Is Dark Silicon Useful?

Harnessing the Four Horsemen of the Coming Dark Silicon Apocalypse

Michael B. Taylor
Computer Science & Engineering Department
University of California, San Diego

ABSTRACT

Due to the breakdown of Dennardian scaling, the percentage of a silicon chip that can switch at full frequency is dropping exponentially with each process generation. This utilization wall forces designers to ensure that, at any point in time, large fractions of their chips are effectively dark or dim silicon, i.e., either idle or significantly underclocked.

As exponentially larger fractions of a chip's transistors become dark, silicon area becomes an exponentially cheaper resource relative to power and energy consumption. This shift is driving a new class of architectural techniques that "spend" area to "buy" energy efficiency. All of these techniques seek to introduce new forms of heterogeneity into the computational stack. We envision that ultimately we will see widespread use of specialized architectures that leverage these techniques in order to attain orders-of-magnitude improvements in energy efficiency.

However, many of these approaches also suffer from massive increases in complexity. As a result, we will need to look towards developing pervasively specialized architectures that insulate the hardware designer and the programmer from the underlying complexity of such systems. In this paper, I discuss four key approaches – the four horsemen – that have emerged as top contenders for thriving in the dark silicon age. Each class carries with its virtues deep-seated restrictions that requires a careful understanding of the underlying tradeoffs and benefits.

Categories and Subject Descriptors B.7.1 [*Integrated Circuits*]: Types and Design Styles

General Terms Design, Performance, Economics

Keywords Dark Silicon, Multicore, Dim Silicon, Utilization Wall, Dennardian Scaling, Near Threshold, Specialization

1. INTRODUCTION

Recent trends in VLSI technology have led to a new disruptive regime for digital chip designers, where Moore's Law continues but CMOS scaling ceases to provide the fruits that it once did. As in prior years, the computational capabilities of chips are still increasing by 2.8× per process generation; but a *utilization wall* [26] limits us to only 1.4× of this benefit – resulting in large swaths of our silicon area remaining underclocked, or dark – hence the term *dark silicon* [19, 8].

These numbers are easy to derive from simple scaling theory, which is a good thing, because it allows us to think intuitively about the problem. Transistor density continues to improve by 2× every two years, and native transistor speeds improve by 1.4×. But energy efficiency of transistors is improving only by 1.4×, which, under constant power-budgets, results in a 2× shortfall in energy budget to power a chip at its native frequency. Therefore, our rate of utilization of a chip's potential is dropping *exponentially* by a jaw-dropping 2× per generation. Thus, if we are just bumping up against the dark silicon problem in last generation's product line, then in eight years, we will be faced with designs that are 93.75% dark!

In the title of this paper, we refer to this widespread disruptive factor informally as the *dark silicon apocalypse*, because it officially marks the end of one reality ("Dennardian Scaling") – where progress could be measured by improvements in transistor speed and count – and the beginning of a new reality ("post-Dennardian Scaling") – where progress is measured by improvements in transistor energy efficiency. In the past, we tweaked our circuits to reduce transistor delays and turbo-charged them with dual-rail domino to reduce FO4 delays. In the new regime, we will tweak our circuits to minimize capacitance switched per function; we will strip our circuits down and starve them of voltage to squeeze out every femtojoule. Where once we would spend exponentially increasing amounts of silicon area to buy performance, **now, we will spend exponentially increasing amounts of silicon area to buy energy efficiency.**

A direct consequence of this breakdown in CMOS scaling is the industrial transition to multicore in 2005. Because filling chips with cores does not circumvent utilization wall limits, multicore is not the final solution to dark silicon [8] – it is merely industry's initial, liminal response to the shocking onset of the dark silicon age. Wikipedia defines liminality as "an in-between situation characterized by the reversal of hierarchies, and uncertainty regarding the continuity of tradition and future outcomes". With multicore, industry as a whole was uncertain as to the ramifications and scale of the power problems it was going to have, but it knew it needed to do something to address the problem. Over time, in this liminal phase, we are realizing more and more the

DAC 2012, June 3-7, 2012, San Francisco, California, USA.

ramifications and the semiconductor community as a whole is coming to a realization of what the new regime holds.

Due to the breakdown of Dennardian Scaling, multicore chips will not be able to scale with die area; the fraction of a chip that can be filled with cores running at full frequency is dropping exponentially with each process generation [8, 26]. This reality will force designers to ensure that, at any point in time, large fractions of their chips are effectively *dark* or *dim* – either idle or significantly underclocked. As exponentially larger fractions of a chip's transistors become dark transistors, silicon area becomes an exponentially cheaper resource relative to power and energy consumption. This shift calls for new architectural techniques that "spend" area to "buy" energy efficiency.

In this paper, we examine some of the potential approaches that are coming to light about the dark silicon regime. We start by recapping the utilization wall that is the cause of dark silicon in Section 2, and by examining why the multicore response to the utilization wall is inherently limited [8]. We will look at recently proposed responses that are emerging as solutions as we transition beyond the transitional multicore stop-gap solution. Looking back, all of these responses appeared to be unlikely candidates from the beginning, carrying unwelcome burdens in design, manufacturing, and programming. None would appear ideal from an aesthetic engineering point of view (hence the analogy to the "four horsemen"). But the success of complex multi-regime devices like MOSFETs has taught us that engineering as a field has an enormous tolerance for complexity if the end result is better. As a result, we believe that future chips will apply not just one of these alternatives, but all of them.

In Section 3 we examine perhaps the most grim of the four candidates, which we refer to as *shrinking silicon*: simply scaling down the size of chips to reduce the amount of dark silicon. Section 4 examines the promise of underclocked, or dim silicon. Section 5 discusses the promise of specialized coprocessors in dark silicon dominated technology. Finally, we examine the promise of new classes of circuits in Section 6, before concluding.

2. THE UTILIZATION WALL THAT CAUSES DARK SILICON

In this section, we show that a *utilization wall* [26] is the cause of dark silicon [19, 8]. Table 1 shows how this utilization wall is derived. We employ a scaling factor, S, which is the ratio between the feature sizes of two processes (e.g., $S = 32/22 = 1.4x$ between 32 and 22 nm process generations.) In both Dennardian and Post-Dennardian (Leakage-Limited Scaling), transistor count will scale by S^2, and transistor switching frequency will scale by S. Thus our net increase in compute performance from scaling is S^3, or $2.8x$.

However, to maintain a constant power envelope, these gains must be offset by a corresponding reduction in transistor switching energy. In both cases, scaling reduces transistor capacitance by S, improving energy efficiency by S. In Dennardian Scaling, we are able to scale the threshold voltage and thus the operating voltage, which gives us another S^2 improvement in energy efficiency. However, in today's Post-Dennardian, leakage-limited regime, we cannot scale threshold voltage without exponentially increasing leakage, and as a result, we must hold operating voltage roughly constant. The end result is that today, we have a shortfall of S^2, or 2× per process generation. *This is an exponentially worsening problem that accumulates with each process generation.*

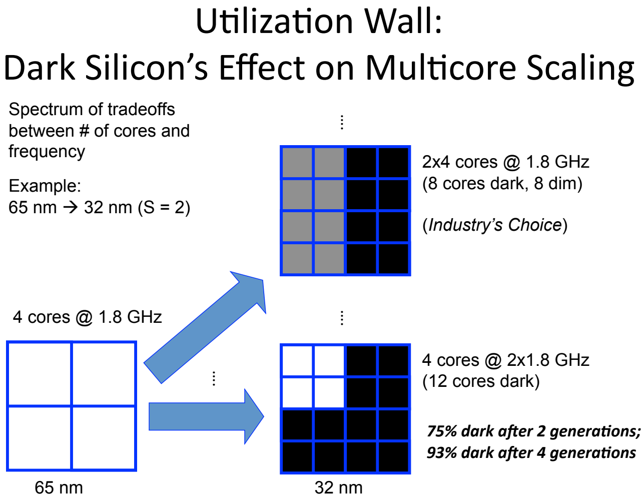

Figure 1: Multicore scaling leads to large amounts of Dark Silicon. (From [8].)

2.1 Silicon's New Potential: 40% energy savings per generation, or 1.4x performance

It is this shortfall that causes problems with multicore as the solution to scaling [8, 26]. Although we have enough transistors to increase the number of cores by 2×, and they would run 1.4× faster, we only have the energy budget to receive a 1.4× improvement. As shown in Figure 1, across two process generations ($S = 2$), we can either increase core count by 2×, or frequency by 2×, or some middle ground between the two. The remaining 4× potential goes unused. This 4× reflects itself in either dark or dim silicon, depending on our preference for frequency versus core count. A quick survey of recent designs such as Tilera TileGx, Intel Gulftown and Nvidia Fermi shows that industry has pursued various combinations of core count increase, and frequency increase/decrease that correlate very closely with the utilization wall. (For subsequent work to [8, 26] on dark silicon and multicore scaling that explores more sophisticated models that incorporate factors such as application space and cache size, see [6, 12, 15].)

3. THE SHRINKING HORSEMAN (#1)

When confronted with the possibility of dark silicon, an immediate response of many chip designers is "Area is expensive. Chip designers will just build smaller chips instead of having dark silicon in their designs!" Of all of the four dark horses, we believe that *shrinking chips* are the most pessimistic outcome, and although all chips may eventually experience "shrinkage", the ones that shrink the most will be those for which dark silicon cannot be applied fruitfully to actually result in a better product, and will rapidly turn into low-margin businesses for which further generations of Moore's Law provide small benefit. To examine this question in further detail, we look at a spectrum of second-order effects associated with shrinking chips.

Transistor Property	Dennardian	Post-Dennardian
Δ Quantity	S^2	S^2
Δ Frequency	S	S
Δ Capacitance	$1/S$	$1/S$
Δ V_{dd}^2	$1/S^2$	1
$\implies$ Δ Power = Δ $QFCV^2$	1	S^2
$\implies$ Δ Utilization = 1/Power	1	$1/S^2$

Table 1: Dennardian vs. Post-Dennardian (leakage-limited) scaling In contrast to the Classical regime proposed by Dennard [4], under the Post-Dennardian regime, the total chip utilization for a fixed power budget drops by a factor of S^2 with each process generation. The result is an exponential increase in the quantity of dark silicon for a fixed-sized chip under a fixed area budget. From [26].

Misconceptions of Dark Silicon. First, it is worth saying that dark silicon does not mean blank, useless or unused silicon – it is just silicon that is not used all the time, or at its full frequency. Even during the best days of CMOS scaling, microprocessor and other circuits were chock full of "dark logic" that is used only for some applications – for example, SIMD SSE units on x86 processors are not used for irregular applications and a doubling of last-level cache conveys benefits only for a narrow band of applications for which 1) cache misses comprise a major percent of program execution time and 2) a large fraction of the working set suddenly fits. For instance, many streaming applications experience no benefit from today's last level caches. And for SSE functional units, and especially last-level caches, this logic is often not used every cycle even in programs that do make use of them, which makes them "dark-silicon friendly".

Going into the future, the exponential growth of dark silicon area will push us beyond logic targeted for direct performance benefits towards swaths of low-duty cycle logic that exists not for direct performance benefit, but for the purpose of improving energy efficiency, which causes an indirect improvement in performance because it frees up more of the fixed power budget.

Cost Side of Shrinking Silicon. Understanding shrinking chips calls for us to consider semiconductor economics. There is a ring of truth to the "build smaller chips" argument – after all, designers spend much of their time trying to meet area budgets for existing chip designs. Smaller chips are generally cheaper, their leakage should be lower depending on power-gate efficiency, and in the small-signal regime of design optimization they are cheaper linearly (or better) with area. But *exponentially smaller chips are not exponentially cheaper* – even if they start out at being 50% of the cost of the system, after a few process generations, the cost of the silicon will be a tiny fraction of packaging and test costs, let alone system, marketing, sales, support and other costs. (For instance, for a typical $100mm^2$ desktop processor die, the silicon cost itself is only $10 or so, but the chip sells at $100-$300.) I/O pad area, design costs, masks costs will fail to be amortized, leading to a rising cost per mm^2 of silicon that ultimately will result in the lack of incentive to move the design to the next process generation. These designs will be "left behind" on older generations. (If this happens at a large scale across too many designs, then fab construction costs would be ammortized more slowly as wafer quantities plummeted, and fab investments would become less attractive relative to alternative investments, signifying an unhappy economic ending to Moore's Law ...)

Revenue Side of Shrinking Silicon. On the other side of the shrinking silicon is the selling price of the chip. In a competitive market, if there is a way to use the next process generation's bounty of dark silicon to attain a benefit to the end product, then competition will force companies to do it. Otherwise, they will generally be forced into the low-end, low-margin, high-competition part of the market and their competitor will take the high end and enjoy high margins and achieve market superiority, much as happened with AMD and Intel in recent years [20]. Thus, in scenarios where dark silicon can be used profitably, decreasing area in lieu of exploiting it would certainly decrease system costs, *but these decreases in area would have much more catastrophic effects on sale price*, if it results in compromised performance or functionality relative to the competition. Thus, the shrinking chips scenario is likely to happen only if we can find no practical use for dark silicon.

Power and Packaging Issues with Shrinking Chips. A major consequence of exponentially shrinking chips is a corresponding exponential rise in power density. Recent work in analyzing the thermal characteristics of manycore chips [14] has shown that peak hotspot temperature rise can be modeled as $T_{max} = TDP \times (R_{conv} + k/A)$, where T_{max} is the rise in temperature, TDP is the target thermal design power of the chip, R_{conv} is the heatsink thermal convection resistance (lower is a better heatsink), k incorporates manycore design properties, and A is the area of the chip. If area drops exponentially, then the second term dominates and chip temperatures will rise exponentially. This in turn will force a lowering of the TDP so that temperature limits are met and reduce scaling below even the nominal 1.4× gain expected from energy efficiency gains. Thus, if thermals drive your shrinking chip strategy, it is much better to hold your frequency constant and increase cores by 1.4× with a net area decrease of 1.4× than it is to increase your frequency by 1.4× and shrink your chip by 2×. On the other hand, there is a concern that even without shrinking chips, the power-density of hotspots is still increasing exponentially, and could be a concern. A recent paper [15] suggests that this is not a significant concern, because as the hotspots shrink, the heat transfer to neighboring non-hotspots becomes proportionally more efficient.

Shrinking chips also present a host of practical engineering issues. Barring scalable innovations in 3-D integration along the lines of through-silicon vias (TSVs), designs would be increasingly pin-limited and would have trouble shrinking even though transistor area is shrinking, since I/O pads have not scaled well with Moore's Law.

4. THE DIM HORSEMAN (#2)

If we move beyond the prospect of shrinking silicon and consider populating dark silicon area with logic that we only use part of the time, then we are faced with two choices: do we try to make the logic in question general-purpose, or special purpose? In this section, we look at low-duty cycle alternatives that try to retain general applicability across many applications. We employ the term *dim silicon* [23, 15] to refer to general-purpose logic that is typically underclocked or used infrequently to meet the power budget.

Dim silicon techniques include scaling up the amount of cache logic, employing near-threshold voltage (NTV) processor designs, using Coarse-Grained Reconfigurable Array (CGRA)-based architectures that attempt to reduce energy by reducing the multiplexing of processor datapaths, and employing temporal dimming techniques.

Near-Threshold Voltage Processors. One recently emerging approach is the use of near-threshold voltage (NTV) logic [5], which operates in the near-threshold regime, providing more less-extreme tradeoffs between energy and delay than conventional subthreshold circuits.

Recently, researchers have looked at wide-SIMD implementations of NTV processors [18, 13] which seek to exploit data-parallelism, the most energy-efficient form of parallelism, and also a NTV many-core implementation [2] and an NTV x86 (IA32) implementation [17].

Although per-processor performance of NTV processors drops faster than the corresponding savings in energy-per-instruction (say a 5× energy improvement for a 8× performance cost), the performance loss can be offset by using 8× more processors in parallel if the workload allows it.

So assuming perfect parallelization, NTV could offer 5× the throughput improvement while absorbing 40× the area – approximately eleven generations of dark silicon. If 40× more free parallelism exists in the workload relative to the parallelism "consumed" by an equivalent energy-limited super-threshold manycore processor, it is a net win to employ NTV in deep-dark silicon limited technology. As we will see with specialization, the more energy-limited the domain (i.e. runs off a small solar panel or battery), the less total parallelism in the workload needed to break even, and thus the broader the applicability across workloads.

NTV presents a variety of circuit-related challenges that have seen active investigation, especially because technology scaling is likely to exacerbate rather than ameliorate these factors. A significant challenge with NTV has been susceptibility to process variability. As the operating voltage is dropped, variation in transistor threshold due to random dopant fluctation (RDF) is proportionally higher, and the variation in operating frequency can vary greatly. Since NT designs expand the area consumption of designs by $\sim 8\times$ or more, variation issues are exacerbated, especially in SIMD machines which typically have tightly synchronized lanes. Recent efforts have looked at making SIMD designs more robust to these variations [24, 18]. Other challenges include the penalties involved in designing SRAMs that can operate at lower voltages and the increased energy consumption due to longer interconnect caused by the spreading of computation across a large number of slower processors.

Bigger Caches. An often proposed dim-silicon alternative is to simply use dark silicon area for caches. We can imagine, for instance, expanding per-core cache at a rate that soaks up the remaining dark silicon area; at a rate of $1.4 - 2\times$ more cache per core per generation. Increased cache sizes can carry both performance and energy benefits for miss-intensive applications, since off-chip accesses are power hungry. The miss-rate of the workload is a key parameter in determining the optimality of increasing cache size.

Going into the future with lower power off-chip interfaces and 3-D integrated memories, the benefits of larger on-chip caches are likely to be reduced; according to a recent study on dark-silicon limited server workloads, one crossover point for server workloads is when caches become large enough that the system ceases to be bandwidth-limited [12] and becomes power-limited.

Coarse-Grained Reconfigurable Arrays. One recurring "dim alternative" is the use of reconfigurable logic. Since the bit-level granularity and long wires of conventional bit-level FPGAs usually incurs high energy overheads, the most promising option is coarse-grained reconfigurable arrays (CGRAs) which have optimized paths for word-level operations. The idea is to naturally lay out the datapaths of the computation in space to avoid the multiplexing costs that are inherent to processor pipelines. The duty cycle of CGRA elements is very low, making it a potential fit exploiting dark silicon. Research in CGRAs has been ongoing prior to the days of dark silicon [28, 7] and continues into the dark silicon era [11]. Commercial success has been limited, but new constraints often make us look at old designs with fresh eyes.

Computational Sprinting and Turbo Boost. Other techniques work through the use of "temporal dimness" as opposed to "spatial dimness", temporarily exceeding the nominal thermal budget but relying on thermal capacitance to buffer against temperature increases, and then ramping back to a comparatively dark state. Intel's Turbo Boost 2.0 [22] uses this approach to boost performance up until the processor reaches, nominal temperature, relying upon the innate capacitance of the heatsink. Computational Sprinting [21] takes this a step further, by proposing the use of phase-change materials to allow chips to exceed their sustainable thermal budget by an order of magnitude or more for sub-second durations, providing a short but substantial computational boost.

5. THE SPECIALIZED HORSEMAN (#3)

As exponentially larger fractions of a chip's transistors become dark transistors, silicon area becomes an exponentially cheaper resource relative to power and energy consumption. This shift calls for new architectural techniques that "spend" area to "buy" energy efficiency. One approach is to use this dark silicon to implement a host of specialized co-processors, each of which is either much faster or much more energy-efficient (100–1000×) than a general purpose processor [26]. Execution hops among coprocessors and general purpose cores, executing where it is most efficient. At the same time, the unused cores are power- and clock- gated to keep them from consuming precious energy.

The promise for a future of widespread specialization is already being realized: we are seeing a proliferation of specialized *accelerators* that span diverse areas such as baseband processing, graphics, computer vision, and media coding. These accelerators enable orders-of-magnitude improvements in energy-efficiency and performance, especially for computations that are highly parallel. Recent proposals [26, 12] have extrapolated this trend and anticipate that in the near future we will see systems that are comprised of more coprocessors than general-purpose processors. In this paper, we term these systems Coprocessor Dominated Architectures, or *CoDAs*.

As the use of specialization grows to combat the problem of dark silicon, we are faced with the reality of a modern-day specialization "tower-of-babel" crisis that fragments our no-

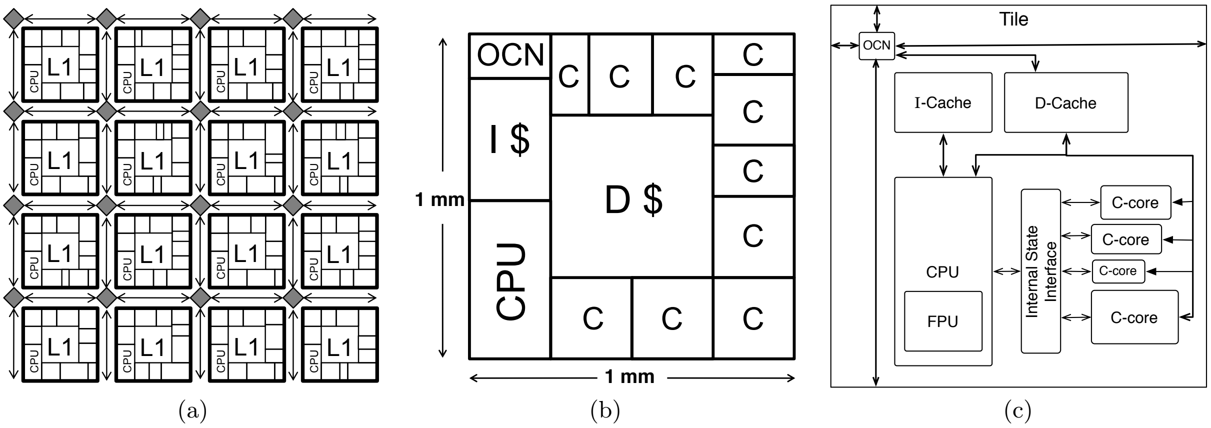

Figure 2: The GreenDroid architecture, an example of a Coprocessor-Dominated Architecture (CoDA). The GreenDroid Mobile Application Processor (a) is made up of 16 non-identical tiles. Each tile (b) holds components common to every tile—the CPU, on-chip network (OCN), and shared L1 data cache—and provides space for multiple c-cores of various sizes. (c) shows connections among these components and the c-cores.

tion of general purpose computation and eliminates the traditional clear lines of communication that we have between programmers and software and the underlying hardware. Already, we see the deployment of specialized languages such as CUDA that are not usable between similar architectures (e.g. AMD and Nvidia). We see over-specialization problems between accelerators that cause them to become inapplicable to closely related classes of computations (e.g. double-precision scientific codes running incorrectly on GPU floating-point hardware that has been specialized for graphics.) We also see problems with adoption due to the excessive costs of programming heterogeneous hardware (e.g., the slow uptake of Sony Playstation 3 due to the difficulty of porting games to exploit the Cell Processor.) Specialized hardware also runs the risk of being obsolete as standards are revised (e.g., an update of the JPEG standard.)

Insulating Humans from Complexity. All of these factors speak to potential exponential increases in the human effort required to both design and program these CoDAs. Combating the tower-of-babel problem requires that we define a new paradigm for how specialization is expressed and exploited in future processing systems. We need new scalable architectural schemas that employ pervasively specialized hardware to minimize energy and maximize performance while at the same time insulating the hardware designer and the programmer from the underlying complexity of such systems.

Overcoming Amdahl-Imposed Limits on Specialization. Amdahl's Law provides an additional roadblock for specialization. The issue is that we need to find broad-based specialization approaches that save energy across the majority of the computation in question, including not only regular, parallel code, but also irregular code. One such CoDA-based system that targets both irregular and regular code is the UCSD GreenDroid processor [8, 9, 25, 10], which is a mobile application processor that targets the hotspots of the Android mobile environment using hundreds of specialized cores called *conservation cores*, or *c-cores* [26, 27, 23]. C-cores are automatically generated from C/C++ source code, and support a patching mechanism that allows them to track software changes. They attain an estimated $\sim 8-10\times$ improvement in energy efficiency, at no loss in serial performance, even on non-parallel code, and without any user intervention required.

In contrast to Near-Threshold Voltage Processors, there is no need to find additional parallelism in the workload in order to cover a serial performance loss. As a result, conservation cores are likely to work across a wider range of workloads including collections of serial programs. However, for highly-parallel workloads where execution time is loosely concentrated, Near-Threshold Voltage Processors may hold an area advantage due to their reconfigurability.

6. THE DEUS EX MACHINA HORSEMAN (#4)

Of the four horsemen, this is by far the most unpredictable. *Deus Ex Machina* refers to a plot device in literature or theatre in which the protagonists seem utterly doomed, and then something completely unexpected and unforeshadowed comes out of nowhere to save the day. In the case of dark silicon, one Deus Ex Machina would be a breakthrough in semiconductor devices. However as we shall see, the breakthroughs that would be required would have to be quite fundamental – in fact most likely would require us to build transistors out of devices other than MOSFETs. The reason is that leakage is set by fundamental principles of device physics, and is limited to a sub-threshold slope of 60 mV/decade at room temperature; that is, in the typical case, a reduction of 10× for every 60 mV that the threshold voltage is above the V_{ss}, which is determined by properties of thermionic emission of carriers across a potential well. Thus, although innovations like Intel's FinFET/Tri-Gate transistor, high-K dielectrics, etc, represent significant achievements maintaining sub-threshold slope close to their historical values, they still remain within the scope of the MOSFET-imposed limits and are one-time improvements rather than scalable changes.

Two VLSI candidates that bypass these limits because they are not based on thermal injection, are Tunnel Field Effect Transistors (TFETS)(e.g., [16]), which are based on tunneling effects, and Nano-Electro-Mechanical switches (e.g, [3, 1]), which are based on physical switches. Both of them hint at orders-of-magnitude improvements in leakage, but remain to be tamed from the wild.

Perhaps one source of our optimism at finding new devices is the efficiency and density of the human brain. The brain integrates 100 trillion synapses that operate at < 100 mv and embody an existence proof of highly parallel, mostly dark operation.

7. CONCLUSION

In this paper, we have examine four possibilities for our dark silicon dominated future. Although silicon is getting darker, for researchers the future is bright and exciting; dark silicon will cause a transformation of the computational stack, and from that transformation will come many opportunities for investigation.

8. ACKNOWLEDGEMENTS

We thank Brucek Khailany (NVidia), Chris Batten (Cornell), Dreslinski (U Mich), Steven Swanson, Jack Sampson and the GreenDroid Team (UC San Diego), Mattan Erez (UT Austin), Dean Tullsen (UC San Diego), Arun Raghavan (U Penn), Milo Martin (U Penn), Doug Burger (Microsoft), and Thomas Wenisch (U Mich) for productive (and challenging) discussions.

9. REFERENCES

[1] Chen et al. "Demonstration of integrated micro-electro-mechanical switch circuits for vlsi applications." In *ISSCC*, Feb. 2010.

[2] D. Fick et al. "Centip3de: A 3930 dmips/w configurable near-threshold 3d stacked system with 64 arm cortex-m3 cores." In *ISSCC*, Feb. 2012.

[3] H. Dadgour, and K. Banerjee. "Design and analysis of hybrid nems-cmos circuits for ultra low-power applications." In *DAC*, june 2007.

[4] R. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc. "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions." In *JSSC*, October 1974.

[5] R. 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*, Feb. 2010.

[6] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger. "Dark silicon and the end of multicore scaling." *SIGARCH Comput. Archit. News*, June 2011.

[7] Goldstein et al. "Piperench: A reconfigurable architecture and compiler." *Computer*, Apr. 2000.

[8] N. Goulding, J. Sampson, G. Venkatesh, S. Garcia, J. Auricchio, J. Babb, M. Taylor, and S. Swanson. "GreenDroid: A mobile application processor for a future of dark silicon." In *HOTCHIPS*, 2010.

[9] N. Goulding-Hotta, J. Sampson, G. Venkatesh, S. Garcia, J. Auricchio, P.-C. Huang, M. Arora, S. Nath, V. Bhatt, J. Babb, S. Swanson, and M. Taylor. "The GreenDroid mobile application processor: An architecture for silicon's dark future." *Micro, IEEE*, March 2011.

[10] N. Goulding-Hotta, J. Sampson, Q. Zheng, V. Bhatt, S. Swanson, and M. Taylor. "Greendroid: An architecture for the dark silicon age." In *ASPDAC*, 2012.

[11] V. Govindaraju, C.-H. Ho, and K. Sankaralingam. "Dynamically specialized datapaths for energy efficient computing." In *HPCA*, 2011.

[12] N. Hardavellas, M. Ferdman, B. Falsafi, and A. Ailamaki. "Toward dark silicon in servers." *IEEE Micro*, 2011.

[13] Hsu, Agarwal, Anders et al. "A 280mv-to-1.2v 256b reconfigurable simd vector permutation engine with 2-dimensional shuffle in 22nm cmos." In *ISSCC*, Feb. 2012.

[14] W. Huang, M. R. Stant, K. Sankaranarayanan, R. J. Ribando, and K. Skadron. "Many-core design from a thermal perspective." In *DAC*, 2008.

[15] W. Huang, K. Rajamani, M. Stan, and K. Skadron. "Scaling with design constraints: Predicting the future of big chips." *IEEE Micro*, july-aug. 2011.

[16] A. Ionescu, and H. Riel. "Tunnel field-effect transistors as energy-efficient electronic switches." In *Nature*, November 2011.

[17] Jain, Khare, Yada et al. "A 280mv-to-1.2v wide-operating-range ia-32 processor in 32nm cmos." In *ISSCC*, Feb. 2012.

[18] E. Krimer, R. Pawlowski, M. Erez, and P. Chiang. "Synctium: a near-threshold stream processor for energy-constrained parallel applications." *IEEE Computer Architecture Letters*, Jan. 2010.

[19] R. Merrit. "ARM CTO: power surge could create 'dark silicon'." *EE Times*, October 2009.

[20] C. Nosko. "Competition and quality choice in the cpu market." 2010.

[21] Raghavan et al. "Computational sprinting." In *HPCA*, Feb. 2012.

[22] E. Rotem. "Power management architecture of the 2nd generation intel core microarchitecture, formerly codenamed sandy bridge." In *Proceedings of Hotchips*, 2011.

[23] J. Sampson, G. Venkatesh, N. Goulding-Hotta, S. Garcia, S. Swanson, and M. B. Taylor. "Efficient complex operators for irregular codes." In *HPCA*, 2011.

[24] Seo, Dreslinski, Woh et al. "Process variation in near-threshold wide simd architecture." In *DAC*, June 2012.

[25] S. Swanson, and M. Taylor. "GreenDroid: Exploring the next evolution for smartphone application processors." In *IEEE Communications Magazine*, March 2011.

[26] Venkatesh, Sampson, Goulding, Garcia, Bryksin, Lugo-Martinez, S. Swanson, and M. B. Taylor. "Conservation cores: Reducing the energy of mature computations." In *ASPLOS*, 2010.

[27] G. Venkatesh, J. Sampson, N. Goulding, S. K. Venkata, M. B. Taylor, and S. Swanson. "QsCores: trading dark silicon for scalable energy efficiency with quasi-specific cores." In *MICRO*, 2011.

[28] Waingold et al. "Baring it all to software: Raw machines." In *IEEE Computer*, September 1997.