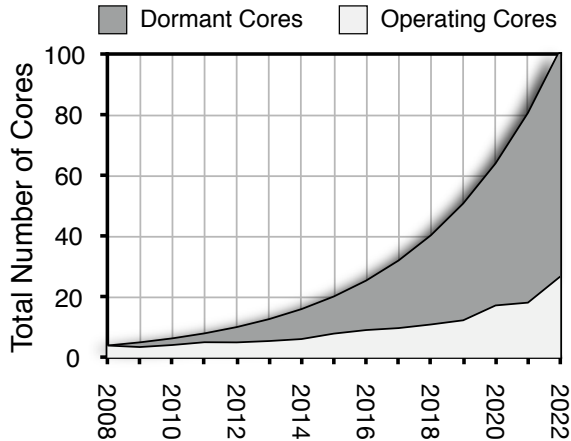


# BubbleWrap: Popping CMP Cores for Sequential Acceleration

Brian Greskamp, Ulya R. Karpuzcu, and Josep Torrellas

Department of Computer Science, University of Illinois at Urbana-Champaign

## Problem: The CMP Power Wall



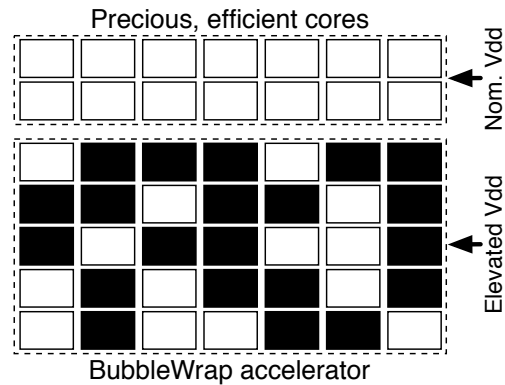
Many-core CMP scaling is about to crash into the power wall. Using data synthesized from the 2008 ITRS update, the figure above shows a growing gap between the number of Nehalem-like cores that fit on a die (about 100 cores by 2022) and the number of cores that can operate simultaneously under a 100W power budget (only 27 cores by 2022). In future designs, the majority of cores will necessarily be *dormant* at any given time in order to meet the power budget. Concerns over programmability and parallelism aside, the future literally provides more cores than we can possibly use. *How can we exploit this surplus of cores to extract more performance?*

## Solution: Expendable Cores

Process variation in future CMPs will render some cores more power efficient than others. The most efficient cores are *precious* because parallel workloads can use them to run as many threads as possible under the power limit. Less-efficient cores can be used as “Bubble Wrap”: They accelerate the sequential phases, sacrificing themselves to defeat Amdahl’s Law by operating at higher-than-normal voltage and frequency. The elevated voltages and temperatures will quickly wear out or “pop” BubbleWrap cores, but this is not a problem, as they can be replaced from a large pool of dormant spares. Meanwhile, the *precious* cores never run above nominal voltage, and so protected from the harsh conditions under which the BubbleWrap cores operate, lead long and efficient lives.

The following figure illustrates our vision of a BubbleWrap CMP from year 2019. It includes 14 *precious* cores that operate only at nominal voltage and are guar-

anteed not to burn out over the service life of the chip. It also includes 35 *BubbleWrap* cores that are microarchitecturally identical to the precious ones but are destined to live fast and die young. Any cores not currently operating are power-gated. The figure shows a chip in mid-life, when some of the BubbleWrap cores are already popped (black).



The system software, responding to user demands, has full discretion on how to use the BubbleWrap cores. It could spend them all in one day if, facing a critical deadline, the user requests performance at all costs. In any case, it continually tracks the aging state of each core, and when one is about to pop, migrates its work elsewhere. When all BubbleWrap cores are popped, the system can still provide a satisfactory and guaranteed performance level over the remaining service life using only the *precious* cores.

If the BubbleWrap cores are used to accelerate a single thread and are rationed out regularly over the service life of the processor, then each of the 35 BubbleWrap cores in our example chip must last only one thirty-fifth of the chip’s overall lifetime — or only about three months for a typical ten year chip service life. This allows a significant increase in supply voltage and therefore frequency. In a 45nm process, a three month service life affords a 15% increase in operating voltage over that admitted by a ten year service life and a 45% increase over nominal voltage [1]. As technology scales, the number of dormant cores increases, and the service life demanded of each BubbleWrap core declines. Consequently, the benefits of BubbleWrap acceleration increase as technology advances. Moreover, BubbleWrap requires no modifications to the core design.

## References

- [1] H.T. Huang et al. 45nm high-k / metal-gate CMOS technology for GPU / NPU applications with highest PFET performance. In *International Electron Devices Meeting*, December 2007.