

Apparate : Evading Memory Hierarchy with GodSpeed Wireless-on-Chip


Nitesh Narayana GS
Universitat Politècnica de Catalunya
Barcelona, Spain
nitesh@ac.upc.edu

Abhijit Das
Universitat Politècnica de Catalunya
Barcelona, Spain
abhijit.das@upc.edu

1 Introduction

We have come a long way since the first digital computer ENIAC was developed in the early 1940s [1]. For example, our mobile phones have 100,000 times more computing power than the Apollo Guidance Computer (AGC) [3] that landed the first humans on the Moon [6]. Over the past few decades, we have witnessed some remarkable technological and architectural advancements, from the inception of transistors to their continuous shrinking to date, from single-core to massive many-core processors, and from large monolithic chips to manageable chiplets. What is very intriguing is how these advancements often align with certain time-tested predictions, like Moore’s Law [18], Dennard Scaling [13], etc.

A current trend shows that the technological and architectural advancements have shifted the fundamental bottleneck of a system design from computation to communication [9, 10, 20]. Chip and package-scale communication will soon start dictating the designs of next-generation computing systems. While everyone has put on their thinking caps to envision how future computing systems will look like, we present a wild and crazy yet calculated guess. We believe that by extrapolating the data from the time-tested predictions and current trends, we could predict or, even better, suggest how computing systems could be designed in 2050¹.

We present **Apparate** ², a prediction-cum-concept to use wireless communication within the chip to evade memory hierarchy for superior performance and efficiency of computing systems. The subsequent sections will describe the motivation, feasibility, and implementation of Apparate.

2 Trends Till Now to Trends Here After

The projection of trends is restricted till 2050 to provide a comprehensive view of the extrapolations. This timeframe strikes a balance, offering a glimpse into the foreseeable future while maintaining anticipation and excitement.

Transistor Scaling: While most computer architects believe Moore’s Law is either nearing or already dead, Intel [4] disagrees. They believe that Moore’s Law will continue to live on with the support of advanced packaging technology and new materials [11]. We borrow their optimism and show

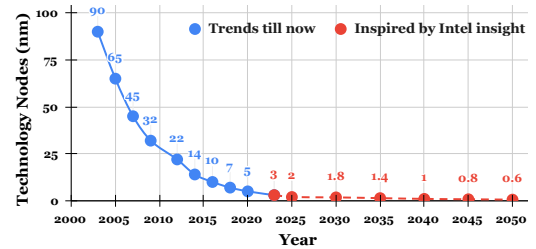


Figure 1. Yet another Moore’s Law graph towards 2050.

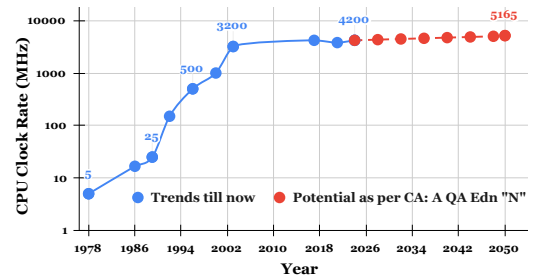


Figure 2. Processor clock rate towards 2050.

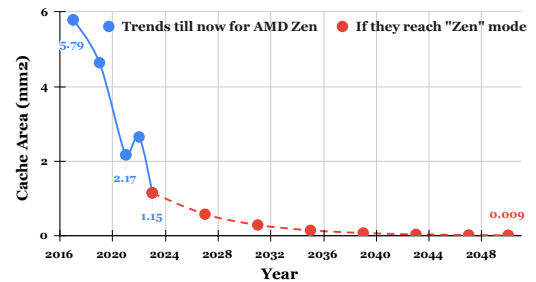


Figure 3. On-chip cache area towards 2050.

in Figure 1 that Moore’s Law might continue but at a slower pace as it will start to decelerate in 2025. Nevertheless, the extrapolated trajectory suggests that the number of transistors on a chip area will continue to increase. *This trend is encouraging and will continue to allow computer architects to explore unconventional and ground-breaking chip design.*

Processor Clock: In Figure 2, we have extrapolated Figure 1.11 of the book CA: A QA Edn “6” [17], showing the growth in clock rate of microprocessors. We observe that after an exponential increase until the last decade, the processor clock rate has now become more or less stagnant. This is mainly due to the breakdown of Dennard Scaling. *This trend implies that there isn’t going to be much of a change in the required L1 cache bandwidth, which is currently at 1 Tbps.*

¹By 2050, the transistor [2] would complete its 100th anniversary.

²Apparate or Apparition [19] is an act of magical transportation from one place to another without any physical means in the Wizarding World [8].

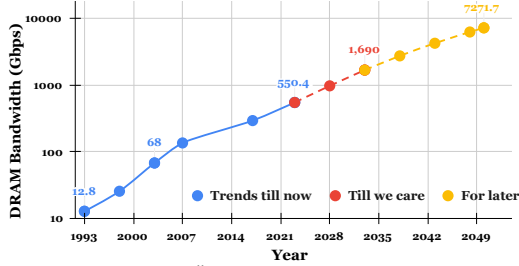


Figure 4. DDR “N” bandwidth towards 2050.

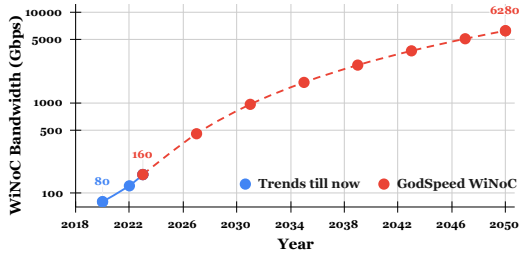


Figure 5. GodSpeed WiNoC bandwidth towards 2050.

On-Chip Cache Area: To identify the trend in on-chip cache area footprint over the years, we examined the AMD Zen processor [5] series. Extrapolating their L1+L2+L3 cache areas in Figure 3, we observe a continued decrease in the footprint. While the L1 cache size remained steady to keep up with the processor speed, L2 and L3 kept increasing. Nevertheless, their overall footprint continued decreasing due to the exponential growth in transistor count within the same chip area. This trend raises an intriguing question: With no area overhead in increasing cache size and a saturated L1 cache bandwidth, can DRAMs replace caches in the future?

DRAM Bandwidth: Our extrapolation in Figure 4 shows that DRAM could achieve 1 Tbps bandwidth well before 2050, with a projection of up to 7 Tbps by that time. A CPU directly communicating with DRAM would ideally remove all the memory hierarchy-related bottlenecks. However, despite this promising outlook, putting DRAMs alongside CPUs poses inherent challenges due to its capacitor-based storage mechanism. This roadblock raises a serious curiosity: Could there be a way to make the CPU directly talk to the DRAM?

Wireless Network-on-Chip: There has been a growing interest in Wireless Network-on-Chip (WiNoC) [12, 14–16]. To shed light on this emerging trend, we have charted the extrapolated trajectories of WiNoC’s bandwidth and area in Figures 5 and 6, respectively. On one side, we observe that WiNoC could attain L1 cache-equivalent bandwidth of 1 Tbps as swiftly as DRAMs. On the other side, we also observe that the WiNoC Transmitter (Tx) and Receiver (Rx) could rival the cache area footprint by 2050. This prompts us to contemplate the unthinkable: Could WiNoC replace caches and be the bridge between the CPU and DRAM?

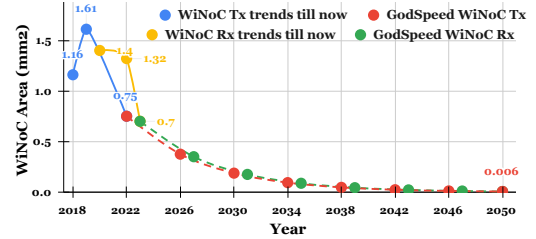


Figure 6. WiNoC Tx and Rx areas towards 2050.

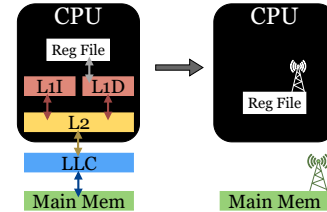


Figure 7. The Apparate concept.

3 Apparate: A Way Forward

We propose a radical departure from traditional CPU architecture and advocate for the evasion of the “latency-hiding” caches from the memory hierarchy. As illustrated in Figure 7, our concept, **Apparate** ✂, replaces caches with wireless transceivers at the register file and the DRAM. When the CPU initiates a memory request, it “*apparates*” from the register file to the DRAM through the transceiver. The future high-bandwidth DDR “N” DRAM will immediately respond with the data, which will “*apparate*” back to the register file. This arrangement will obviate the need for memory hierarchy³, thereby removing its associated bottlenecks.

In terms of feasibility, Apparate prompts several intriguing questions. Below, we offer a teaser of potential questions and answers, leaving the wild exploration to the readers!

- *What happens to “cache” coherency?* As caches will be phased out, the onus for maintaining coherency will fall onto the main memory. Consequently, coherence tables will find their new home in the main memory.
- *How hot can the chip now get?* Regardless, it should be more manageable with no caches to heat things up!
- *Will we need prefetching, replacement, etc.?* We may no longer need conventional complexities like cache prefetching, replacement, etc. This will liberate valuable chip space, allowing for new innovations!

4 Conclusion

We hypothesise that computer architecture is poised to transition into an era dominated by WiNoC technology. Therefore, investing thought into Apparate will not only shape the future design of computing systems but also provoke a fundamental question: Do we truly need what we already have? As demonstrated through Apparate, it becomes evident that traditional caches⁴ may no longer be essential in the future.

³A long-standing dream of most computer architects!

⁴And hence as a butterfly effect other structures related to caches.

Acknowledgments

We extend our sincere gratitude to the WACI initiative of ASPLOS for giving us a forum that recognises thoughts and ideas that are indeed wild and crazy! Special thanks to our friend Hrishikesh R Menon [7] for his creative support.

References

- [1] 1945 (accessed April 18, 2024). ENIAC. <https://en.wikipedia.org/wiki/ENIAC>.
- [2] 1947 (accessed April 18, 2024). Transistor. <https://en.wikipedia.org/wiki/Transistor>.
- [3] 1966 (accessed April 18, 2024). Apollo Guidance Computer. https://en.wikipedia.org/wiki/Apollo_Guidance_Computer.
- [4] 1968 (accessed April 18, 2024). Intel. <https://en.wikipedia.org/wiki/Intel>.
- [5] 2017 (accessed April 18, 2024). AMD "Zen" Core Architecture. [https://en.wikipedia.org/wiki/Zen_\(microarchitecture\)](https://en.wikipedia.org/wiki/Zen_(microarchitecture)).
- [6] 2019 (accessed April 18, 2024). Apollo 11 anniversary: Could an iPhone fly me to the moon? <https://www.independent.co.uk/news/science/apollo-11-moon-landing-mobile-phones-smartphone-iphone-a8988351.html>.
- [7] 2019 (accessed April 18, 2024). Hrishikesh R Menon. <https://www.linkedin.com/in/hrimenon/>.
- [8] 2019 (accessed April 18, 2024). Wizarding World: The Official Home of Harry Potter. <https://www.wizardingworld.com/>.
- [9] 2022 (accessed April 18, 2024). Single-Chip Processors Have Reached Their Limits. <https://spectrum.ieee.org/single-chip-processors-have-reached-their-limits>.
- [10] 2023 (accessed April 18, 2024). Interconnect is the Root of Generative AI. <https://embeddedcomputing.com/technology/ai-machine-learning/interconnect-is-the-root-of-generative-ai>.
- [11] 2023 (accessed April 18, 2024). Moore's Law. <https://www.intel.com/content/www/us/en/newsroom/resources/moores-law.html>.
- [12] Sergi Abadal, Albert Cabellos-Aparicio, Eduard Alarcon, and Josep Torrellas. 2016. WiSync: An Architecture for Fast Synchronization through On-Chip Wireless Communication. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems*. 3–17.
- [13] Robert H Dennard, Fritz H Gaensslen, Hwa-Nien Yu, V Leo Rideout, Ernest Bassous, and Andre R LeBlanc. 1974. Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions. *IEEE Journal of Solid-State Circuits* 9, 5 (1974), 256–268.
- [14] Vimuth Fernando, Antonio Franques, Sergi Abadal, Sasa Misailovic, and Josep Torrellas. 2019. Replica: A Wireless Manycore for Communication-Intensive and Approximate Data. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems*. 849–863.
- [15] Antonio Franques, Apostolos Kokolis, Sergi Abadal, Vimuth Fernando, Sasa Misailovic, and Josep Torrellas. 2021. WiDir: A Wireless-enabled Directory Cache Coherence Protocol. In *Proceedings of the International Symposium on High-Performance Computer Architecture*. 304–317.
- [16] Amlan Ganguly, Kevin Chang, Sujay Deb, Partha Pratim Pande, Benjamin Belzer, and Christof Teuscher. 2010. Scalable Hybrid Wireless Network-on-Chip Architectures for Multicore Systems. *IEEE Trans. Comput.* 60, 10 (2010), 1485–1502.
- [17] John L Hennessy and David A Patterson. 2017. *Computer Architecture: A Quantitative Approach* (6th ed.). Elsevier.
- [18] Gordon E Moore. 1965. Cramming More Components onto Integrated Circuits. *Electronics* 38, 8 (1965), 1–4.
- [19] Joanne K Rowling. 2000. *Harry Potter and the Chamber of Secrets*. Bloomsbury Publishing.
- [20] Sayeef Salahuddin, Kai Ni, and Suman Datta. 2018. The Era of Hyper-Scaling in Electronics. *Nature electronics* 1, 8 (2018), 442–450.