



The Role of Active Disks in Edge Computing

Ziqiang (Edmond) Feng

Joint work with Shilpa George, Roger Iyengar, Haithem Turki,
Padmanabhan Pillai*, Jan Harkes, Mahadev Satyanarayanan
CMU and *Intel Labs

12/2019 Open Edge Computing Workshop



Visual Data on the Edge

- Bandwidth, privacy, regulations, etc.
 - ☞ *Long-term storage and computation on the edge*
- Both **cost** and **efficiency** are crucial



A Closer Look at Storage

Storage cost efficiency

☞ Store data in **encoded** formats
(JPEG, PNG, H.264, ...)

☞ Store on **disks** rather than SSDs
(4x cost-per-bit diff)

(Un)surprisingly, disk read and image decoding become the performance bottleneck in image analytics.

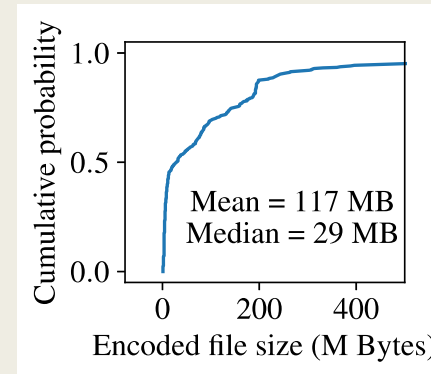
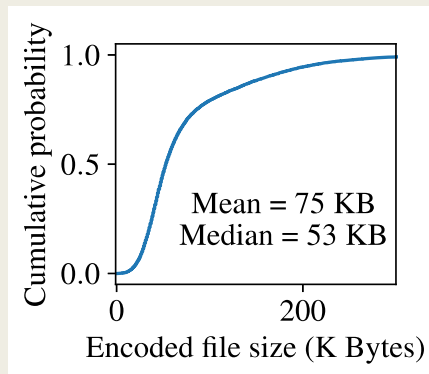
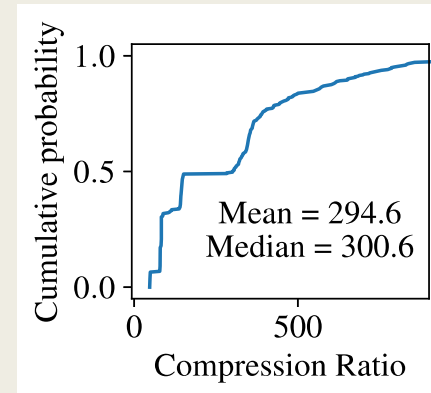
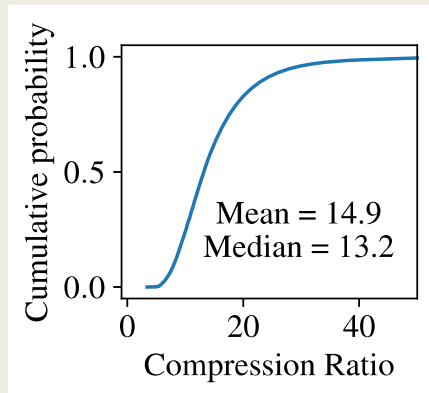
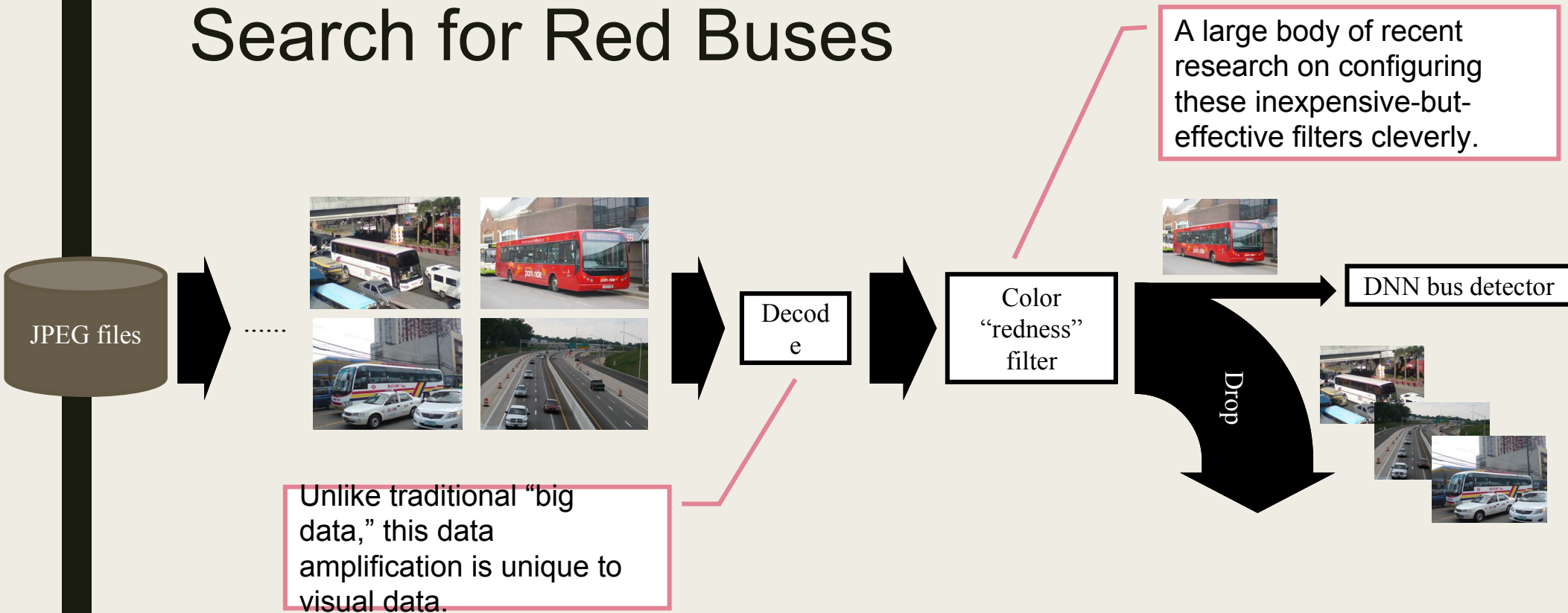


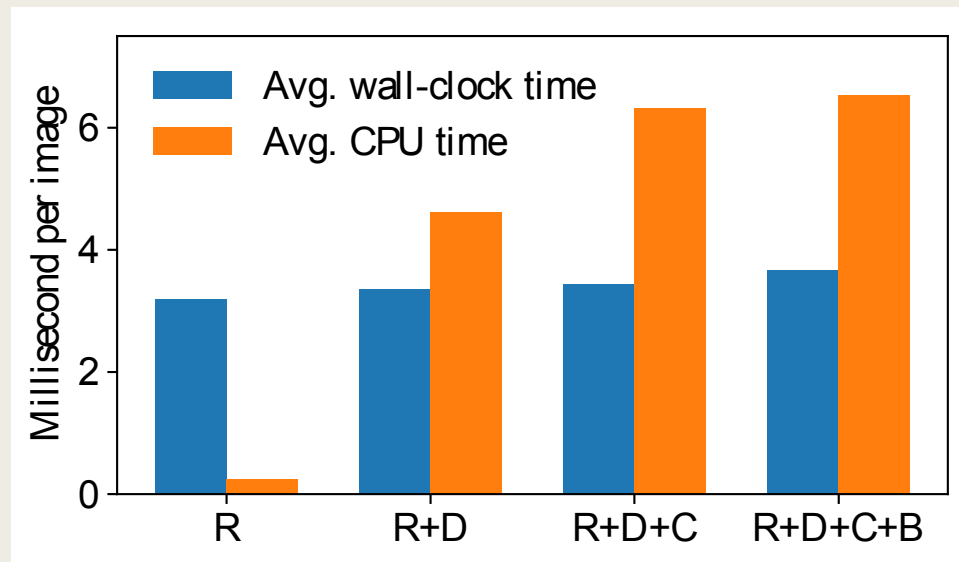
Image Data (JPEG)
YFCC100M

Video Data (H264)
VIRAT

Image Analytics Pipeline Example: Search for Red Buses



Cost of Disk Read and Image Decoding



R: read files from disk

D: decode JPEG into pixel arrays

C: color filter

B: bus detection using DNN

4 cores/8 threads multithreading.

50,000 images.

Only <2.5% of images are processed by bus detector.

Can the Disk Help?

Potential benefits

- Free up host CPU cycles for later stages of the pipeline.
- Maximize utilization of the disk's internal read speed.
- More energy efficient.

“Active disk” = executing application logic on disk

Isn't This An Old Idea?

- Active disks' concept dates back to 1990s
 - *Many research efforts*
 - *Never gained traction in commercial products*
 - *All hail Moore's Law on commodity CPUs*
 - *Evolving workload*
- A new opportunity fueled by new trends
 - *Big visual data + Deep learning + NVMe & hardware accelerators*
 - *Image decoding, as the first step in any pipeline, is necessary, standardized, and thus **future-proof**.*

Two Challenges of Decode-on-Disk

- Decoding using on-drive processor is slow (compute)

Fixed-function hardware decoder

- Decoded data is 15x larger (bandwidth)

NVMe bus

Hardware vs. Software Decode

Decoding Device	JPEG Decode MPixel/s
Host CPU 2.3 GHz (measured, single thread)	60
Disk CPU 1.0 GHz (extrapolated)	26
FPGA (Xilinx)	140
FPGA (Intel)	73

Similar arguments for video decoder and face detector. Refer to our paper for more details.

The Case for NVMe-attached Active Disks

SATA	500 – 700 MByte/s
NVMe	> 1,000 MByte/s
HDD “internal” read	100 – 300 MByte/s
SSD “internal” read	> 500 MByte/s

←
×15 (*compression ratio*)
= 1500~4500 MB/s

The Maker's Advantage: Matching Processing with Bandwidth

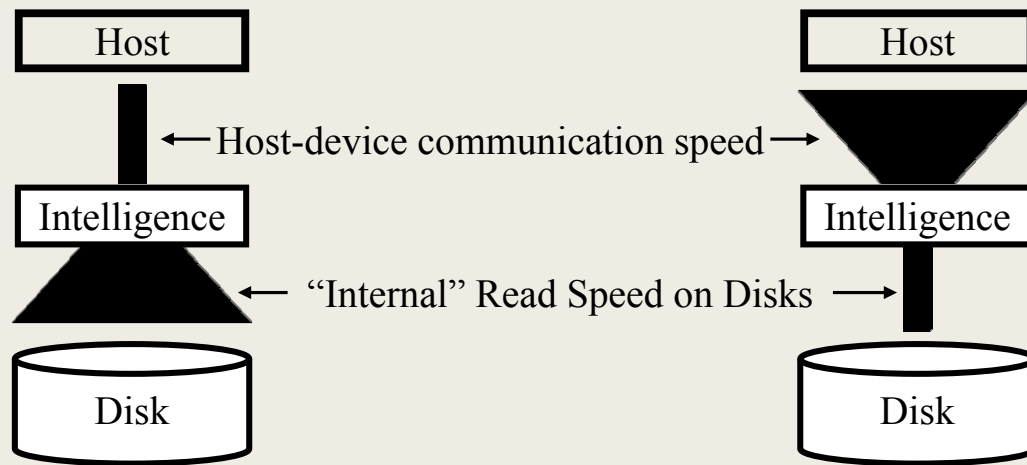
$$\text{Num of decoders} = \frac{\min\left(\frac{\text{Disk internal read speed}}{\text{Avg. *encoded* object size}}, \frac{\text{NVMe speed}}{\text{Avg. *decoded* object size}}\right)}{\text{Object throughput per decoder}}$$

Object Store API with Decode-on-Read

```
FetchAndDecodeObject {  
    int64 object_id,  
    int32 opcode,  
    iovec* where_to_put_decoded_object,  
    iovec* where_to_put_original_object  
}
```

Different opcode to specify on-drive operation and what to return

Useful for sending “passing” images over the network without re-encoding



Previous research on active disks

This work

Optimizing for Batch Processing

- Image analytics is batch processing:
order doesn't matter
- Disk can take the liberty to re-order objects for best performance
 - ***Not today's disks' request scheduling***
 - ***Hide geometry and mechanical from the host***
- Requires a batch iterator interface:
(ensuring exactly-once semantics)

```
IterateCollection {  
    int64 collection_id,  
    int32 opcode,  
    int64 logical_index,  
    int64* flags,  
    int64* returned_object_id,  
    iovec* where_to_put_decoded_object,  
    iovec* where_to_put_original_object  
}
```

Evaluation Methodology

- **Timing-accurate emulation** of active disks
 - *Used an actual disk to get mechanical timings*
- Measured application-level performance with **real** application code running on **real** host CPU and GPU (TensorFlow, OpenCV, ...)
- Data sets
 - *YFCC100M (image)*
 - *VIRAT (video)*


(Refer to our paper for details)

Default Settings

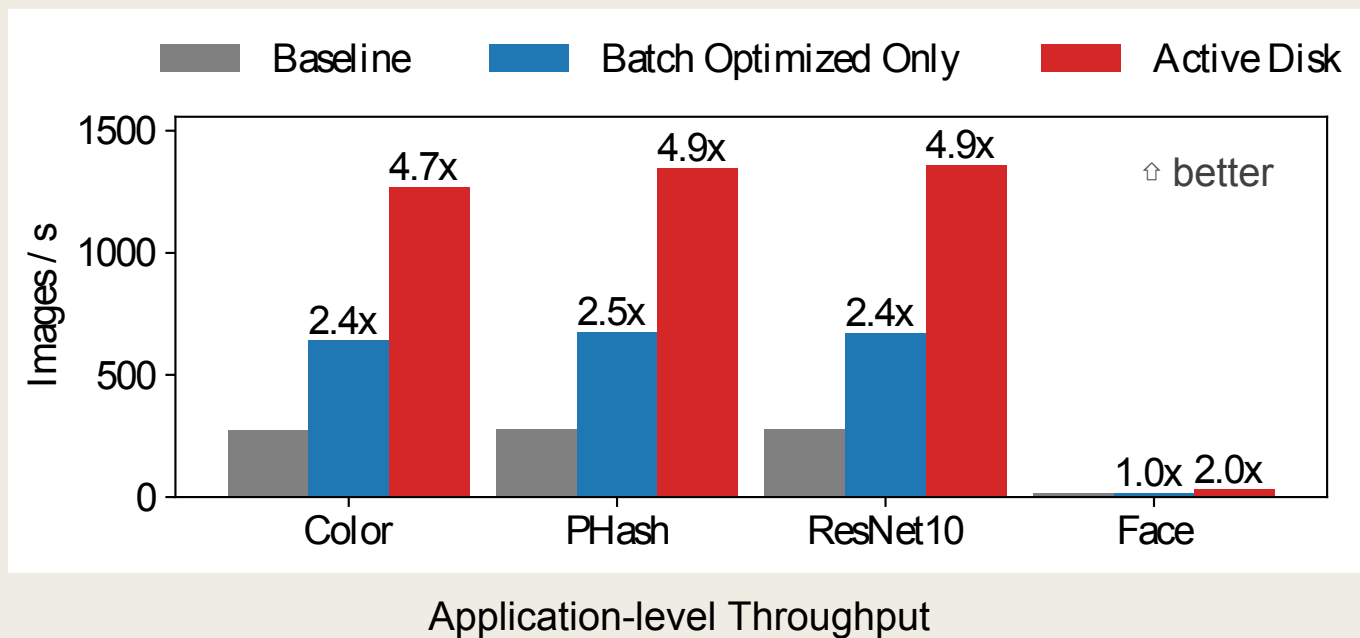
Host	
CPU	4 cores / 8 threads @ 2.30 GHz
DRAM	64 GB
GPU	NVIDIA GTX 1080 Ti

SATA Disk (Baseline)	
RPM	7,200
Throughput (bulk)	187 MB/s
Throughput (JPEG)	98 MB/s

Active Disk (Emulated)	
Host-Disk Bus	2,000 MB/s
JPEG Decoder	5x 140 MPixel/s
Face Detector	30 FPS
Video Decoder	480 FPS @ 720p

 Useful reference
for actual
implementation of an
active disk

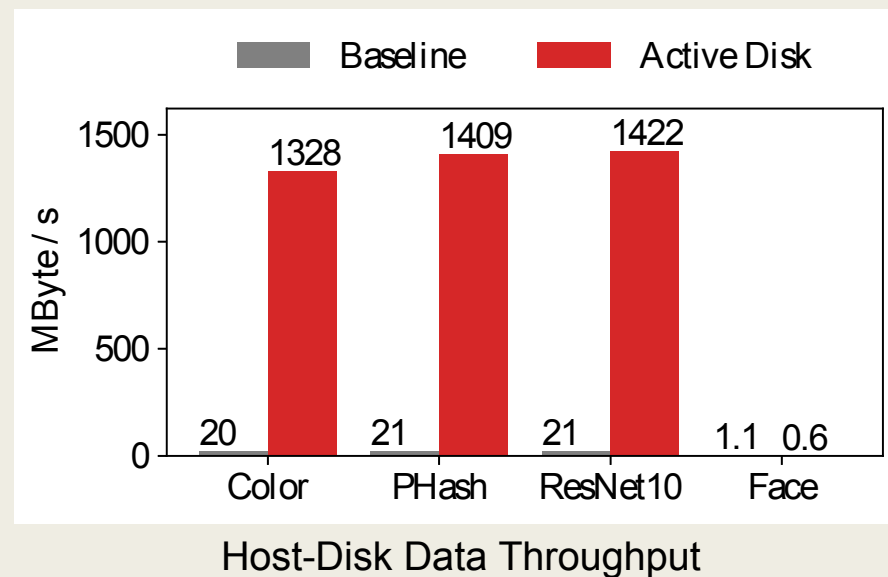
Microbenchmark: Early-discard Filters



Filters: Color filter, Perceptual hashing, Tiny neural network (ResNet10), Face detection.

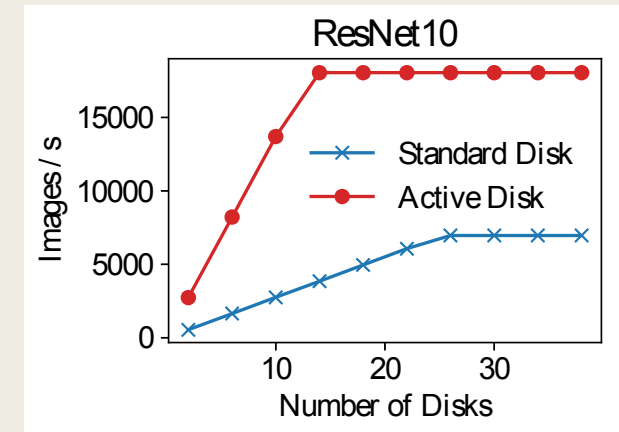
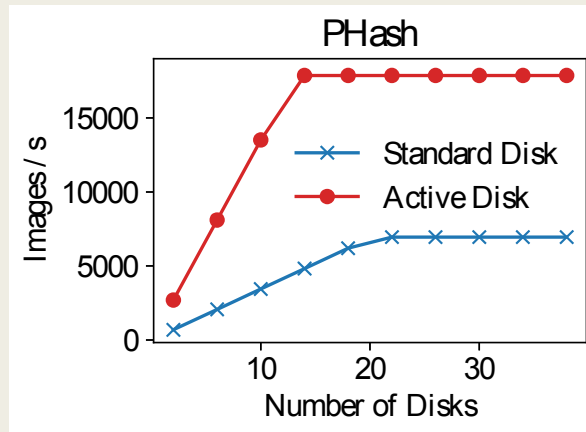
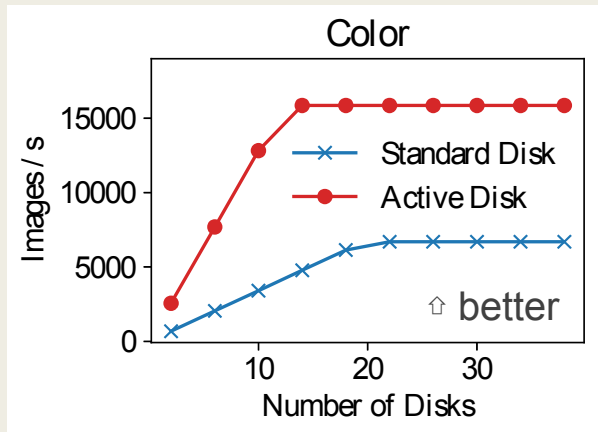
Batch optimization is approximated using FS info.

Is NVMe Necessary and Sufficient?



Yes, with ~1,500 MB/s bandwidth requirement

Let's Add More ~~Servers~~ Disks



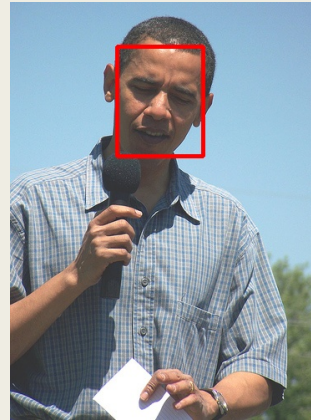
Extrapolated Application Throughput when Adding More Standard/Active Disks to Our Dual-socket 36-core Server

Whole Analytics Pipelines



Red Bus

Color filter + object detection
(Image)



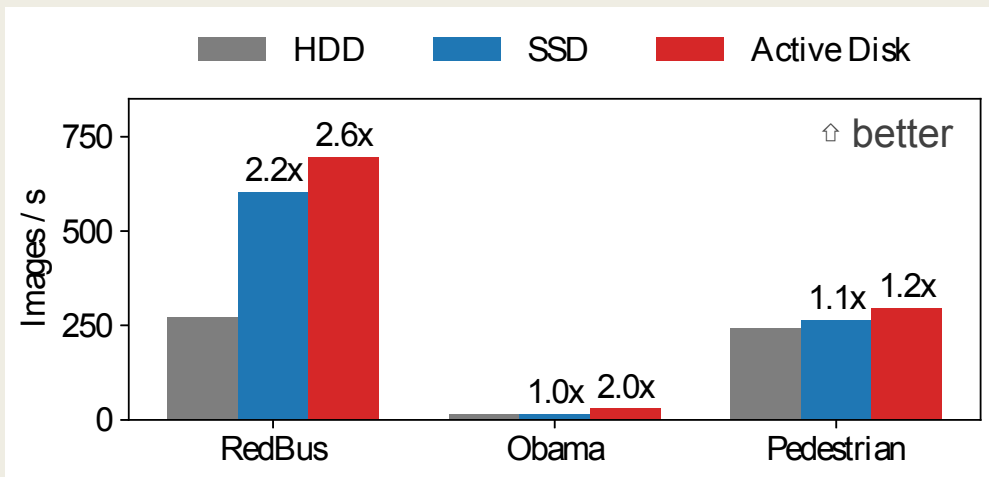
Obama

Face detection
+ identification
(Image)



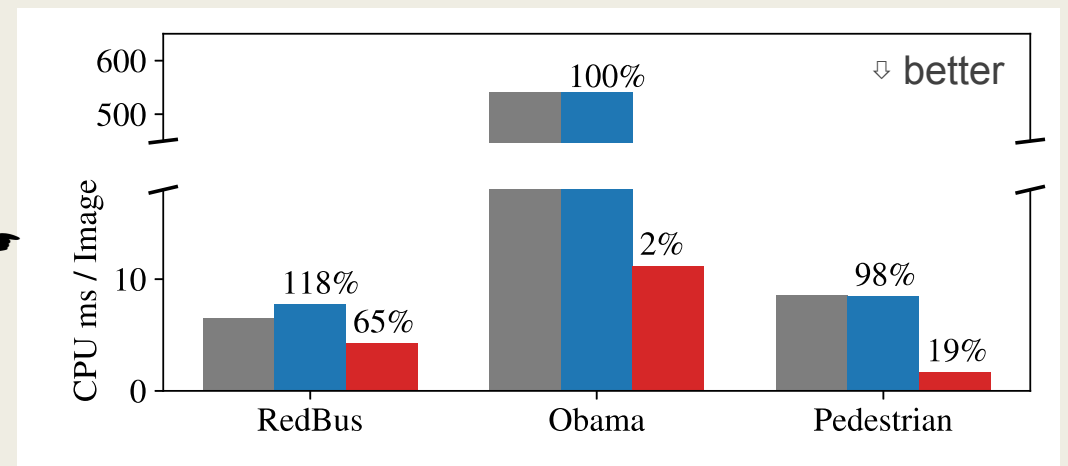
Pedestrian

Frame sampling + Difference detection
+ object detection
(Video)



Application-level Throughput

CPU Cost on Host

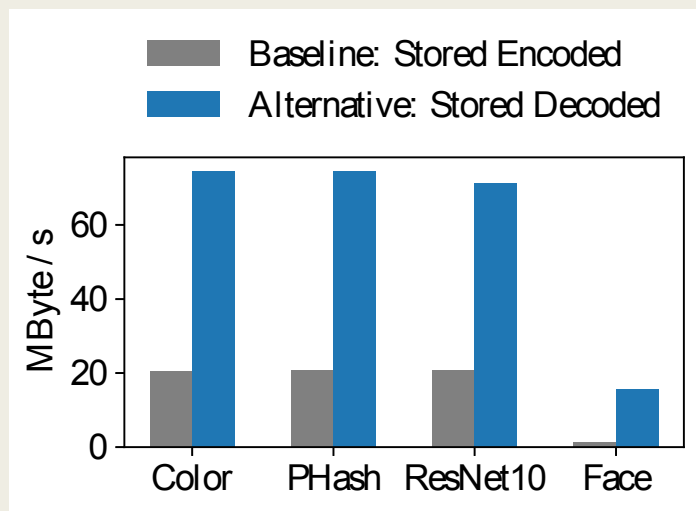


Conclusions

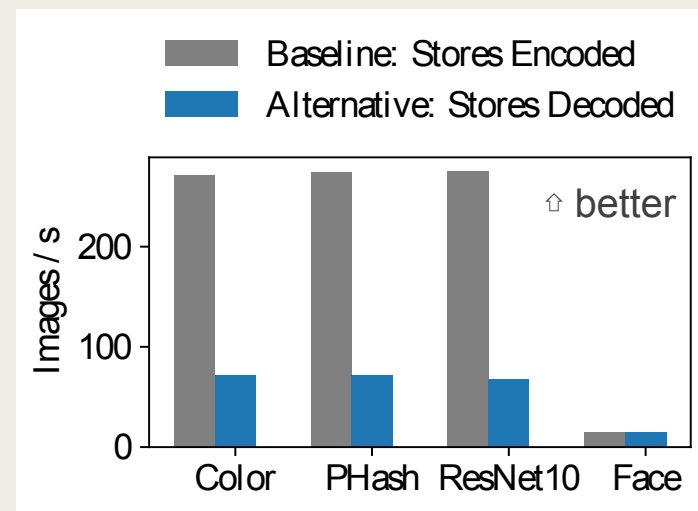
- A new design of active disk
 - *NVMe*
 - *Decode-on-read*
 - *Optimized batch read*
- Speed up image analytics pipelines by up to 2.6x

THANK YOU.

How About Simply Storing the Decoded?



Host-Disk Data Throughput



Application-level Throughput

3.7x higher host-disk throughput ➤ 15x lower object throughput ➤

Comparing to Host-side Accelerator (GPU, Co-processor, etc.)

- GPUs are expensive. You want to reserve their cycles for more valuable general-purpose computations (e.g., DNN)
- Active disks' processing capability:
 - *Scales with data naturally*
 - *Can be engineered to match closely with the disk's internal bandwidth*
- No on-drive batch optimization

Questions to Answer

- Can active disks improve application-level performance?
- Is NVMe necessary and sufficient?
- How much processing is needed on disk to deliver benefits?
- How many active disks can be connected to one server?
- How do active disks compare to alternative solutions?