# REFRESH ENABLED VIDEO ANALYTICS (REVA) : IMPLICATONS ON POWER AND PERFORMANCE OF DRAM SUPPORTED EMBEDDED VISUAL SYSTEMS

Siddharth Advani[1], Nandhini Chandramoorthy[1], Karthik Swaminathan[1], Kevin Irick[2], Yong Cheol Peter Cho[3], Jack Sampson[1] and Vijaykrishnan Narayanan[1]

[1]The Pennsylvania State University, [2]SiliconScapes LLC, [3]ETRI, South Korea

## Introduction

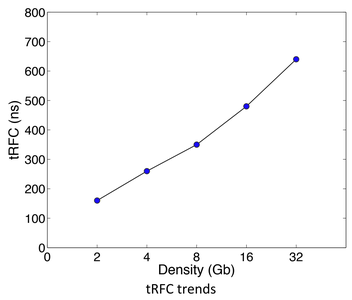
Wearable video systems require capabilities for real-time video analytics and prolonged battery lifetimes for wide adoption.

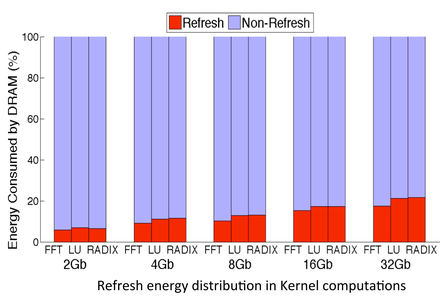
Application-specific video accelerators have shown to be extremely efficient in terms of power and performance to handle real-time workloads.

However, the memory system contributes between 10-30% of the overall power of embedded video systems and mobile phones. In particular, the **refresh power** in DRAMs has become a major source of concern and continues to increase with every new generation of devices.

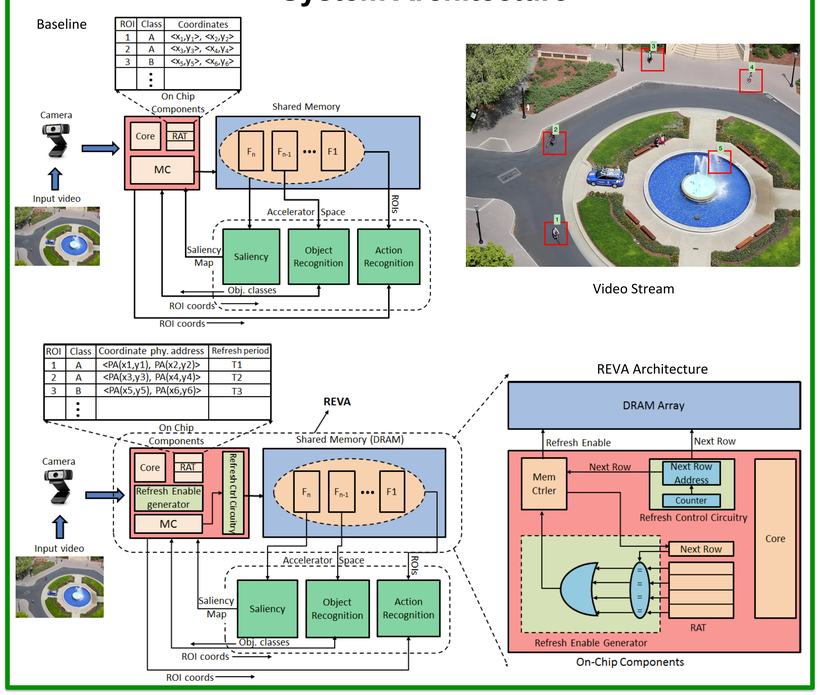
We propose a **Refresh Enabled Video Analytics (REVA)** system that reduces refresh power by using semantic knowledge to modulate the refresh periods at a row-level granularity.

## Motivation




## System Architecture



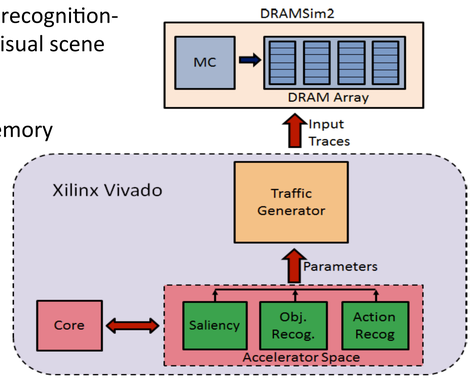

## Experimental Setup

We consider a detection-object recognition-action recognition pipeline for visual scene understanding.

We used Vivado to generate memory traces based on our accelerator parameters.

These traces act as input to DRAMSim2 to enable power modeling.

Our evaluations were on a 8 Gb single rank DDR3-1600 device.

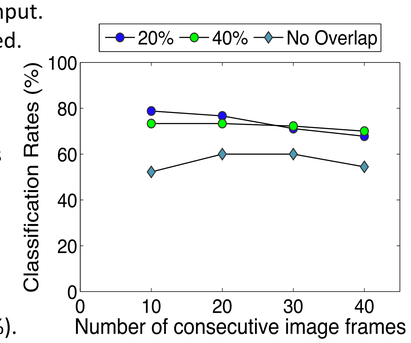


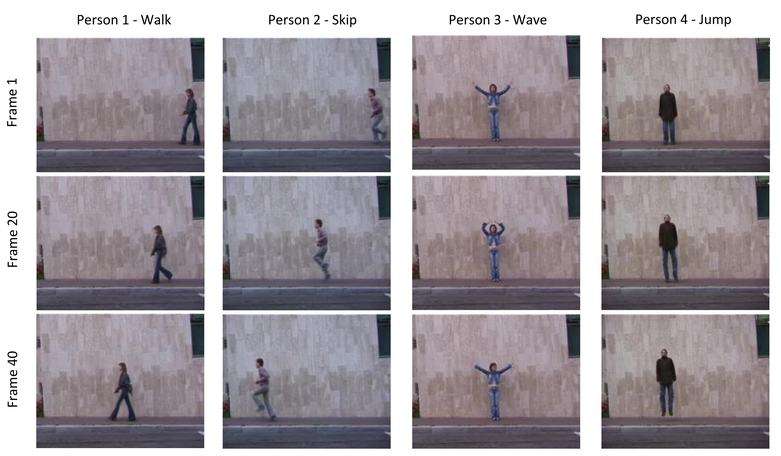

| | |
|---|---|
| DRAM row buffer policy | Open page, Closed after four accesses |
| DRAM Configuration | DDR3-1600, 8 Gb, 1 channel, 1 rank, 8 banks/rank |
| Timing parameters (ns) | $t_{RP}$ = 11, *tRCD* = 11, *tRFC* = 350, *tREFI* = 7800 |
| Current parameters (mA) | $I_{DD0}$ = 110, $I_{DD1}$ = 135, $I_{DD2P}$ = $I_{DD2Q}$ = 40, $I_{DD2N}$ = 42, $I_{DD3N}$ = 45, $I_{DD4W}$ = 280, $I_{DD4N}$ = 270, $I_{DD5}$ = 215, $I_{DD6}$ = 12 |

## Sensitivity Analysis

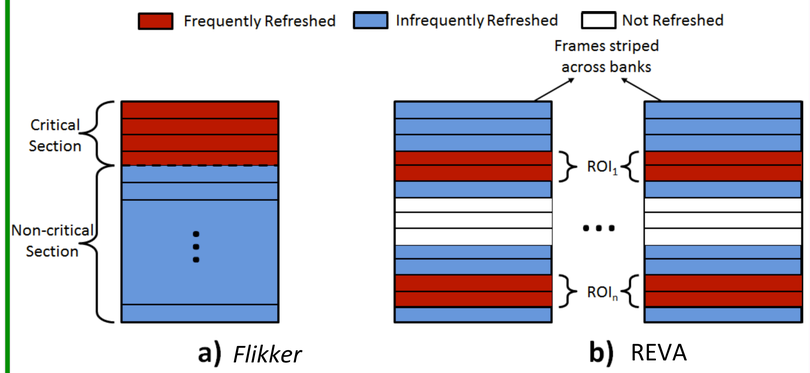
In a purely streaming embedded system refresh can be turned off completely. However, there can be instances when a burst of RoIs is generated and the pipeline cannot sustain a high throughput. RoIs will then be needed to be buffered. In action recognition consecutive RoIs are used for classification.

We evaluated different configurations of action recognition on the Weizmann dataset. The results indicate that a purely streaming (no overlap of video segments) configuration affects accuracy considerably (by around 10%).




## Results

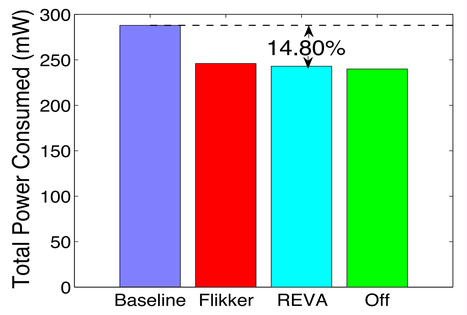



We compare REVA with state-of-the-art architectures like *Flikker*. Other refresh related work includes RAIDR and Elastic Refresh but is orthogonal to our approach.

REVA is dynamically capable of changing the refresh value of an RoI, by simply modifying the refresh period in the corresponding RAT entry.

This avoids any additional writes needed to re-write the RoI into frequently refreshed memory.

The expected gains from REVA's finer grained approach always provide at least as much power savings as *Flikker* while being feasible for a larger set of tasks.




## Conclusions

We demonstrate a **Refresh Enabled Video Analytics** system, which is tuned to optimize energy utilization of the memory by exploiting data characteristics in vision-based applications. We show that depending on the RoI stored in memory, the need for refresh varies from row to row across each memory bank. By adjusting refresh rates selectively depending on the need for re-use of different RoIs, we demonstrate **88%** improvement in refresh power and **15%** improvement in overall power over existing DRAM schemes.

## References

[1] S. Liu, Pattabiraman K., T. Moscibroda, and B. Zorn, "Flikker:Saving DRAM Refresh-power through Critical Data Partitioning," in ASLPOS, 2011.
[2] J. Liu, B. Jaiyen, R. Veras, and O. Mutlu, "RAIDR: Retention-Aware Intelligent DRAM Refresh," in ISCA, 2012.
[3] J. Stuecheli, D. Kaseridis, H. Hunter, and L. John, "Elastic Refresh: Techniques to Mitigate Refresh Penalties in High Density Memory," in MICRO, 2010.
[4] R. Peters and L. Itti, "Applying computational tools to predict gaze direction in interactive visual environments," ACM Trans. on Applied Perception, 2007.
[5] J. Mutch and D. G. Lowe, "Object class recognition and localization using sparse features with limited receptive fields," IJCV, 2008.
[6] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in ICCV, 2007.

This work was supported in part by NSF Expeditions in Computing Award 1317560, and NSF award 1213052. The work was also supported by infrastructure provided by NSF Award 1205618. This work was partly supported by the IT R&D program of MOTIE/ KEIT 1004884 Automotive ECU SoC and Embedded SW for Multi-domain Integration.