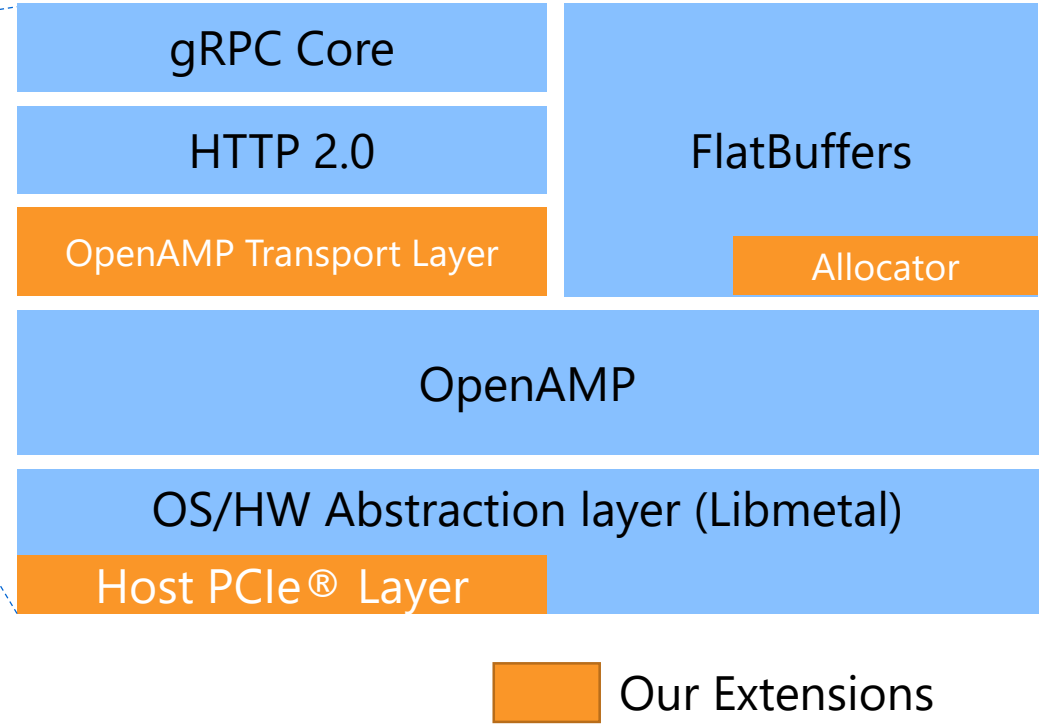
Hypervisor

- Specialized in **profiling** and **monitoring**
- Event-based & sampling-based profiling of SW functions, memory/bus, accelerators
- Error Detection/Trap

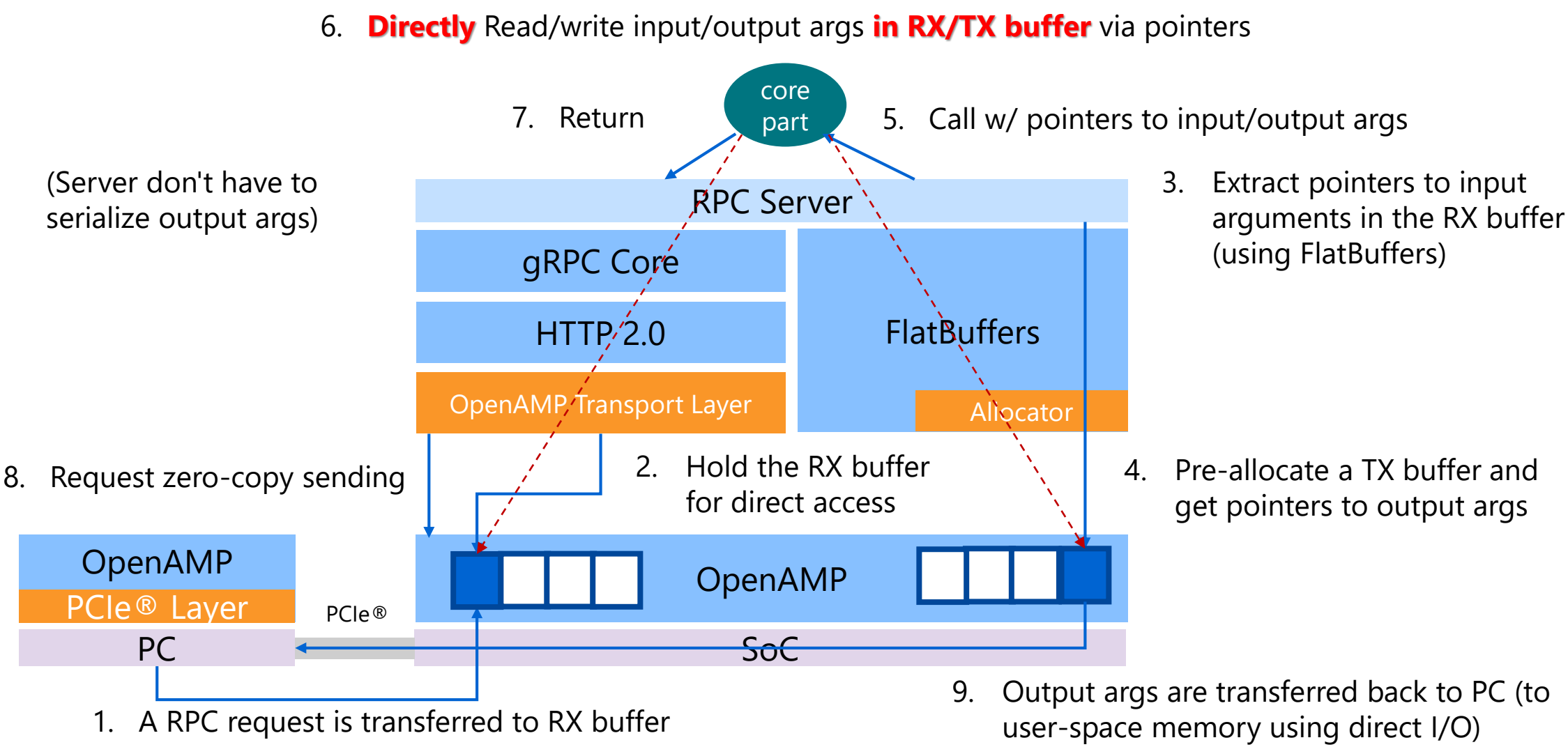
Zero-copy RPC over PCIe®



- Careful Integration of gRPC, FlatBuffers, OpenAMP zero-copy API, and direct I/O PCIe® driver
- New extensions to improve their cooperation

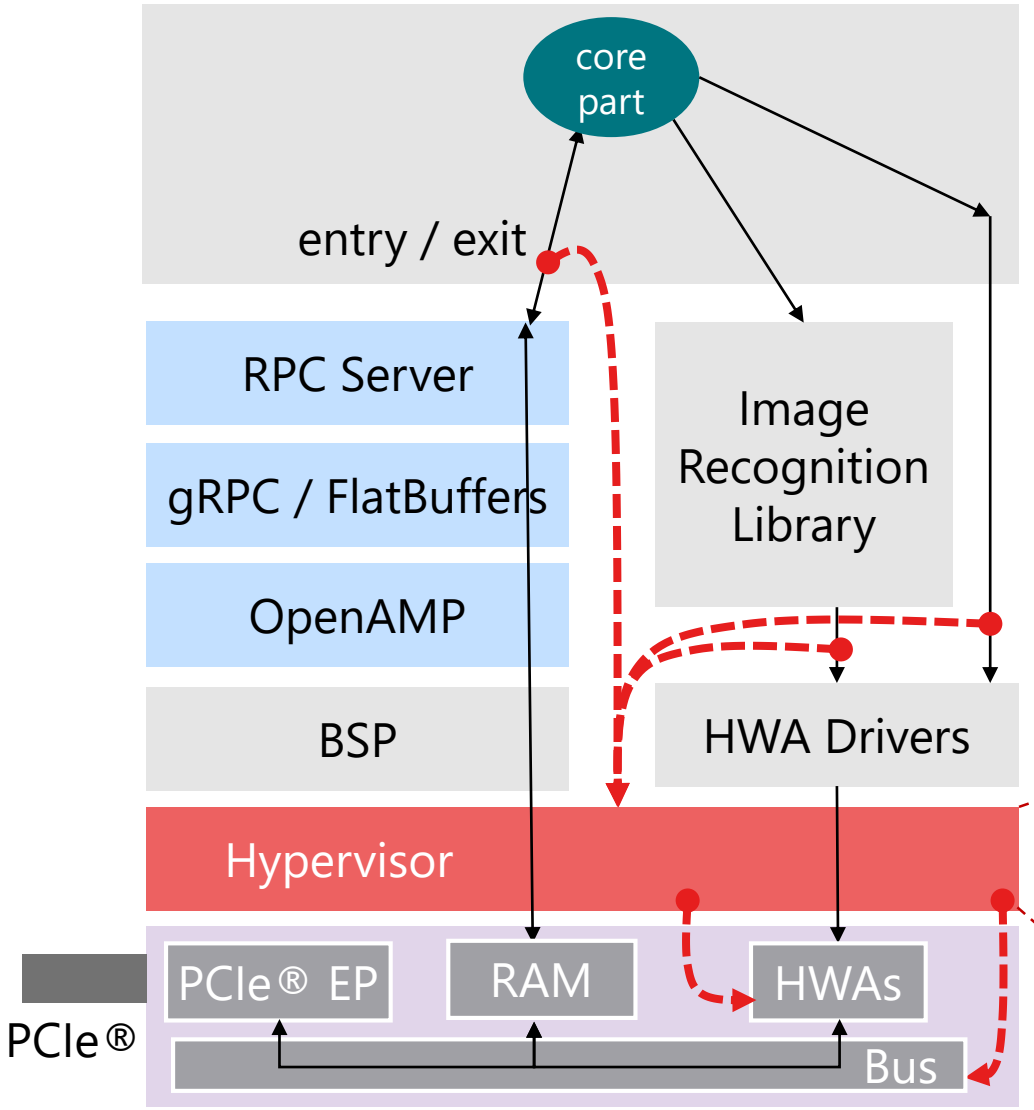


Zero-copy RPC: Direct accesses to RX/TX buffers via pointers

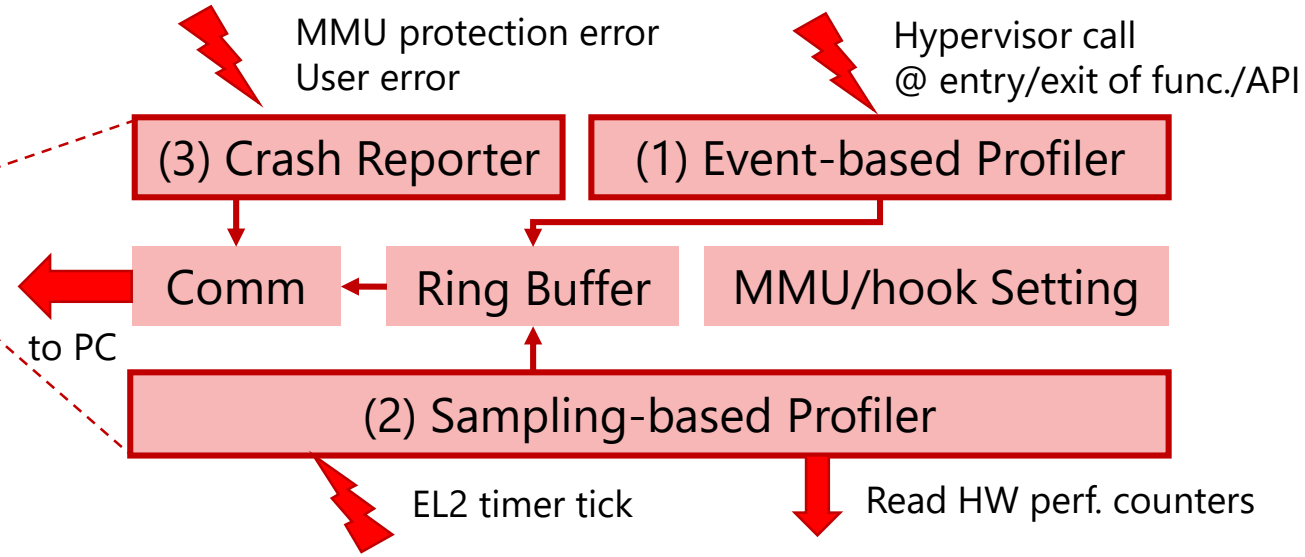


FlatBuffers & OpenAMP collaboration achieves truly zero-copy RPC

Lightweight Hypervisor



- Dedicated to profiling and monitoring
  - No support for multiple VM
- Use Arm® AArch64 virtualization features
  - Second state translation of MMU, interrupt virtualization, etc.
- Hook insertion at run-time
  - Can be applied for pre-compiled binary app



Experiment Setup: 5-class DNN semantic segmentation

HDMI output

PC

PNG @ CityScapes 3x2048x1024

imread / crop

3x1280x900

imshow

SoC

resize @ **AFFINE HWA**

3x368x273

DNN HWA

5x46x35

resize @ **AFFINE HWA**

5x1504x1152

crop / make label image

1x1280x900

convert to color img

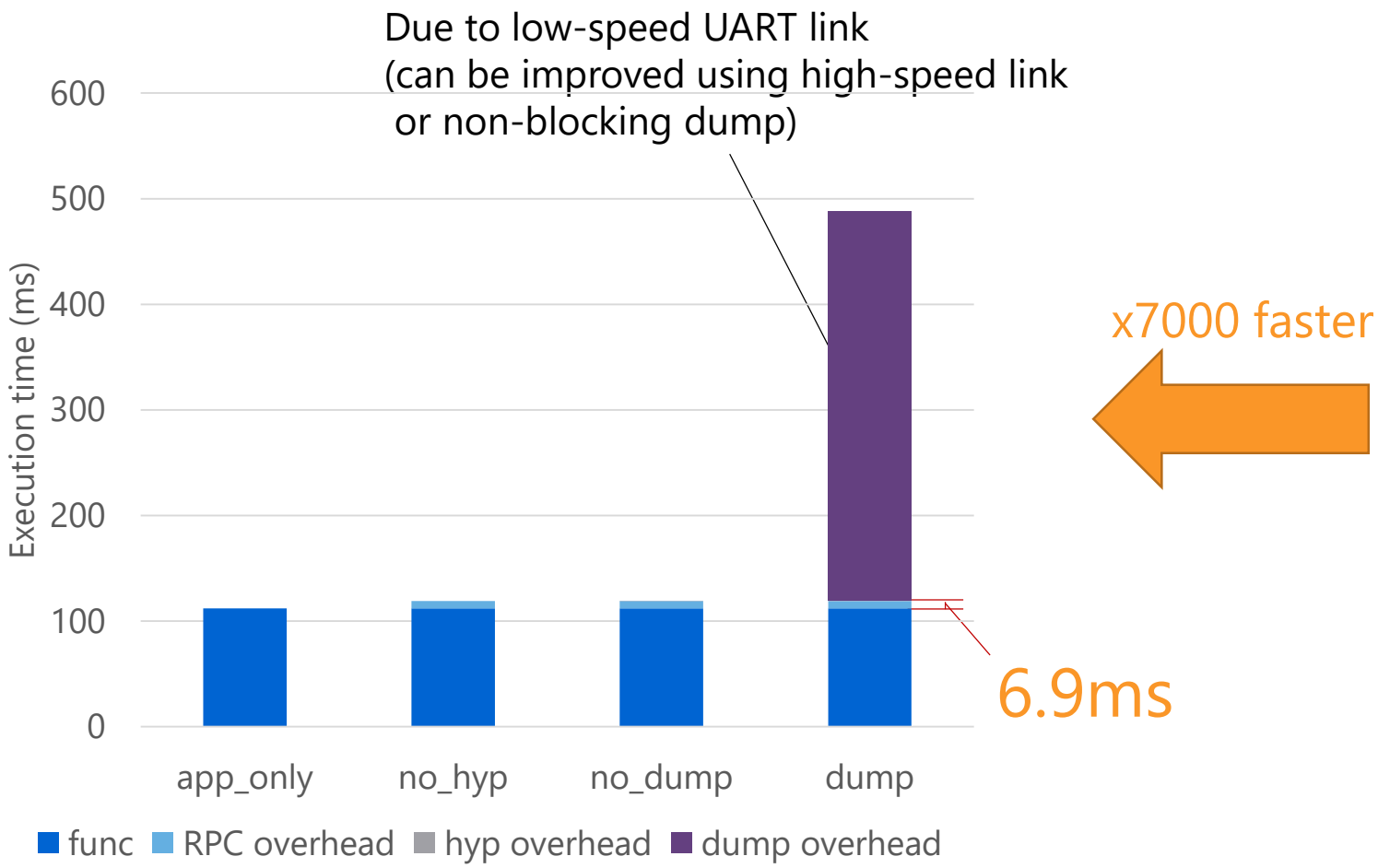
3x1280x900

blending @ **DSP**

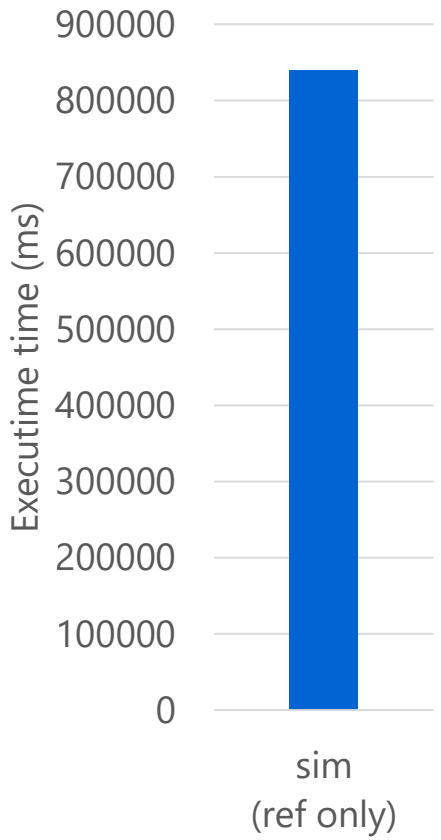
3x1280x900

Experiment Result: Overhead

6.9ms offload latency @ 3x1280x900 & negligible monitoring overhead

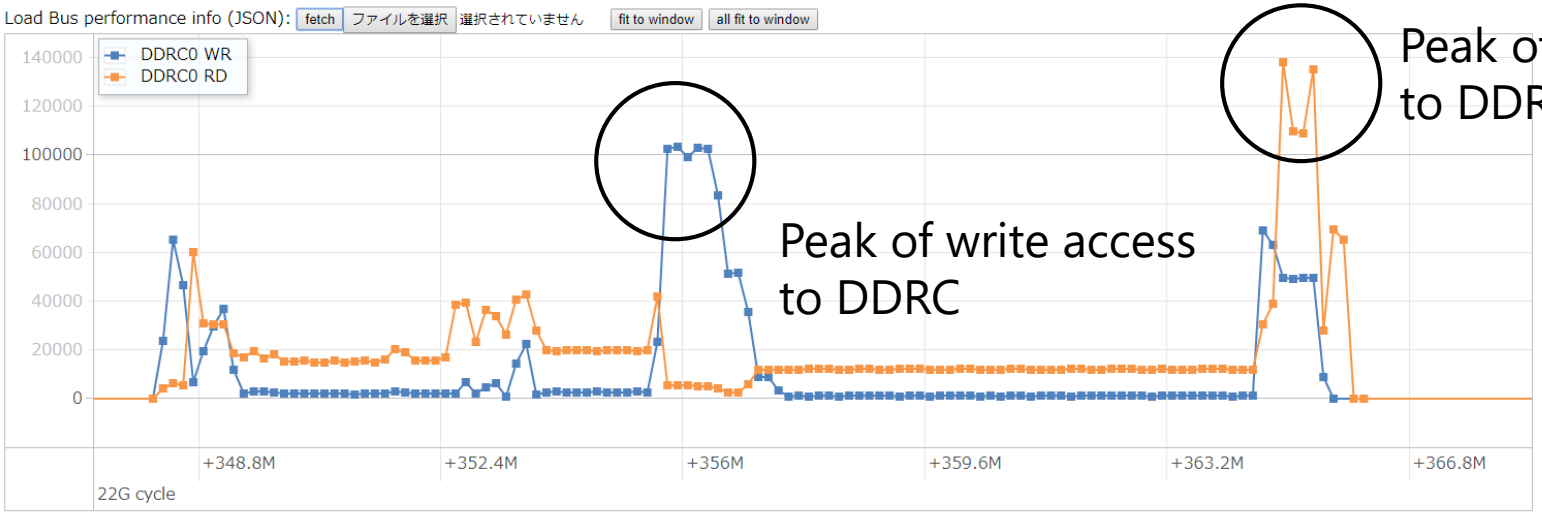


x7000 faster



Experiment Result: Visualized profile

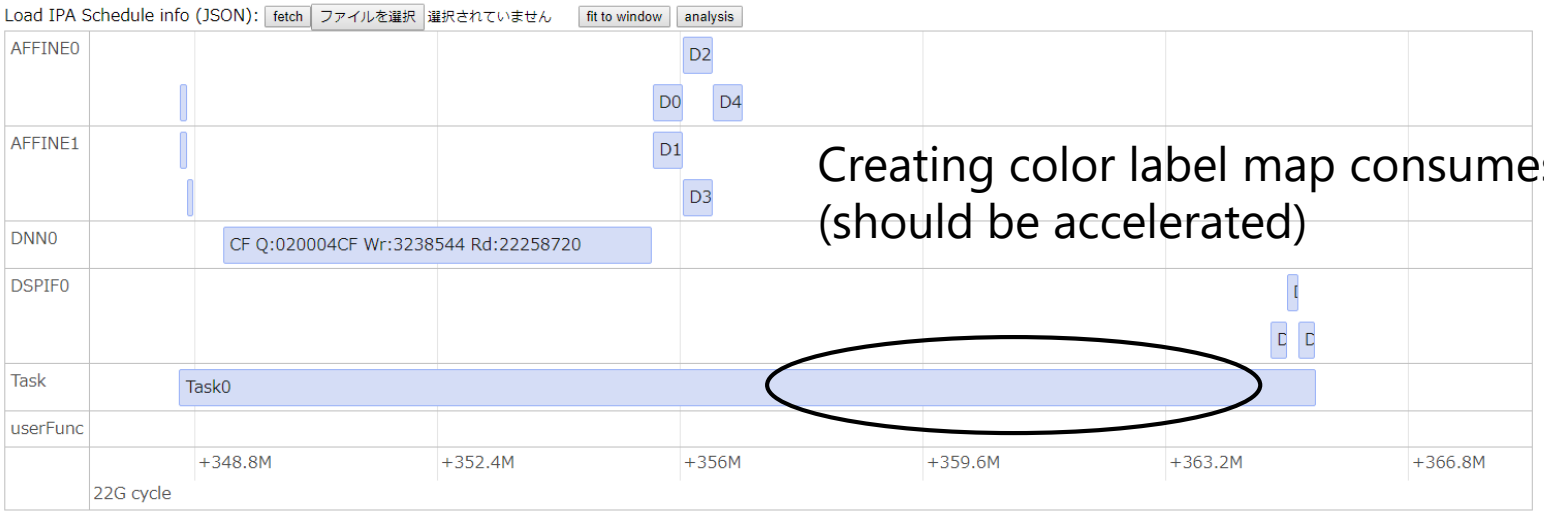
Bus Performance Chart



Peak of read access to DDRC

Peak of write access to DDRC

IPA Schedule Chart



Creating color label map consumes a lot of CPU time (should be accelerated)

Summary

Difficulty of early accuracy and performance estimation for embedded ADAS applications

- Due to limited precision and different performance curve of target SoC

Our solution: PC x SoC hybrid execution environment

- Achieves low-overhead accuracy and performance estimation
  - Offloading hot-spots of application to target SoC
  - Profiling and monitoring on target SoC
- Proposed techniques
  - Zero-copy RPC over PCIe®
  - Lightweight hypervisor specialized in profiling and monitoring

Experiment

- Our solution has enough capability for early estimation using DNN semantic segmentation
  - achieves 6.9ms offload latency @1280x900 image which is small enough to run application in real-time
  - probes user SW, HWAs, and bus behavior on target SoC with tiny overhead

Proposed hybrid environment makes early estimation possible