System level synergies for modern memory systems

Krishna T. Malladi

***Abstract*— This brief explores some of the recent advances in memory systems, while examining the scope for future optimizations.**

I. Current Memory Systems

First, let us explore some of the present challenges and trends of memory systems. As applications become data hungry, memory systems in the future could require a combination of scaling techniques from both technology and architectural layers. On the technology front, scaling will continue to ensure memory devices provide large capacity. However, as we get to small transistor feature sizes, reliability of memory bits becomes an important factor []. Error correction at the architectural layer could be one direction to complement scaling and ensure bit reliability.

On the performance front, memory latency and bandwidth have been the traditional challenges and often referred to as "memory wall". Much of the work in computer architecture has focused on caching, efficient prefetching in order to hide the latency effects of frequent memory accesses. As modern caches begin to saturate and workloads evolve to become large and random, the focus of system performance has once again shifted to memory system optimizations []. These approaches include changing the core micro architecture and organization in the DRAM die while attempting to limit the area overhead []. Even though many recent approaches seem rational, it has been difficult to obtain momentum, in part, due to memory industry fabrication overheads. It is possible to overcome these performance challenges in by utilizing intelligence in the memory deice and making the related changes to provide larger speedup.

II. Processing near memory

Processing near memory has been an active area of research for more than 2 decades. However, the programming model changes necessitate a large system level performance advantage to be viable. To realize this, many prior proposals assume the presence of more than $2\times$ internal memory performance. Such architectures provide performance advantages by overcoming the bottlenecks of memory interface. However, the reality is that many wide interface memory devices today expose the large bandwidth to the host. As new workloads like machine learning emerge, it becomes even more crucial to expose even more bandwidth to the host. However, such bandwidth matching also reduces the performance advantages of placing those processing tasks close to memory. As the processing systems close to memory are often thermal and power limited, it becomes challenging to pack useful compute.

Future work need to be cognizant of such practical challenges. However, as future computer architecture advances depend on domain specific architectures, new opportunities emerge to revisit processing near data. One could also consider integrating such processing abilities directly into the DRAM chip to maximize performance benefits. Some examples include [1] and [2]. While the former provides large reconfigurable arrays, the latter performs specialized machine learning operations directly using DRAM arrays.

References

[1] M. Gao et al, "DRAF: A Low-Power DRAM-based Reconfigurable Acceleration Fabric", in Proceedings of the 43rd International Symposium on Computer Architecture (ISCA), June 2016.

[2] S. Li et al, "DRISA: A DRAM-based Reconfigurable In-Situ Accelerator", in Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), October 2017.