DOI:10.1145/2408776.2408796

Technical Perspective
Is Dark Silicon Real?

By Pradip Bose

THE MICROPROCESSOR CHIP R&D community has been well aware of the so-called "power wall" challenge for over a decade now. Researchers have focused mainly on creative techniques to improve power-performance efficiency. Developers have adopted many of those ideas, and through engineering innovations have been able to keep the economics of technology scaling largely justifiable up until now. Industry witnessed a clear paradigm shift (as a response to the looming power wall) when the single-core processor chip era gave way to the multicore era at the beginning of the current century. Power (and power *density*) limits, coupled with the steady demise of idealized Dennard scaling rules, made it difficult to keep increasing the clock frequency. Also, limits in instruction-level parallelism (ILP) made it difficult to keep increasing single-core instructions-per-cycle (IPC), without spending an inordinate amount of area and power. However, even though we embarked upon the multicore trail, the power wall was never forgotten. We knew that sacrificing single-thread performance in favor of generational increases in chip throughput would not make the power wall go away forever! It would loom large again as the core count kept increasing. Because, fundamentally, the memory and I/O bandwidth demands dictated by the need to "feed" so many cores costs power and pins that we do not have. Also, delivering the current to feed an increasing number of cores at a lower voltage than before makes the designer hit a chip C4 current limit wall that is difficult to ignore.

The following work by Esmaeilzadeh et al. is a landmark paper that opens our eyes to the unrelenting power challenge we face in the multicore era. Most interestingly, it raises the specter of *dark silicon*: lots of processor cores, but very few that can be powered on or utilized at any given time. Not that the authors see this as a desirable feature of future designs; but they certainly raise a very valid question about the future viability of the basic multicore paradigm. Just like ILP limits make it ever harder to boost single-thread IPC at affordable power and complexity, thread-level parallelism (TLP) at the chip level gets ever harder because of the limited parallelism in so-called parallel applications. And, even for some scientific applications that are embarrassingly parallel or for commercial server workloads with large TLP, on-chip shared hardware resource contention and size limits make it more difficult to extract that parallelism at affordable power. So, even if we are able to go on doubling the number of cores each technology generation, we have two basic problems, as clearly enunciated by the authors: for a fixed chip power budget and area, even a very aggressive investment in application parallelism enhancement does not help one get even respectably close to the targeted 2X (throughput) performance growth per generation; and even if cooling and power delivery technologies improve to allow a large increase in the chip power budget, real application parallelism levels would not allow targeted performance scaling in most cases—not by a long shot. The paper's elegant analytical formalism shows that under ITRS projections, as we approach the 8nm technology node, over half the chip will remain unutilized (and consequently "dark"). In a sense, this is regardless of whether one views the problem from the perspective of a power wall constraint or from one that focuses first on the effective TLP limit constraint.

The effective parallelism content of real application workloads is often small enough that strong single-thread performance remains a crucial factor to combat the (serial) Amdahl bottleneck. The paper, therefore does consider heterogeneous (or asymmetric) multicores in the analysis in a quest to find an optimistic outlook for the future. However, the combination of realistic chip power limits and real application parallelism limits makes it hard or impossible to sustain historically established performance growth rates using the multicore paradigm as we currently know it.

Is the specter of progressively darker silicon real? Or, are there technological or design breakthroughs around the corner to help us circumvent such a scenario, at least in the short term? Alternatively, if that specter is indeed real from a utilization efficiency viewpoint, but not directly from a power limit perspective, are there other ways the "idle" cores can be used to provide functionality that is not traditional "performance"? For example, can available idle cores be used to enhance reliability or security? The paper does briefly journey into optimistic dreamland to give the reader a hint about promising new innovations that could potentially be disruptive in the face of the specter that seems to be haunting us at this time.

This paper is not just a doomsday predictor. It raises our awareness of the problem through scientific quantification; but it should also serve as a springboard for innovative research, especially for computer architects. However, the architect cannot hope to invent in a vacuum; the needed innovations will surely come, but only by adopting a holistic, cross-layer view of the full system—from devices, through circuits, microarchitecture, system architecture, and the software stack. Researchers are well-aware of this urgent need, thanks to papers like this one; the industrial development teams cannot wait to take advantage of the next generation of holistic, cross-layer system architectural thoughts, models, and design tools. C

Pradip Bose (pbose@us.ibm.com) is a research staff member at IBM T.J. Watson Research Center where he manages the Reliability- and Power-Aware Microarchitectures department.