Adventures in Moving a Computer Vision Solution from Cloud to Edge

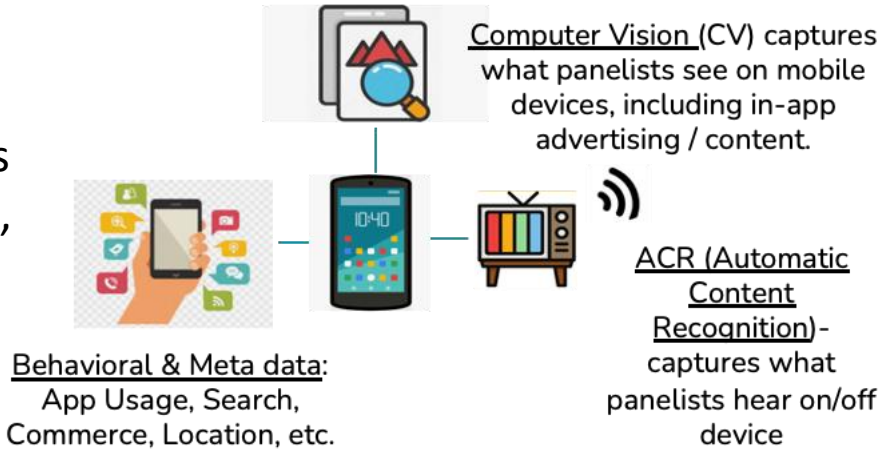
Nate D'Amico

CTO and Head of Product
MetaConsumer Inc.

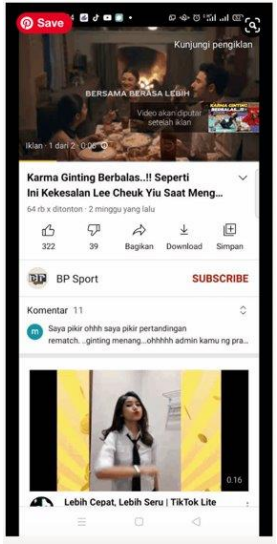
META
CONSUMER

MetaConsumer Overview: Computer Vision Approach to Media Measurement

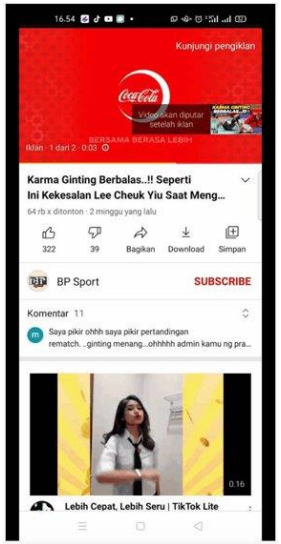
- Single source media measurement using computer vision
  - Panelists are recruited/paid to participate in a study by downloading our app
  - Single source is measuring multiple channels for a single panelist (i.e., Facebook, YouTube, TikTok, web, etc.)
- Visual recognition to “see what the panelist sees”
- Audio recognition to “hear what the consumer hears”
- Sensor and meta data capture



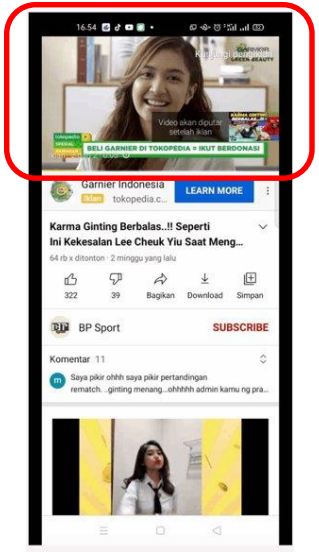
MetaConsumer Overview: Visual Recognition Example



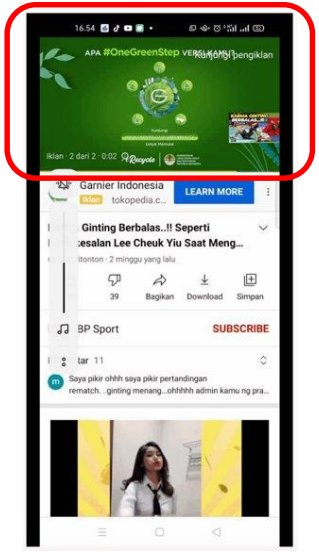
-8 Seconds



-4 Seconds



Time = 0 Seconds



+4 Seconds

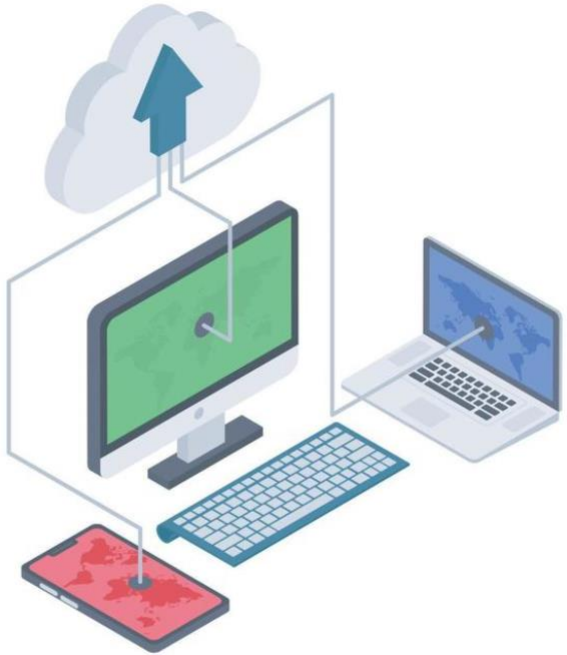
+8 Seconds

MetaConsumer:
Solution & Architecture Overview

- Passive monitor on mobile device collects **a lot** of data with aggressive coverage goals
  - Every second of screen time
  - As many hours as possible of ambient TV/Radio audio capture
- Daily, the passive monitor generates ~500 MB - ~1 GB locally on the mobile device

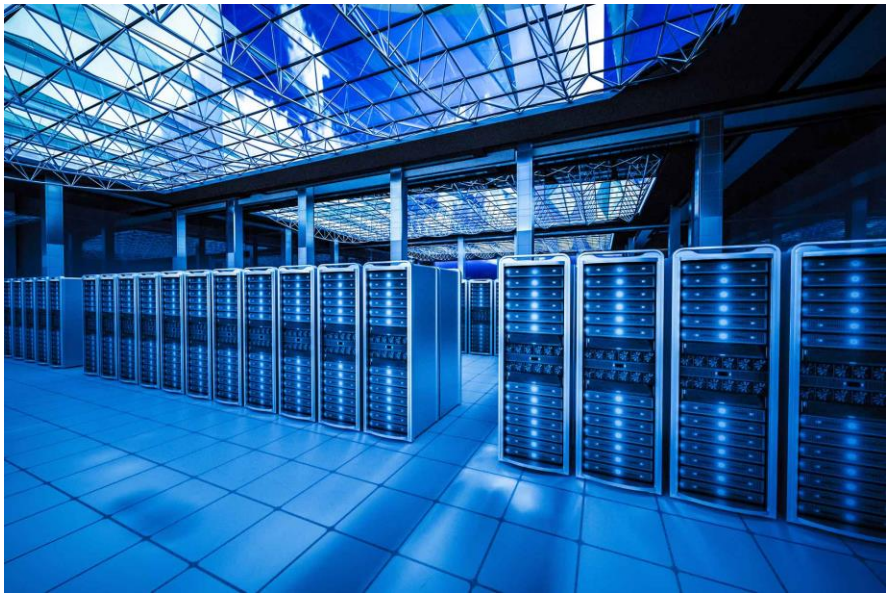
MetaConsumer:
Solution & Architecture Overview

- Encrypt collected data and store locally in files on panelists device
- Under proper mobile conditions (Wi-Fi connectivity and > X% battery life) pack data into zip files and push to S3
- Vision servers running on Ubuntu x86 instances pull jobs from work queue and process uploaded zip file contents (visual and audio files)



MetaConsumer:
Solution & Architecture Overview

- Majority of our workload is computer vision tasks
  - Billions of visual/audio searches
  - **Heavy** CPU cycles, light on memory once indexes are loaded
  - 1000’s of cores per batch cycle
  - 100% AWS Spot instances



Growing Pains:
More Projects, More Data

- Running in multiple regions/countries
- Wi-Fi upload requirements makes it harder to recruit panelists in certain regions in the world, making panel more expensive to run
- Out of control compute costs, reaching ~65% of expenses for a project.
- Data privacy concerns for certain clients in certain regions
- New partner demands, new/more data

Cloud & Solution Technical Challenges

- Privacy concern even with encryption usage, at point of vision processing data is decrypted
- Intermittent Wi-Fi upload failures, more battery drain on mobile devices
- Lots of effort to automate and baby sit Spot requests to maintain $0.01/CPU/hour targets
- Computer vision workload very hard to optimize for parallel compute (GPU, FPGA, etc.)

We Started Looking into the Future

- What scale will we need to operate at?
  - Region by region, which regions, etc.
- Will privacy be a blocker at some point?
- What are the average class of mobile devices in each region?
- What does it look like adding desktop, tv, out-of-Home, VR/AR, etc.?
- Will measurement be solely individual or also whole household?

Journey to the Edge: Weighing Benefits with an Edge Approach

- Privacy concern even with encryption usage, at point of process data is decrypted
- Intermittent Wi-Fi upload failures, more battery drain on mobile
- Lots of effort to automate and baby sit Spot requests to maintain $0.01/CPU/Hour targets
- Computer vision workload very hard to optimize for parallel compute (GPU, FPGA, etc.)

✓ Keep all data on users device and edge device

✓ Local Wi-Fi or Bluetooth transfers immediate

× Trading Spot headaches for edge device logistics, deployment, upgrade management, etc.

✓ No immediate need to optimize, can run 24/7

Journey to the Edge: Starting Limitations

- We are software developers, limited to no hardware experience
- No $$$ to hire outside experts to help and no time to ramp up ourselves
- Look at single board computers and related components
- Design updated human-in-the-loop workflows that can engage with edge content
- Possible wide variants on where a home tv monitor would sit in relation to the tv, makes selecting best/cheapest camera module difficult
  - Interesting but difficult data collection and research project to measure camera abilities
- Introduction of new vision pipelines for tv monitoring at the edge

Journey to the Edge: Starting with Raspberry Pi

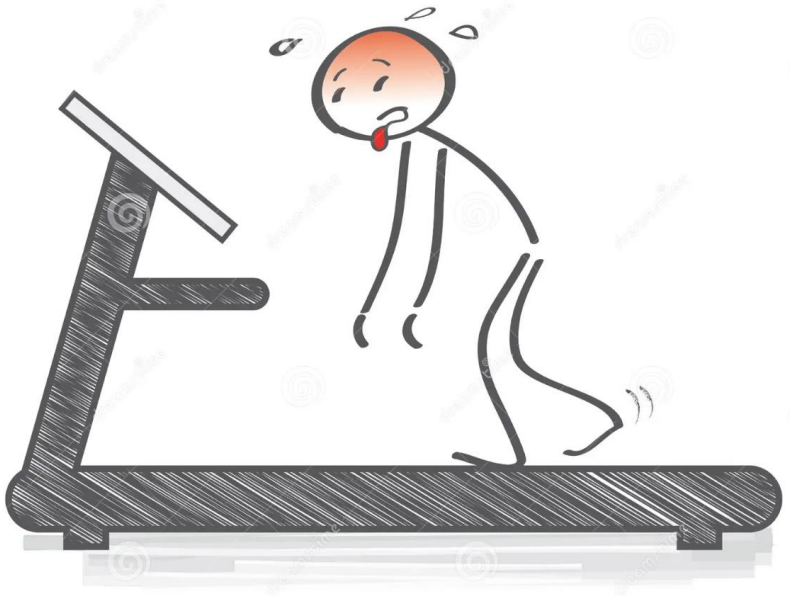
- Limited experience dictated single board computer starting approach
- Raspberry Pi easiest starting choice
  - Ubuntu support and we are already using Debian packaging in the cloud
  - Cheap & easy setup
  - Supply issues tripped us up a bit, unwilling to spend 2x-4x the $$ even if only in the $100s of dollars of waste

Journey to the Edge: Hello World

- Raspberry Pi B+ lying around
- Start with computer vision porting
  - Got vision stack running for reproducing our AWS Spot instance set up
  - Our computer vision stack has **lots** of libraries it depends on
  - Build & dependency work for ARM for libraries that are not pre-packaged (e.g., .deb available)

Journey to the Edge: Leverage the Cloud for Iterations

- Very slow compilation on edge device made the iterative development process untenable
- AWS Graviton to the rescue
  - Previous experience on another project with gen 1 & 2
  - Software iterations in the cloud, push updates to edge
- Open questions/issues
  - Dependency/versioning issues for ARM
  - Floating point calculation deltas



Journey to the Edge:
RISC-V Ready for Main Stage

- Lots of movement in RISC-V space
  - Big community focus on compilation and getting libraries supported
  - Vector extension
  - Custom extensions supported by third parties

Journey to the Edge: Dependency Issues & Legacy Code

- Our computer vision stack has **lots** of libraries it depends on
  - Build & dependency work for ARM for libraries that are not pre-packaged (e.g., .deb available)
  - Much harder for RISC-V efforts
  - Repeat AWS/Graviton Iterations
    - Scaleway introduces RISC-V cloud instances

Journey to the Edge:
Current View & Approach Moving Forward

- Leverage a RISC-V single board computer setup OR design “simple” one leveraging Andes or similar RISC-V processor with vector + custom extension support
- LFEdge/Eve for possible device management
- Looking at 2 edge approaches
  - Home TV Monitor device (edge server + camera/mic sensors)
  - Old/Spare mobile device runs “Privacy Service”, can run 24/7 to offload from panelists main device vs uploading to cloud for processing
- We are also in data collection mode for tv monitoring
  - Aid in camera module selection
  - Have base data sets for new vision pipelines

Conclusions

Leverage cloud services wherever possible for design, iterations, experiments

Use edge work to clear out technical debt and refresh services and dependencies

RISC-V feeling ready for prime time

Thank You – Relevant Links

Graviton (ARM Servers)
https://aws.amazon.com/ec2/graviton/

Scaleway (RISC-V instances)

https://scaleway.com

AWS FPGA Instance Type
https://aws.amazon.com/ec2/instance-types/f1/

Andes (RISC-V: AX45MP)
https://www.andestech.com/en/products-solutions/andescore-processors/riscv-ax45mp/

LFEdge (edge management)

https://lfedge.org/projects/eve/