Conference/Talk Summary

*Prof. Onor Mutlu gave a captivating broad summary of the current bottleneck (due to data-heavy operations) in computer architecture and explained the change of paragdism that this happening from processor-centric to data-centric designs. The end goal is still the same : saving-up on energy, faster computations. He presented a few of the solutions that he has been working on, focusing on data-centric architectures.*

*He vastly expanded on the memory bottleneck and a few of the introduced solutions have already made their way onto the market or are being worked on and improved at the time being. As the density of DRAM increases, the reliability decreases on the other hand which is why we need to revisit the previous design and processing the data where it makes the most sense, as imagined years ago in A logic in memory IEEE 1970).*

Strength of the talk and mechanisms

* The presenter guided the public troughout the span of the talk and tried not to leave anyone behind
* There was no focus on a single-aspect, which could make us lose track of the other systems layers. This lead to a better understanding of the improvements that can be performed everywhere.
* Real examples (UPMEM, Samsung In-Memory) and extra material (with scientific ressources) for the curious interested attendees.
* Several angles of attacks have been presented to reach the ultimate goal : data-driven, data-aware and data-centric
* Mention of self-optimization of the memory architectures which was not discussed in the presentation

Weaknesses of talk and mechanisms

* It is hard to define time-scales for changes of paragdism, as a question pointed out.
* Lack of precise, detailed examples with specific applications, and more of a whirwind tour over the solutions due to the time constraints, which doesn’t bring many new inputs for the already aware people
* Brief overview of DRAM which to the some people lacking background could be prejudiciable
* No questioning on wether or not the computed data is relevant, which has also been touched in the QA session but the definitive answer if more up to the end user and not if it’s inherently a good thing to discard or scrap out data based on its utility.

Detailed comments

Challenges presented in this 1-hour overview are already are of importance, but will also continue to be so in the near future as well as in the long-term. DRAMs is the most widespread technology for memories but is still facing issues such as Rowhammer (2014) that drive researchers to find mechanisms /new ideas to mitigate currents problems and prevent them in the future. The lesson here is that nothing should ever be taken for granted.

I appreciated that the audience was guided and not left behind, thanks to the explanations on DRAM (voltage/rows/columns), however there was no mention of DIMM, Channels, Banks divisions, and it can be tough for a novice to understand the implications of such issues on Rowhammer. Introducing numbers of impacted products, users, and also adding a perspective on how probable it is that the flaw is leveraged by an attack could improve the slightly the overall informative talk.

Expanded summary

The presenter showed graph-processing examples which are enabled by stacked ICs on top of each other, shortening the paths between the CPU and Data. Other archictectures suchs as GPUs have allowed new-ways of accessing and computing the data, which lead to the push of ML as we know it today.

Examples that were seen: data initialization and bulk copy operations which are usually high latency, high bandwith and cache polluting, can be moved inside DRAMs bypassing the CPU. (details : idea is that we have two consecutive row activations and we make use of the row buffer)

Analog nature of DRAM cells (SIMDRAM, Ambit Architecture) allow any bitwise operations thanks to AND, NOT, OR, thus enabling processing in memory.

Lessons learned

The broad overview of the given talk has given a more interesting approach, and it is sometimes a good refresher to have another look at the pyramid bottom to enable bigger changes that lead to even bigger, fundamental architecture improvements.

Attending regularly such talk is necessary to keep an open-mind on recent advancements in the field, and enables fruitful collaborations between researchers that share their work.