Incidental Computing on IoT Nonvolatile Processors

Kaisheng Ma[1], Xueqing Li[1], Jinyang Li[2], Yongpan Liu[2], Yuan Xie[3], Jack Sampson[1], Mahmut Taylan Kandemir[1], and Vijaykrishnan Narayanan[1]

{kxm505, lixueq, sampson, kandemir, vijay}@cse.psu.edu, lijy15@mails.tsinghua.edu.cn, ypliu@tsinghua.edu.cn, yuanxie@ece.ucsb.edu

[1]*Dept. of Computer Science and Engineering, The Pennsylvania State University* [2]*Dept. of Electronic Engineering, Tsinghua University*

[3]*Dept. of Electrical and Computer Engineering, University of California at Santa Barbara*

Original paper was published in MICRO 2017 [1].

By this year 2017, there are 28B IoT devices, and by year 2020, there will be 50.1B IoT devices [2]. The per year increase is more than 20%. By year 2025, they will generate 11 trillion dollars value [2]. The shift from battery-powered systems to self-powered systems promises to fuel the next revolution in the *Internet of Things* (IoT). The ability to power IoT devices using ambient, scavenged energy liberates them from the lifetime, deployment, and servicing limitations of a fixed battery.

While ambient energy sources are notoriously fickle, concurrent advances in energy harvesting, ultra-low power computation, and non-volatile memory have enabled a new generation of processors, known as *non-volatile processors* (NVPs), which tightly integrate non-volatile memory elements into the logic fabric of the processor, thereby enabling almost instantaneous stopping and starting of execution via parallel distributed backup and restore functionality for processor state. For NVPs with microarchitectural hardware-managed backup, systems can make persistent progress even if only one instruction successfully completes between power interruptions.

Prior efforts on hardware-managed NVPs that perform local computation have focused on enhancing the efficiency of converting harvested energy into persistently executed instructions [4, 5]. These techniques primarily focused on (1) reducing the number and overheads of backups and restores and (2) adapting the compute architecture to exploit dynamic variations in incoming power.

However, the forward progress metric used in these works does *not* directly capture higher level application semantics regarding the "utility" of the work performed: *In many IoT applications, temporal and interactivity requirements can make the quality of partial results, or even the existence of any response at all, more important than the fraction of instructions needed to eventually produce a "best-quality" result.* Adding a "quality knob" provides flexibility in an NVP, where the need to make conservative decisions regarding energy reserves for backup operations can otherwise impose substantial overheads on execution. In an NVP, if the effort needed to ensure preservation of data is sufficiently reduced, some power emergencies may be avoided, improving response timeliness. Moreover, in addition to natural synergies with power management, accepting variable quality responses frees a harvesting system to apportion effort with respect to the continued relevance of the data being processed: If an NVP has been without power for some substantial time, resuming work on the input it was processing when power failed may have lower utility, from an application perspective, than moving on to processing the newest input.

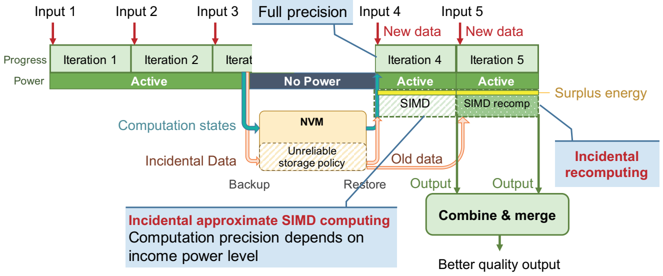

Figure 1: Incidental approximation concept

In this paper, we introduce **incidental approximate computation** to address opportunistic responsiveness versus quality tradeoffs under unstable power income, and implement and evaluate an instantiation of the incidental computing approach based on memory and datapath approximation within an NVP.

Noticing such a phenomenon that *In many deployment scenarios, catching up quickly after a power failure may take priority over the quality of response*, and another phenomenon that data importance drops over time, we first introduce the **incidental computing** concept for NVPs as shown in Figure 1. Instead of rolling back after power failure, incidental computing *rolls forward* to process the most recent and (most of the time) most important new data. If there is additional power available beyond that needed to process the new data, then older data will be processed at reduced quality. Incomplete executions from before a power failure are regarded as "incidental" and their importance drops over time. For the energy-harvesting NVP scenario explored in this paper, this is done through bitwidth-oriented approximation techniques in the datapath, memory, and backup-recovery modules to divide power and resources and provide differential guarantees of output quality between the current and prior computations. We also propose **incidental recomputing**, wherein the quality of older computations targeted for incidental computing can be gradually improved iteratively if picked up over multiple incidental computing passes.

[1] Kaisheng Ma, Xueqing Li, Jinyang Li, Yongpan Liu, Yuan Xie, Jack Sampson, Mahmut Taylan Kandemir, and Vijaykrishnan Narayanan. 2017. Incidental computing on IoT nonvolatile processors. In Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-50 '17). ACM, New York, NY, USA, 204-218.

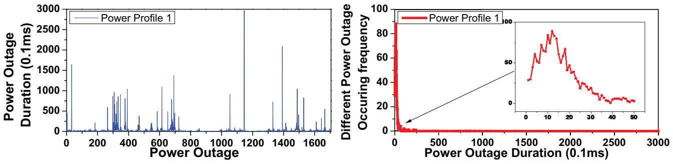

Figure 2: Power outage duration (left) and statistics (right)

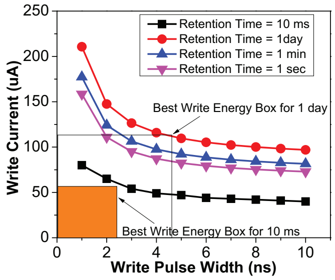

Figure 3: STT-RAM Write energy & retention time[3, 8]

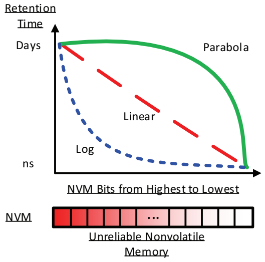
Retention Time
Days
ns
Parabola
Linear
Log
NVM Bits from Highest to Lowest
NVM
...
Unreliable Nonvolatile Memory

Figure 4: Retention Time Shaping (RTA)

Current NVPs [4, 6, 7] utilize nonvolatile technologies with maximum retention times on the order of a decade or more, and parameters tuned to maximize both retention and reliability. However, most power emergencies in wearable harvesting devices last just a few ms, and are rarely more than a fraction of a second, as shown in Figure 2. Approximate computing provides an opportunity to substantially mitigate these overheads by relaxing the reliability of the "lower order" NVM bits used to back up data during power emergencies, and using commensurately less energy for backup and recovery operations, as shown in Figure 3. We propose **incidental backup** with several retention time matching models and supporting write circuits that can reduce the energy of backup operations through matching the retention time to the combination of the duration of power emergencies and the impacts of reduced fidelity to overall approximation quality. We consider three retention time reduction functions to shape the retention time in a way that reduces from the most significant bit to the least significant bit, as shown in Figure 4.

Collectively, the incidental approximation approaches improve forward progress by 4.28x improvement within tolerable quality loss. 4X-5X forward progress vastly extend the application domain of energy harvesting IoTs, making some applications traditionally impossible due to limited harvested energy and high processor starting threshold, now possible.

REFERENCES

[1] J. Haj-Yihia, E. Weissmann, V. S. Degalahal, N. Shulman, T. Kuzi, I. Franko, A. Gur, and E. Rotem, "Autonomous c-state algorithm and computational engine alignment for improved processor power efficiency," Jul. 2 2014, uS Patent App. 14/322,185.

[2] S. Jankowski, J. Covello, H. Bellini, J. Ritchie, and D. Costa, "The internet of things: Making sense of the next mega-trend," *IoT primer*, 2014.

[3] A. Jog, A. K. Mishra, C. Xu, Y. Xie, V. Narayanan, R. Iyer, and C. R. Das, "Cache revive: architecting volatile stt-ram caches for enhanced performance in cmps," in *Proceedings of the 49th Annual Design Automation Conference*. ACM, 2012, pp. 243–252.

[4] Y. Liu, Z. Wang, A. Lee, F. Su, C. P. Lo, Z. Yuan, C. C. Lin, Q. Wei, Y. Wang, Y. C. King, C. J. Lin, P. Khalili, K. L. Wang, M. F. Chang, and H. Yang, "A 65nm reram-enabled nonvolatile processor with 6x reduction in restore time and 4x higher clock frequency using adaptive data retention and self-write-termination nonvolatile logic," *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 84–86, Jan 2016.

[5] K. Ma, Y. Zheng, S. Li, K. Swaminathan, X. Li, Y. Liu, J. Sampson, Y. Xie, and V. Narayanan, "Architecture exploration for ambient energy harvesting nonvolatile processors," *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, pp. 526–537, 2015.

[6] S. Senni, L. Torres, A. Gamatié, and G. Sassatelli, "Non-volatile processor based on MRAM for ultra-low-power IoT devices," *ACM Journal of Emerging Technologies in Computing (JETC)*, vol. 00, no. 00, 2017.

[7] S. Senni, L. Torres, G. Sassatelli, and A. Gamatie, "Non-volatile processor based on mram for ultra-low-power iot devices," *J. Emerg. Technol. Comput. Syst.*, vol. 13, no. 2, pp. 17:1–17:23, Dec. 2016.

[8] C. W. Smullen, V. Mohan, A. Nigam, S. Gurumurthi, and M. R. Stan, "Relaxing non-volatility for fast and energy-efficient stt-ram caches," in *High Performance Computer Architecture (HPCA), 2011 IEEE 17th International Symposium on*. IEEE, 2011, pp. 50–61.