AMNESIAC: Amnesic Automatic Computer

Trading off Computation for Storage for Energy Efficiency

Ulya R. Karpuzcu
Department of Electrical and Computer Engineering
University of Minnesota, Twin Cities
ukarpuzc@umn.edu

Problem: Growing (Static) Power

Processors of today are about to face a power wall: Each technology generation, more and more functionality – i.e. more and more computation engines – can be crammed into the unit chip area. Power consumption increases accordingly, however, cooling limitations prevent a proportional expansion of the power budget. As a result, we can simultaneously power-on, thus utilize, only a progressively diminishing fraction of the computation engines that we can integrate on chip. A growing fraction of the chip area will remain necessarily un-powered, aka *dark* [3].

We cannot engage a larger fraction of the chip area into computation without reducing power consumption. A promising way to achieve this is reducing the operating voltage V_{dd} aggressively to stay only slightly above the threshold voltage V_{th}. This unconventional regime of operation, *Near-threshold Voltage Computing* (NTC) renders power savings of 10-50$\times$ [1] which permit more computation engines to fit in the chip power envelope.

Power savings increase with the proximity of the operating voltage to the threshold voltage. As the V_{dd} reaches V_{th}, both components of power, dynamic and static, reduce, however, not at the same pace: static power decreases less. Hence, the share of static power grows, as depicted in Figure 1: For any given technology generation, as the operating voltage decreases from its nominal value by $0.75\times$, $0.5\times$, and $0.4\times$ (the latter two corresponding to NTC), the share of static power quickly increases. Aggravated by shrinking feature sizes, variability in design parameters intensifies this effect. Even worse, the impact of variability increases with decreasing V_{dd}. *Accordingly, any static-power-heavy operation (mainly storage) becomes more expensive than any dynamic-power-heavy operation (mainly computation).*

Solution: Trade-off Computation for Storage

At as low operating voltages as near-threshold, computation becomes cheaper than data storage in terms of power consumption. Hence, *re-computing data becomes cheaper than storing and retrieving pre-computed data.* Accordingly, we can maximize energy efficiency by re-generating – i.e. re-computing – data instead of storing and retrieving pre-computed data. Due to (i) re-computation consuming less power than storage; (ii) any power/performance overhead associated with data retrieval from storage – i.e. communication with memory – being minimized if not eliminated, an *amnesic* machine is expected to operate more energy-efficiently.

An amnesic machine accommodates even less storage than any program to execute on the machine demands. This limited storage of the machine represents the *short-term* memory of the amnesic system, which lacks long-term memory by construction. When compared to its non-amnesic, classic counterparts, an amnesic machine can facilitate more computation engines to occupy the area once devoted to (long-term) memory. *How to unlock the energy efficiency premise of amnesic operation?*

Amnesic Machines: Working Hypotheses

- **How to determine the size of the short-term memory?** The algorithm – how machine operations are composed in performing program-specific tasks – dictates the maximum amount of memory each program requires. An amnesic machine would perform best under amnesic algorithms – i.e. algorithms of minimal memory demand to exploit re-computation. In mapping conventional algorithms to amnesic machines, producer-consumer chains should be tracked to identify data inputs necessary for re-computation. Such inputs can either reside in short-term memory, or be re-computed themselves as outputs of previous steps in computation.

- **How to bound the performance overhead of amnesic execution?** Re-computation incurs a performance overhead, and eventually may turn out to be slower than retrieval of pre-computed data. This compromise should be carefully analyzed. Ideally, the machine should be able to dynamically adjust the level of amnesia, i.e. how much storage should be allocated as short-term memory – equivalently, to what extent data should be re-computed.

- **How to orchestrate re-computation and data storage/retrieval?** Depending on how often a given chunk of data gets re-used, and how long it would take to re-compute it, amnesic operation may not always be desirable. We can rely on a learning controller with prediction capabilities to impose amnesic operation only if it offers higher energy efficiency.

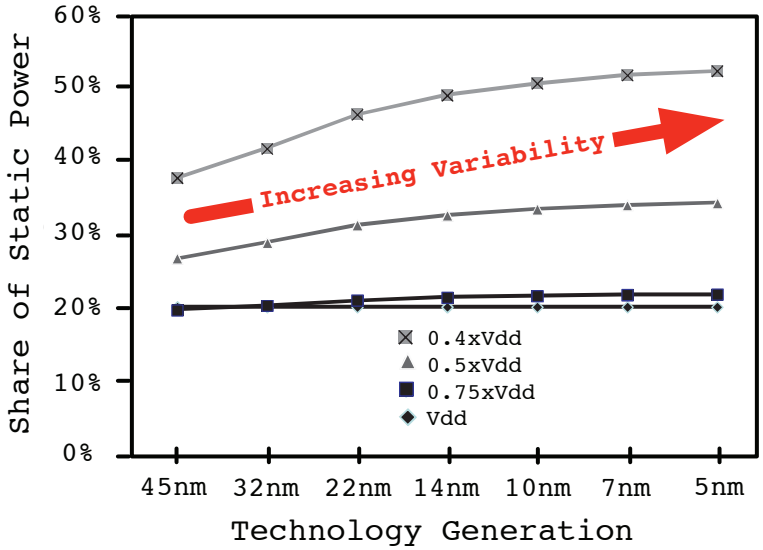

Figure 1: Evolution of % share of static power in total power consumption with technology generations [2].

- **Emerging memory technologies?** The discussion so far assumed state-of-the-art memory technologies. Emerging memory technologies to practically eliminate static power have been proposed, however, current manufacturing restrictions complicate tight integration with computation engines. Even if such restrictions were overcome, amnesic operation would still be beneficial: it can both reduce the memory footprint and minimize power-hungry and slow communication with memory.

- **Big data?** Future programs are likely to deal with increasing amounts of data. Increasing amounts of data, however, do not necessarily translate into a requirement for larger and larger storage. The key lies in how much data get processed at a given time. This amount would dictate a loose upper bound on the size of the short-term memory of an amnesic system.

References

[1] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge. Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits. *Proceedings of the IEEE*, 98(2), February 2010.

[2] Himanshu Kaul, Mark Anders, Steven Hsu, Amit Agarwal, Ram Krishnamurthy, and Shekhar Borkar. Near-threshold Voltage (NTV) Design: Opportunities and Challenges. In *Design Automation Conference*, June 2012.

[3] M.B. Taylor. Is Dark Silicon Useful? Harnessing the Four Horsemen of the Coming Dark Silicon Apocalypse. In *Design Automation Conference*, June 2012.