Report - A Landscape of the New Dark Silicon Design Regime

1. **Provide a short (100-word) summary of the paper.**

The article addresses the effects of CMOS scaling in the modern era, especially the pareto optimality between performance and energy efficiency goals in successive generations. The timeline spans Dennard scaling with S^2 energy efficiency from voltage scaling, and post-Dennard scaling with constant voltage limited by leakage. It then dives into the 4 directions of adapting to the dark silicon: the pessimistic shrinking of area, underclocking/less-duty cycle circuits, dedicating parts of the chip to specialized accelerators and exploring novel devices to replace some/all the MOSFETs currently in use. It concludes with drawing some fascinating insights about brain, a very dark computer which performs massively parallel analog operations at a very low voltage.

2. **What did you learn from the paper?**
   - NTV Circuits: while it comes with a lot of variability due to PVT intolerance, it can amortize some massively parallel operations (like current day's data-driven vectorized loads) if the required accuracy can be met.
   - DVFS: Intel's Turbo Boost operated at various frequencies based on the number of cores being run, which in turn depends on the load. This Dynamic Voltage and Frequency Scaling technique increases power proportional to the cube of the increase in frequency.
   - Exceeding the thermal budget: Intel's Turbo Boost 2.0 and Arm's big.LITTLE employ several high-energy, high-performance cores thereby going over the thermal budget for short durations (in bursts) and reverting back to low-energy, low-performance cores especially to execute the "race-to-finish" class of computations. The effect of this is mitigated by thermal engineering including having some phase-change materials that can help maintain the chip's temperature.
   - Specialized hardware comes with its own cost: the "tower of babel" where configuring them along with the general-purpose computing units becomes increasingly complicated.
   - GreenDroid's C cores: an attempt to overcome Amdahl's limits on specialization, this improves the inherently parallel utilization without impacting the performance of serial workloads. This is a great general-purpose like parallelizing PE, as NTVs might still hold area-advantage for highly parallel loads.
   - Brain performs fine analog operation with it synapses handling large fan-outs & fan-ins of ~7000 nodes per neuron, with each node being spatially quite far apart.

3. **What did you like about the paper?**
   - The paper concisely briefs about the directions of research work to deal with the post-Dennard era of dark silicon.
   - The insights about the brain makes me think about some of the fundamental building blocks that make up the systems today. The transistors of choice being MOSFETs has its own inherent limitations. While it is nearly impossible to move the entire hardware stack to circuits/architecture with the emerging novel devices to handle such large fan-ins & fanouts etc, it is important to draw as much inspiration as we could which can lead to a greater design space exploration for one of the horsemen – the specialized accelerators, at the very least.
4. **What did you dislike about the paper?**

There isn't anything that I majorly dislike about the paper itself. However, some of the things that would've aided the insights gained from the paper:

- While the solutions within any particular "horseman" had its own metrics for comparison and coming up with an effective FoM, it is probably essential to come up with a fundamental metric/framework to compare all the solutions like the Amdahl's law does for the general-purpose architecture.
- Also, a classifying the caches as "Low-duty cycle" and hence dark blocks might be application specific and depends on the level of cache itself. Ofcourse, the higher the level (L3 & beyond, if any), possibly lower its duty cycle. But arbitrarily adding levels to the hierarchy might give incrementally diminishing fruits wrt cache-hits etc. A better context to this claim might improve the clarity. Certain cache/scratchpad memory like in accelerators may be continuously replaced resulting in tight coupling with the processors itself.

5. **The paper talks about "Shrinking" as a horseman of the apocalypse. Do you agree? Why or why not?**

Yes, I share similar views as of the author regarding "Shrinking". The dark silicon regime offers plenty of opportunities for architectural innovation. However, making a smaller chip not only takes away that avenue but also makes it thermally less stable. The Power/Area increases rapidly across generations and cooling costs involved itself might overshoot the chip area budget that as saved. Rising temperatures of the chip also increases variability and affects the life-span of the circuits involved. Also, as described in the paper, the economics of making cheaper chip in terms of area might still end up inferior to a competitor offering noticeably better performance for a little extra cost.