A Case for Temperature-Driven Task Migration to Balance Energy Efficiency and Image Quality of Vision Processing Workloads

Venkatesh Kodukula, Sai Bharadwaj Medapuram, Britton Jones, Robert LiKamWa
Arizona State University, Tempe, AZ
vkoduku1@asu.edu,bmedapur@asu.edu,bsjones7@asu.edu,likamwa@asu.edu

ABSTRACT

Many researchers in academia and industry [4, 8] advocate shifting processing near the image sensor through near-sensor accelerators to reduce data movement across energy-expensive interfaces. However, near-sensor processing also heats the sensor, increasing thermal noise and hot pixels, which degrades image quality. To understand these implications, we perform an energy and thermal characterization in the context of an augmented reality case study around visual marker detection. Our characterization results show that for a near-sensor accelerator consuming 1 W of power, dynamic range drops by 16 dB, image noise increases by 3 times, and the number of hot pixels multiplies by 16, degrading image quality. Such degradation impairs the task accuracy of interactive perceptual applications that require high accuracy. The marker-detection fails for 12% of frames when degraded by 1 minute of 1 W near-sensor power consumption.

To this end, we propose temperature-driven task migration, a system-level technique that partitions processing between the thermally-coupled near-sensor accelerator and the thermally-isolated CPU host. Leveraging the sensor's current temperature and application-driven image fidelity requirements, this technique mitigates task accuracy issues while providing gains in energy-efficiency. We discuss challenges pertaining to effective, seamless migration decisions at runtime, and propose potential solutions.

CCS CONCEPTS

• **Computer systems organization → Heterogeneous (hybrid) systems**;

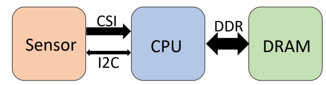
KEYWORDS

continuous sensing, mobile systems, task migration

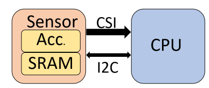
ACM Reference Format:
Venkatesh Kodukula, Sai Bharadwaj Medapuram, Britton Jones, Robert LiKamWa. 2018. A Case for Temperature-Driven Task Migration to Balance Energy Efficiency and Image Quality of Vision Processing Workloads. In *Proceedings of 19th International Workshop on Mobile Computing Systems & Applications (HotMobile '18).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3177102.3177111

HotMobile '18, February 12–13, 2018, Tempe , AZ, USA

ACM ISBN 978-1-4503-5630-5/18/02…$15.00
https://doi.org/10.1145/3177102.3177111



(a) Traditional vision pipeline has energy-expensive camera and memory interfaces.



(b) Near-sensor processing greatly reduces data traffic, relieving energy-expensive interfaces.

Figure 1: Traditional and near-sensor vision pipelines

1 INTRODUCTION

With rapid advances in computer vision, vision-based workloads are increasingly compute- and memory-intensive. Processing such workloads on traditional CPU/GPU based systems (Fig. 1a) is energy-inefficient and slow, due to limited spatial parallelism and energy-expensive DRAM transactions. This has motivated a significant trend towards shifting processing nearer to the image sensor to reduce the overhead of transferring imaging data from the sensor to the application processor. Some works propose streaming DRAM-less vision accelerators [8], while others propose mixed-signal processing [2, 15] to reduce the data size of the imaging output. Advances in 3D stacked fabrication have opened further possibilities for near-sensor processing with minimal footprint, integrating pixel arrays, memory, and processing circuits in different layers. The smaller interconnects and fine-grained parallel processing of stacked architectures raises performance and energy-efficiency, while maintaining small physical area. Due to these factors, commercial devices have employed stacked technology to integrate image signal processing, e.g., demosaicing and white balance, inside compact sensor modules. Whether stacked or non-stacked, the envisioned near-sensor processing architecture would resemble Fig. 1b.

However, near-sensor processing raises sensor temperature, especially for stacked sensor solutions. This is particularly problematic for image sensors, whose quality degrades at raised temperatures. This limitation has blocked the potential expressiveness and performance of near-sensor accelerator implementations from coming to fruition. In § 3, we characterize that after 1 minute of 1 W of near-sensor power draw, sensor temperature rises by 24 °C, which increases the standard deviation of thermal noise by 3X, drops dynamic range by 16 dB, and multiples the number of hot pixels by 16. In addition to degrading user experience, low quality negatively affects the task accuracy of vision applications. In the above conditions, our pose estimation study fails to detect markers in 12% of frames when images are subject to the thermal noise of near-sensor processing. Under normal noise conditions, i.e., when processing far from the sensor, the same pose estimation detects markers on all frames.

To this end, we propose a balance of near-sensor processing and far-sensor processing through temperature-driven task migration. The migration will shift processing nearer to the sensor to promote energy-efficiency by reducing data traffic. As the sensor heats up, the migration shifts processing far from the sensor to promote image fidelity by allowing the sensor to cool down. This strategy will maintain image quality and task accuracy while providing energy-efficiency gains of near-sensor processing. In § 4, we discuss system-level challenges towards temperature-driven task migration.

2 BACKGROUND AND RELATED WORK

Thermal noise in image sensors Image sensors are susceptible to thermal issues, as an increase in temperature reduces signal fidelity of the sensor output due to thermal noise [10]. Thermal noise is fundamentally present across any carrier of charge. For sensors, the mean-square thermal noise voltage is characterized as kT/C, i.e., proportionally related to temperature and inversely proportional to capacitance. Image sensors are particularly susceptible to thermal noise in low light environments, where the image signal is weak, due to fewer photon arrivals.

Furthermore, the dark current of a photodetector creates another temperature-dependent source of noise, which doubles with every 7 K rise in temperature [9]. Dark current manifests as "hot pixels" in the image frame, abnormally bright pixels in otherwise dark scenes. As disclosed in datasheets, many sensors are designed with 60 °C - 70 °C as a hard limit for guaranteed fidelity, but image quality degrades on approach to the limit.

Task migration for the mobile cloud Due to limited energy budgets on mobile systems, many works explore offloading the computation to the cloud. These works are fundamentally different in terms of application partitioning strategy, offloading decision, framework mechanisms, i.e., virtual machine cloning and code offloading. In Mirror Server [20], the entire application is offloaded to a virtual machine. In CloneCloud [6], partitioning is done at a thread-level granularity, based on static program analysis. The offloading decision is taken dynamically based on application status. In MAUI [7], partitioning is done at method-level granularity and the MAUI profiler and solver dynamically make offloading decisions for remote-able methods. These techniques inform our understanding of potential task migration strategies for heterogeneous distributed systems.

Thermal-driven migration for processor Thermal management techniques for microprocessors aim to quell concerns related to hardware reliability and cooling cost. Dynamic voltage and frequency scaling trades energy savings for performance. To reduce performance loss, dynamic thermal management can use trigger, response, and initiation mechanisms [5], hybridized thermal stress-aware adaptation [18], and stochastic techniques for thermal safety [12].

Many OS and compiler-based thermal-aware task scheduling techniques also serve to manage temperature through software-hardware cooperation of thermal-aware priority queues [14], passive load balancing and active migration techniques [16], and mechanisms around task queues [3].

These techniques inspire investigation towards thermal-aware migration for sensors, augmented by the sensor's sensitivity to thermal coupling from near-sensor processing.

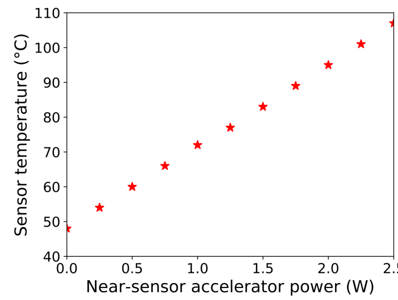


Figure 2: Sensor steady state temperature increases with near-sensor processing power. 1 W raises temp. by 24 °C for an ambient temperature of 27 °C.

3 IMPLICATIONS OF NEAR-SENSOR PROCESSING

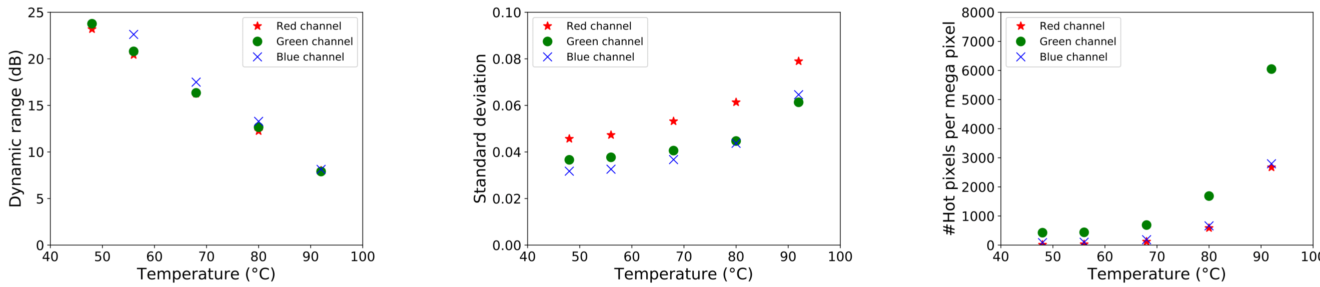
Here we present a characterization to understand the energy and thermal implications of near-sensor processing. In particular, we examine the potential effectiveness of near-sensor processing around a marker-based pose estimation application to understand the relationship between energy-efficiency, image fidelity, and task accuracy.

We have three objectives. First, we investigate the potential energy-efficiency of near-sensor processing. Second, we identify image artifacts generated by the heat of near-sensor processing. Finally, we study how these artifacts affect the task accuracy of an application.

3.1 Proximity for energy-efficiency

Near-sensor processing increases energy-efficiency by reducing expensive data movement across memory and camera interfaces. Here, we discuss these interfaces and their influence on a case study: visual marker-based pose estimation.

3.1.1 Traditional vision pipelines are energy-inefficient due to expensive interfaces. There are three main components in an imaging pipeline: image sensors, processing units, and memory. The physical interfaces between these components consume substantial power. Transmitter and receiver buffers of high speed interfaces use power-hungry operational amplifiers, whose power consumption linearly increases with data rate. Furthermore, to maintain high speeds without increasing voltage, differential interfaces use complementary signals to reduce the influence of noise, drawing more power than their single-ended interface counterparts. Altogether,



(a) Raised noise floors cause the dynamic range to decrease 2 dB with every 3 °C rise in temperature.

(b) Increased influence of thermally sensitive image noises increases the standard deviation of an image.

(c) Raised dark current triggers twice as many hot pixel aberrations with every 6 °C rise in temperature.

Figure 3: Raised temperatures degrade image quality due to thermal noise and hot pixel generation.

this results in energy-expensive high-speed interfaces; common LVDS interfaces consumes approximately 30 pJ/bit.

As shown in Fig. 1a, an image sensor and application processor communicate with each other via camera serial interface (CSI) for frame data and I^2C bus for camera control and configuration. Meanwhile, DRAM and CPU communicate via a DDR interface for read/write/control operations. In vision applications, e.g., marker-based pose estimation, these interfaces consume substantial power dissipation, which we examine using Microsemi's power estimator [1]. For high precision, the application may need to process frames at high resolutions when visual markers are small and/or far from the camera. Streaming at 4K resolution at 60 frames per second will result in data rates on the order of Gbps.

For the camera interface, we evaluate a common CSI interface with 1 clock and 4 data lanes; we find that this consumes 135 mW of power. Since I^2C is used only for control/configuration, it operates at low frequency – 400 kHz – and consumes only 12 mW of power. For the DRAM interface, a common DDR3 configuration will use a width of 32 bits, ECC, and on-die 120 Ω termination. In this configuration, the DDR interface consumes around 600 mW of power. While CSI and DDR interfaces are active throughout the runtime of the application, the I^2C bus stays active just before the start of the application. Consequently, the camera and memory interfaces sum up to 835 mW of substantial power. Thus, interfaces cause traditional vision pipelines to be energy expensive, motivating a pursuit of optimization.

3.1.2 Near-sensor processing reduces the burden of energy-expensive interfaces. To avoid the burden of energy-expensive interfaces, near-sensor processing units should process vision workloads. This will reduce – and in some cases, eliminate – the use of energy-expensive DDR and CSI interfaces.

Near-sensor processing pipelines architecturally differ from traditional vision pipelines in that in-sensor memory replaces off-sensor DRAM. This allows the vision accelerator to operate on image data before the data crosses the off-sensor interface. The accelerator could be fixed-function hardware or a general-purpose processing unit, as long as it satisfies the computational needs of the near-sensor processing. With recent advances towards commercial in-sensor accelerators [13], such an architecture is practically realistic.

The near-sensor processing unit would substantially reduce interface data rate. In the case of marker-based pose estimation application, the output from the accelerator would need only be the translational and rotational estimates of a pose, whose 6 floating point numbers would occupy 24 bytes of data per frame. For optical see-through devices, the camera frame itself is not necessary. For video see-through devices, camera frames can be sent at lower resolutions and lower frame rates than those used for pose estimation. As opposed to the Gbps burden of camera frame data, which requires a CSI interface, the reduced data output would allow the use of the efficient I^2C interface for dramatically reduced energy consumption.

3.2 Proximity degrades task accuracy

While near-sensor processing is energy-efficient, it increases sensor temperature through thermal coupling. Higher temperatures significantly degrade image quality, due to thermal noise and hot pixel generation, impairing the task accuracy of vision applications. Here, we present a thermal characterization of near-sensor processing. Specifically, how does...

- ... near-sensor processing affect sensor temperature?
- ... sensor temperature affect image quality?
- ... image quality affect task accuracy?

3.2.1 Proximity generates heat. Any form of processing dissipates power, which creates heat. Due to their small, silent form factors, mobile devices use passive cooling to dissipate heat through conduction to the skin of the device. In smartphones, the sensor and CPU are separated via ribbon cable, which limits thermal coupling; CPU temperature does not substantially influence sensor temperature. However, the proximity of a near-sensor accelerator will generate strong thermal coupling of the sensor with accelerator power.

To characterize this temperature-power relationship, we use Therminator [19], a compact thermal simulator for smartphone hardware. The simulator uses a specification file to describe the device layout, comprising the size and location of different components, and a power trace file to profile component power. We modify the files to place the processor and sensor in close proximity, separated by 17 mm.

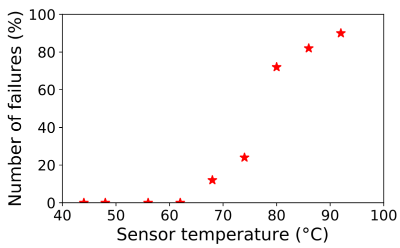


Figure 4: Degradation in image fidelity at higher temp. increases failure rate of marker detection.

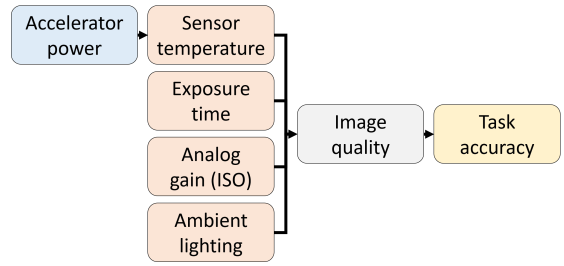


Figure 5: Dynamically changing sensor and environment conditions influence task accuracy. This poses a challenge to migration decisions.

In this placement, the steady state temperature of sensor linearly increases with the processing power of the near-sensor accelerator, as shown in Fig. 2. For every 1 W of near-sensor processing power, within 60 seconds, there is a 24 °C rise in sensor temperature. Stacked image sensor configurations will generate even greater temperature dependencies.

3.2.2 Heat degrades image fidelity. High temperatures degrade image quality, due to thermal noise and hot pixels. While denoising algorithms can remove some of these artifacts, their effectiveness needs to be weighed against their power consumption and performance overhead. We leave an extensive analysis of denoising solutions for near-sensor processing as future work. Here, to quantify noise, we perform experiments [1] to capture raw images under controlled lighting as we heat the sensor to different temperatures.

For our image capture platform, we use the AR0330, a Bayer-filtered 3.2 Mp image sensor, integrated with Microsemi Smart-Fusion2 advanced development kit. The solution uses DRAM for temporary image storage and an FPGA for processing. The FPGA hosts a Cortex-M3, which we use to configure the image sensor's registers, e.g., for 32 ms exposure time and 4X analog gain. We remove the camera lens to reduce temperature-related distortion in captured images.

We construct a platform to observe how reported pixel values change with temperature. To control lighting, we vary a pulse width modulated LED, adjusting its position for uniform sensor illumination. To control temperature, we use a Hitachi RH650V heat gun and a FLIR ONE thermal camera to raise and monitor temperature. We raise sensor temperature to 100 °C and capture frames as the sensor cools down. We cycle the temperature 10 times per lighting environment.

To generate a baseline "noiseless" image for comparison, we average 10 frames captured after the sensor settles to an ambient steady-state temperature of 44 °C. Using pixel-wise comparisons of captured frames against the baseline image, we characterize the effect of temperature on pixel value. To do this, we first group pixel locations by color channel and reported baseline pixel value. Then, for a given temperature, we treat each pixel of each captured frame as a sample of a distribution, grouped by the baseline pixel value of its pixel location. This characterization reports three important image degradations created by high temperatures.

[1]Data available at `https://github.com/kodukulav/research/thermal`

Dynamic range reduction: Dynamic range measures a sensor's ability to capture dark and bright portions of a scene. As noise floor increases with temperature, the dynamic range of reported values shrinks. Our measurements show that high-temperature pixels report raised values, separated from the baseline value by a temperature-dependent offset. As this forces pixel values to start above zero and saturate at lower luminance, the effective range of reported pixel values shrinks. We measure that dynamic range drops by 2 dB for every 3 °C rise in temperature as shown in Fig. 3a.

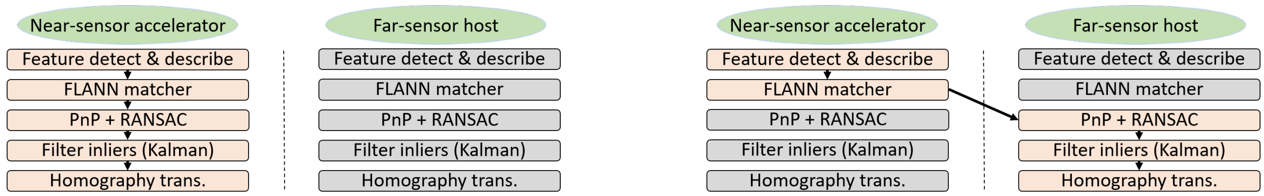
Image noise increase: Thermally sensitive noise sources cause pixels to report deviated values. We quantify this by calculating the standard deviation of pixel values in our grouped distributions. Our measurements confirm that noise sharply increases with temperature, as shown in Fig. 3b.

Hot pixel generation: As discussed in the background, hot pixels appear where dark current is high. These aberrations increase with temperature, as dark current exponentially increases with temperature. We count abnormally large pixel values in dark images to measure the rate of hot pixel generation. Our measurements confirm that the number of hot pixels doubles for every 6 °C rise, as shown in Fig. 3c.

3.2.3 Degradation in image fidelity impairs task accuracy. Raised temperatures lead to pixel-level artifacts that degrade image quality, impairing the task accuracy of vision applications. To study the consequences of image quality on task accuracy, we insert noise into images around marker-based pose estimation, implemented around OpenCV's tutorial[2]. The code uses OpenCV calls to extract visual features and descriptors, to associate image features with reference template features through Flann-based matching, and estimate the pose of the camera through a Perspective-n-Point algorithm. This reports camera pose with respect to a physical marker.

For our characterization, we use our Microsemi imaging setup to capture images of a 7.5 in. x 10 in. marker from various perspectives. We then use noise models interpolated from our characterization, adding noise to captured images to simulate high temperature captures from the same camera pose. As shown in Fig. 4, at a fixed distance of 100 cm pointed at the target, the number of marker detection failures sharply increases with temperature. Thus, to keep estimation failures below a threshold, sensor temperature must remain low.

[2]`http://docs.opencv.org/3.2.0/dc/d2c/tutorial_real_time_pose.html`



(a) All stages run on near-sensor accelerator, resulting in no communication overhead. This thermally strains the sensor.

(b) For thermal relief, we migrate three stages from near-sensor acc. to far-sensor host, using data transfer overhead.

Figure 6: Two different partition splits

Our measurements indicate: (*i*) processing far from the sensor is energy-inefficient, due to power-expensive interfaces, and (*ii*) processing close to the sensor degrades task accuracy due to image degradation. Thus, near-sensor processing and far-sensor processing schemes create a possibility for energy efficiency and task accuracy tradeoffs.

4 THERMAL-DRIVEN TASK MIGRATION

Our characterization reveals that near-sensor processing and far-sensor processing offer tradeoffs between energy-efficiency and image fidelity. This makes a strong case for thermal-driven task migration: shift workloads towards near-sensor processing for energy-efficiency and towards far-sensor processing for image fidelity.

Systems can partition workloads by different granularities: by programs, threads, classes, and methods. An effective granularity will allow the system to expressively and efficiently migrate tasks. In vision, as most processing can be represented using a connected graph [11], it is simplest to partition the task between operational stages, as in [17]. For example, as illustrated in Fig. 6b, for marker-based pose estimation, the system can run feature detection and FLANN stages on the near-sensor processing unit and the remainder of the stages on the far-sensor host. To allow the system to change the partition split at runtime, we assume the system architect will provision for flexible near-sensor operation. Whether the near-sensor operation consists of fixed-function units or general-purpose instructions, the system will need to be able to run different streams of operations on the near-sensor processing unit.

In the remainder of this section, we discuss thermal-driven task migration challenges related to decision-making and performance assurance for migrating between partition splits.

4.1 Situationally aware partition decisions

Determining an effective partition split must balance the thermal asymmetry of the distributed processing units and the costs of data transaction. For example, partitioning at an early stage grants thermal relief, but only at the expense of communication cost. On the other hand, partitioning at the latter stages of the pipeline ensures less thermal relief for less communication overhead.

To approach this challenge, we draw inspiration from related works that dynamically partition applications to offload mobile tasks to remote servers for energy efficiency [6, 7, 17]. Along similar lines, a dynamic partitioning of the connected graph could optimize workload placement for energy-efficiency and/or performance while still satisfying sufficient task accuracy. Thus, we plan to characterize the performance, energy, and communication overhead of partitioning in the context of processing vision workloads across near-sensor and far-sensor processing units. Such partitioning should also utilize local optimizations, e.g., dynamic voltage-frequency scaling on the accelerator or host.

As in prior offloading works, we envision a runtime that will dynamically decide which task stages should be offloaded, driven by inputs from a task profiler and an optimization solver. Unlike other migration works, however, partitioning between near- and far-sensor processing must be guided by several conditions that affect image quality and task accuracy, as illustrated in Fig. 5. Partitioning decisions will directly influence near-sensor processing activity, which raises sensor temperature over time. However, image quality also depends on the lighting environment; dark scenes require sensors to use large exposure times and high analog gains, increasing motion artifacts and noise sensitivity. Given the mobile nature of vision applications, the split decision will need to continually adapt to continuously changing conditions.

4.2 Seamless task migration

Vision processing can be pipelined to allow different stages to process in parallel at high performance. When the system opts to change the partition split, e.g., from Fig. 6a to Fig. 6b, the migration will need to shift stages to run on the far-sensor host. This presents a challenge: how will the system provide the pipeline data of previous stages to a newly migrated partitioning scheme? Ideally, this should be seamless; there should be no drop in pipelined processing performance.

To complete task migration to a different partitioning split, the system must synchronize computational states between the near-sensor processing unit and the far-sensor host. This includes any dependencies generated by previous operational stages. It also includes the output of the last stage before the partitioning split. In Fig. 6b, the latter three stages will not be able to run until they are provided with the output of the previous stages. In our case study, this constitutes 24 KB, which would take 17 ms to transfer from near-sensor to far-sensor. Thus, we will face the challenge of mitigating latency as the near-sensor accelerator communicates data and the far-sensor host fills the pipeline.

Efficient pipeline utilization has been well explored, e.g., in branch prediction. Along similar lines, to keep pipeline fully occupied, we can speculatively predict thermal emergencies and begin to fill the pipeline to minimize performance delays.

Future work: predictive scheduler for migration

We plan to study effective mechanisms towards a scheduler that uses sensor temperature, environment conditions, and application fidelity requirements to guide partitioning decisions between near-sensor and far-sensor processing. For smooth performance, our scheduler will predict migration points in advance, and use early communication to hide migration latency. Through this predictive scheduler, the system will implement thermal-driven task migration to balance energy efficiency and task accuracy for vision workloads.

5 CONCLUSION

Near-sensor processing is the key to energy-efficient imaging and vision, as evidenced by recent academic and industrial efforts towards stacked image sensors. However, we show that near-sensor processing degrades vision tasks due to thermal noise, placing hard limits on the adoption of near-sensor processing. Thus, to balance efficiency and accuracy, we propose thermal-driven task migration to dynamically shift tasks between the thermally coupled near-sensor accelerator and the far-sensor host, based on environmental conditions. We will build on our early work through a deeper implementation-based study of thermal-driven task migration mechanisms. We will also pursue a richer investigation into the broader implications of near-sensor processing on a wider variety of vision tasks. Thermal-driven task migration will enable a future of energy-efficient continuous mobile vision through powerful near-sensor processing.

ACKNOWLEDGMENTS

The authors are grateful for comments made by anonymous reviewers and the paper shepherd Dr. Nic Lane. The authors thank Microsemi for their generous support through SmartFusion2 hardware kit and software licenses. This material is based upon work supported by the National Science Foundation under Grant No. 1657602.

REFERENCES

[1] Microsemi power estimators and calculators. https://www.microsemi.com/products/fpga-soc/design-resources/power-calculator.

[2] A. Alaghi, C. Li, and J. P. Hayes. Stochastic circuits for real-time image-processing applications. In *Proc. of Design Automation Conference (DAC)*, 2013.

[3] A. S. Arani. Online thermal-aware scheduling for multiple clock domain CMPs. In *IEEE Int. SOC Conference*, 2007.

[4] B. Barry, C. Brick, F. Connor, D. Donohoe, D. Moloney, R. Richmond, M. O'Riordan, and V. Toma. Always-on vision processing unit for mobile applications. *IEEE Micro*, 2015.

[5] D. Brooks and M. Martonosi. Dynamic thermal management for high-performance microprocessors. In *Int. Symp. High-Performance Computer Architecture (HPCA)*, 2001.

[6] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti. Clonecloud: elastic execution between mobile device and cloud. In *Proc. ACM Conf. Computer systems*, 2011.

[7] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl. Maui: making smartphones last longer with code offload. In *Proc. ACM Int. Conf. on Mobile systems, applications, and services*, 2010.

[8] Z. Du, R. Fasthuber, T. Chen, P. Ienne, L. Li, X. Feng, Y. Chen, and O. Temam. ShiDianNao: Shifting vision processing closer to the sensor. In *Proc. Int. Symp. on Computer Architecture.*

[9] A. El Gamal and H. Eltoukhy. CMOS image sensors. *IEEE Circuits and Devices Magazine*, 21(3), 2005.

[10] J. Fraden. *Handbook of Modern Sensors: Physics, Designs, and Applications.* Springer New York, 2010.

[11] Itseez. OpenVX Computer Vision Acceleration Standard. https://www.khronos.org/openvx/.

[12] H. Jung and M. Pedram. Stochastic dynamic thermal management: A Markovian decision-based approach. In *Int. Conf. Computer Design*, 2007.

[13] Y. Kagawa, N. Fujii, K. Aoyagi, Y. Kobayashi, S. Nishi, N. Todaka, S. Takeshita, J. Taura, H. Takahashi, Y. Nishimura, et al. Novel stacked cmos image sensor with advanced cu2cu hybrid bonding. In *Proc. of Electron Devices Meeting (IEDM)*, 2016.

[14] A. Kumar, L. Shang, L.-S. Peh, and N. K. Jha. System-level dynamic thermal management for high-performance microprocessors. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, 27, 2008.

[15] R. LiKamWa, Y. Hou, J. Gao, M. Polansky, and L. Zhong. Redeye: analog convnet image sensor architecture for continuous mobile vision. In *Proc. of Int. Symp. on Computer Architecture*, 2016.

[16] A. Merkel, F. Bellosa, and A. Weissel. Event-driven thermal management in SMP systems. In *Second Workshop on Temperature-Aware Computer Systems (TACS)*, 2005.

[17] M.-R. Ra, A. Sheth, L. Mummert, P. Pillai, D. Wetherall, and R. Govindan. Odessa: enabling interactive perception applications on mobile devices. In *Proc. ACM Int. Conf. Mobile systems, applications, and services*, 2011.

[18] K. Skadron. Hybrid architectural dynamic thermal management. 2004.

[19] Q. Xie, M. J. Dousti, and M. Pedram. Therminator: a thermal simulator for smartphones producing accurate chip and skin temperature maps. In *ACM/IEEE Int. Symp. Low Power Electronics and Design (ISLPED)*, 2014.

[20] B. Zhao, Z. Xu, C. Chi, S. Zhu, and G. Cao. Mirroring smartphones for good: A feasibility study. In *Proc. Int. Conf. on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, 2010.