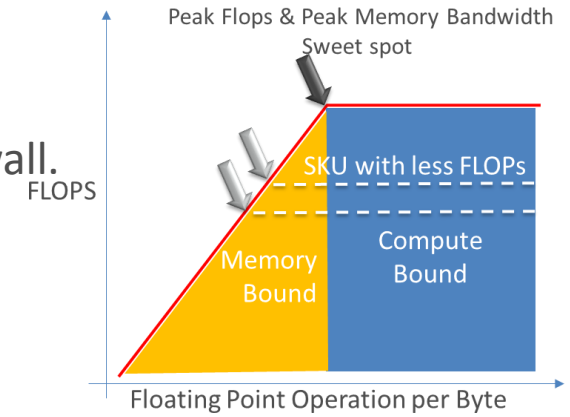
SVA System Vertrieb Alexander GmbH

A VIEW ON MEMORY TECHNOLOGIES B.HOMÖLLE

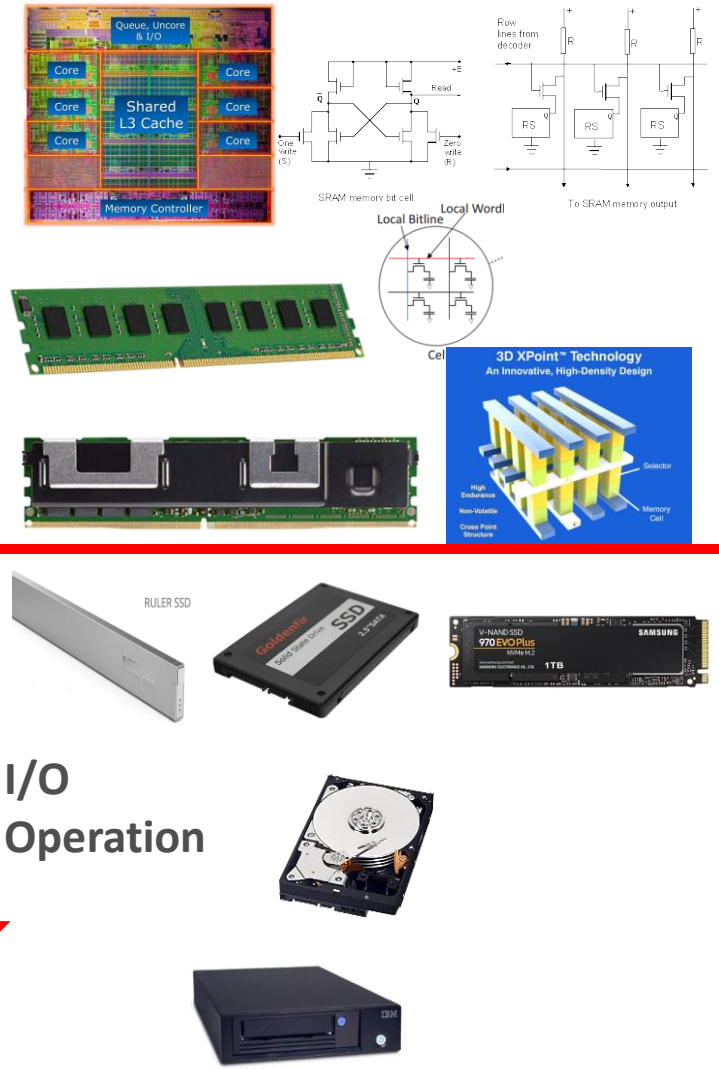
PROBLEM STATEMENT MAIN MEMORY

- Ideal Memory is:
  - Infinite in capacity; All data for complex meshes can be hosted in the main memory.
  - Has throughput like hell; no memory wall ahead slowing down mesh generation and visualization.
  - Has negligible Latency; Runs with CPU core speed to get full performance from the execution unit.
  - Is persistent for decades; can cover memory and storage aspects . No I/O operation to get data from storage.
  - Is low-cost; allows to spend more money for more calculation units.
  - Has low Power; minor amount of energy for data in motion and almost no energy for data at rest.
- Reality Today
  - Main memory is still just the buffer/cache for data to be computed.
  - Arithmetic intensive workloads such as mesh algorithm will easily hit the memory wall.
  - Speed of memory falls behind CPU calculation capabilities (more and more cores).
  - Persistence memory are only at the beginning (or in come back after decades).
  - Memory in $/GB is getting cheaper however per core demand is increasing.
  - Memory power(W) is at the same level of CPUs in a system, burning a lot of energy(Wh).



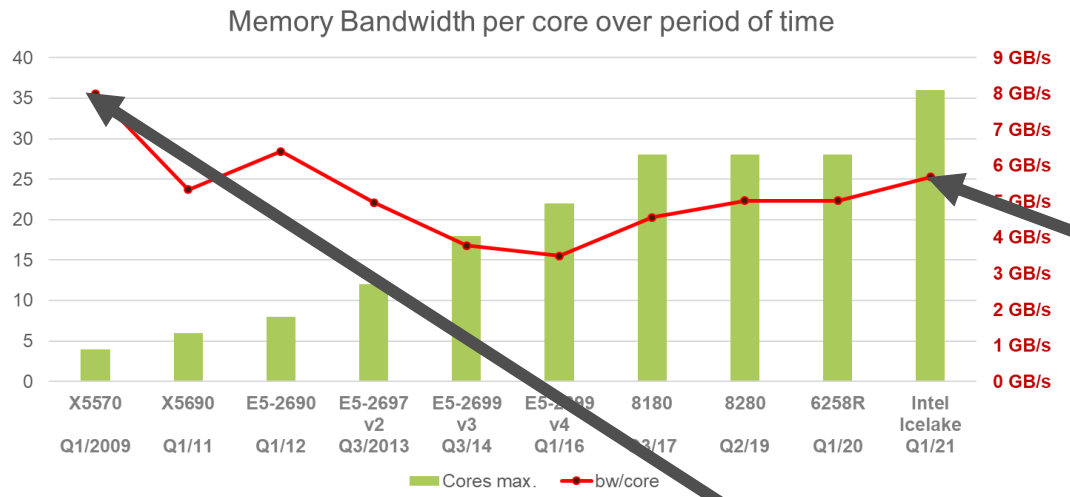
CURRENT SERVER MEMORY/STORAGE ECO SYSTEM

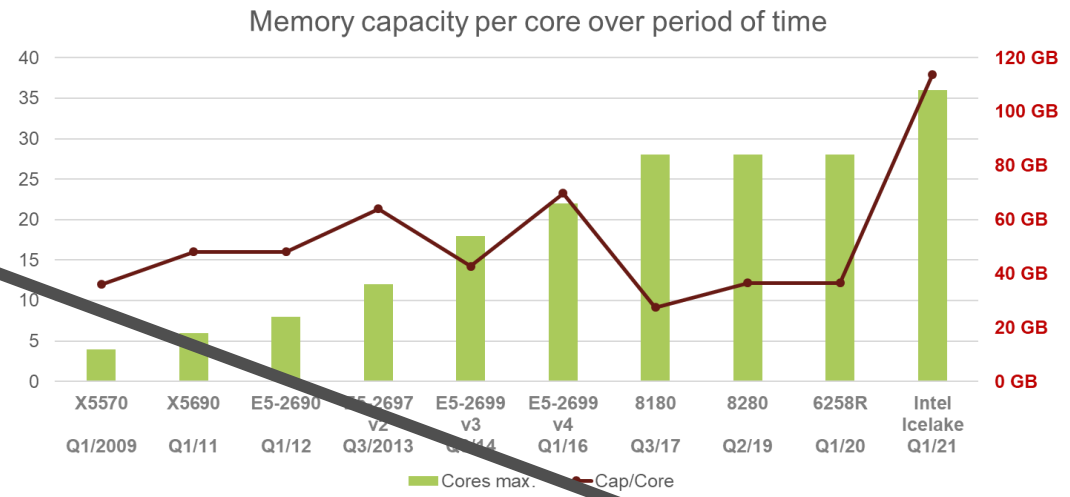
- Rough Sorting
  - SRAM mainly used inside the CPU for Caches. Byte addressable
  - DDR4 DRAM the Main Memory Byte addressable
  - DCPMM the new kid on the block. Byte addressable and persistent
  - NAND (NVMe or SATA) more and more replacement for HDDs as 1st tier storage. Block addressable
  - HDD still best in $/GB 2nd – 3rd tier. Block addressable
  - Tape still the best for long term storage.

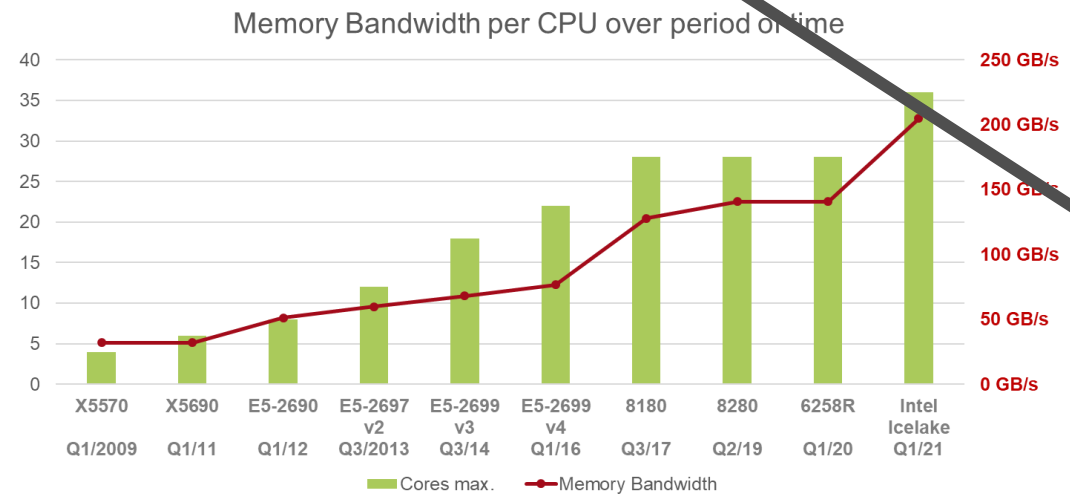


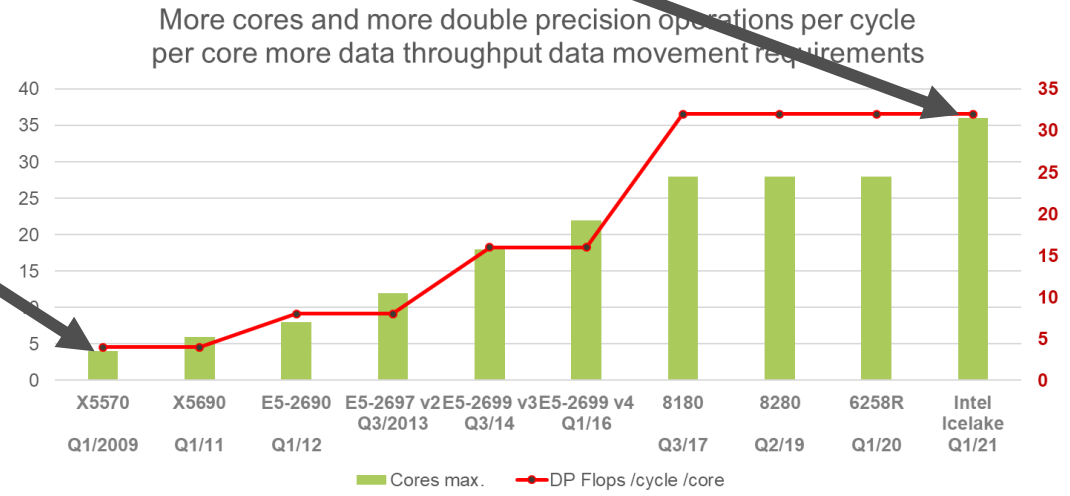
Tech.	Capacity	Latency	Persistence Media	Price $/GB
SRAM	Tiny KB - MB	Sup ns to nsec.	Only during power is on.	Highest
DRAM	Small Few GBs	50 to 100 nsec.	Only during power is on.	High
PCM DCPMM	Medium Few TBs	100-300 nsec.	Yes	Moderate
NAND	Medium TBs	50-300 usec.	Yes	Good
HDD	Medium TBs	Few Milli-seconds	Yes	Better
Tape	High Hundreds of TB	Milli-second to minutes	Yes	Best

MEMORY FIGURES (INTEL CPUS AND DDR ONLY)









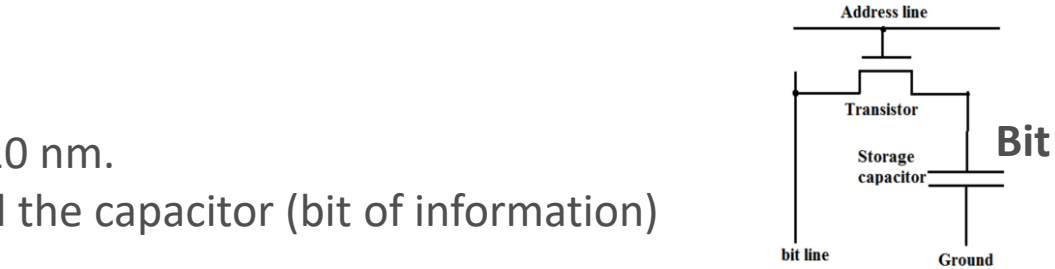
Please Note: the numbers based on datasheet figures. Real system will show lower values!

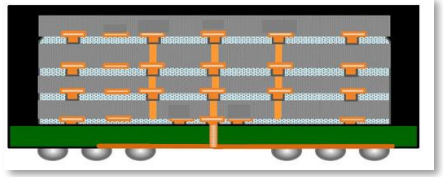
WHAT'S AVAILABLE FOR MAIN MEMORY

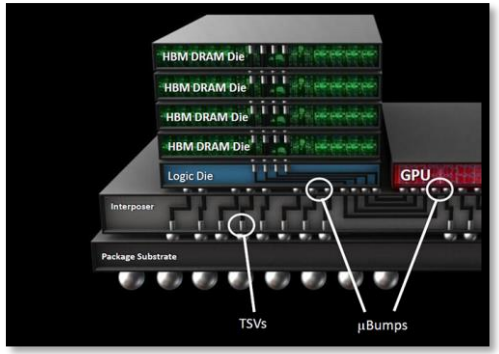
SVA

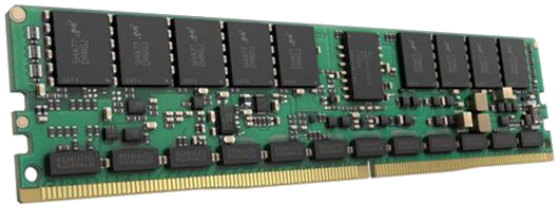
DRAM AND ITS FLAVORS

Expected physical limit for DRAM cells is about 10 nm.
It is non persistent, need refresh cycles to reload the capacitor (bit of information)









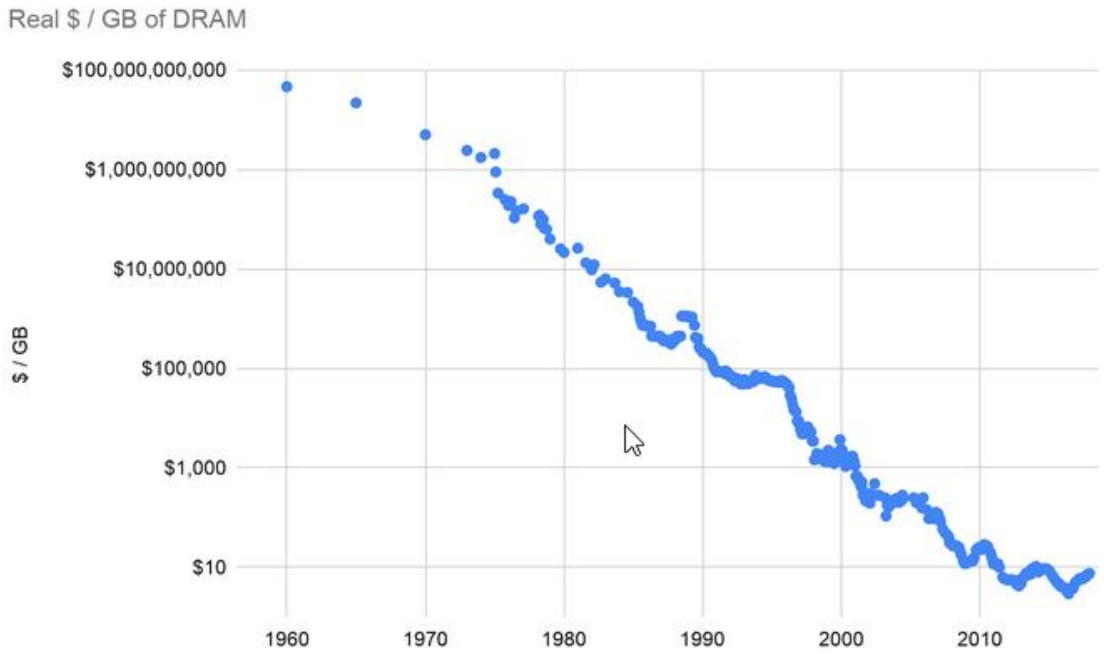
- **DDR4 3DS –**
  Three Dimensional Stacking Through Silicon Via (TSV)
  Max Die 16Gb
  With DDR4 (16Gb x 4) up to **256GB** per DIMM, @3200MHz ~**25.6 GB/s**

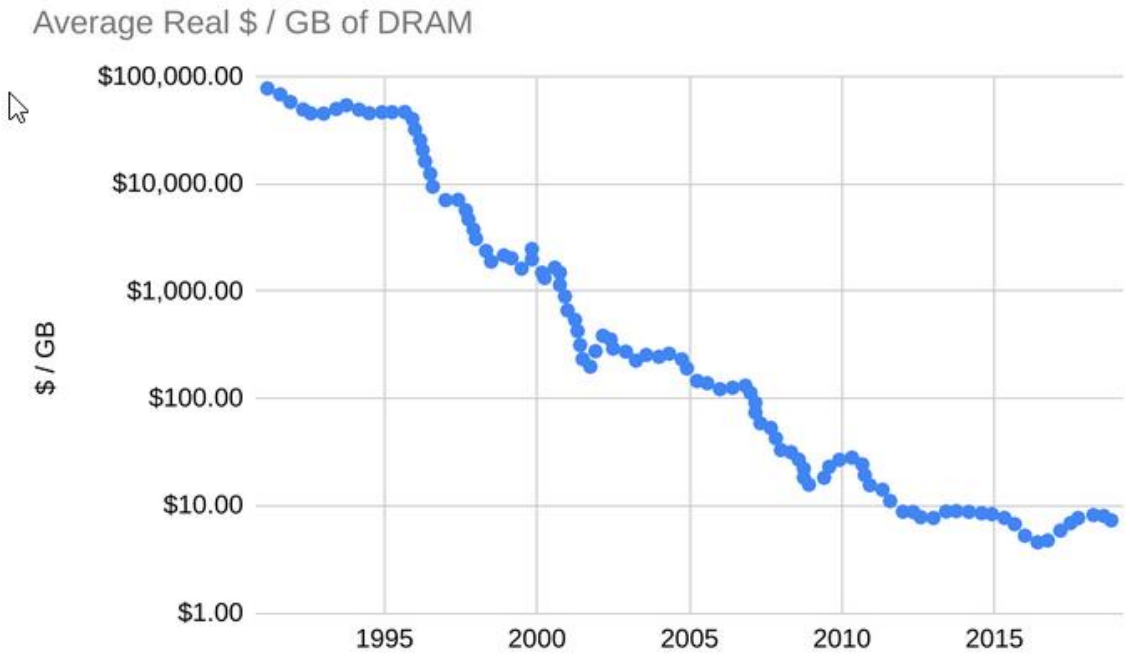
- **High Bandwidth Memory (HBM)**
  A high-bandwidth Interposer connects HBM to CPU or GPU on the same package. Today HBM2 with **4/8GB** capacity and **307GB/s** bandwidth, HBM3 with **8/16GB** capacity and up to ~**512GB/s** bandwidth HBM3+ in 2022 , HBM4 in 2024 and multiple **TB/s** and **≥128GB** per module

- **IBM OMI –DIMMs**
  (Open Memory Interface)
  Interface 8 x 25,6GHz/sec. using separate buffer chip
  Own form factor, plan to make it a JEDEC standard.
  Capacity up to **4TB @320GB/s** or **up 650GB @800GB/s System**

- **DDR5**
  Higher usable throughput ~1.3X compared with DDR4, Lower latency, Max Die 64Gb.
  Support persistent memory protocol.
  Two channels 2x 32+8 Bit (ECC)
  Up to ~**67,2GB/s @8400MHz**
  Next generation DDR6 expected after 2025

$ PER GBYTE

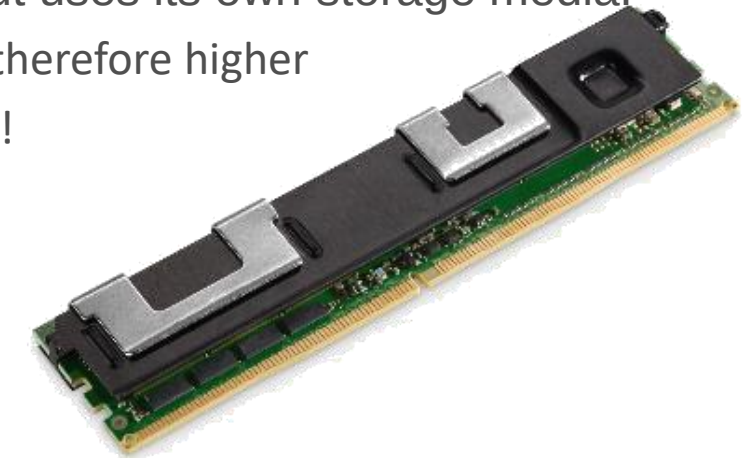




Source .CPI Home." U.S. Bureau of Labor Statistics. U.S. Bureau of Labor Statistics. Accessed April 23, 2020. https://www.bls.gov/cpi/

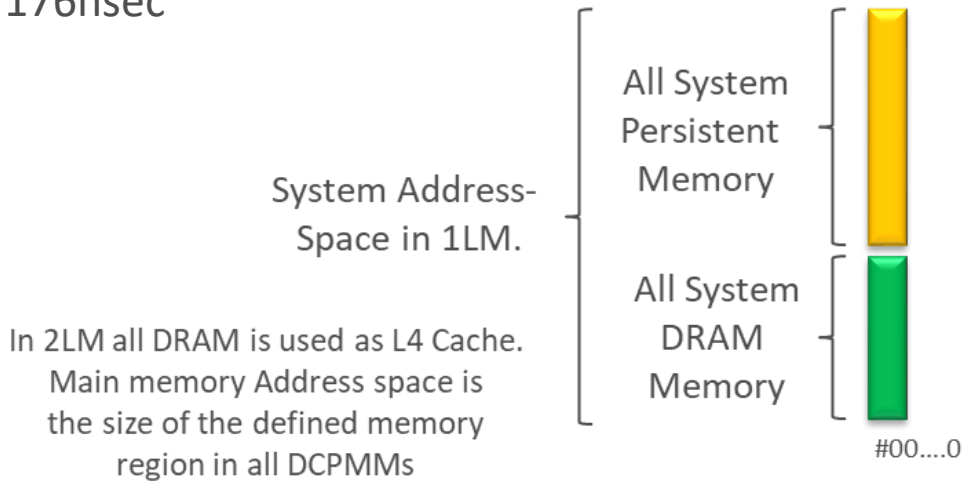
INTEL® OPTANE™ DC (DCPMM)

- DCPMM is not DRAM and not Nand! The media is persistent like flash but uses its own storage media.
  Unlike DRAM, its solid-state cross-point design allows for future shrinkage and therefore higher capacity per DIMM at lower cost. System with DCPMM need always also DRAM!
  Has its own buffer chip which converts internal access cycle to DDR-T

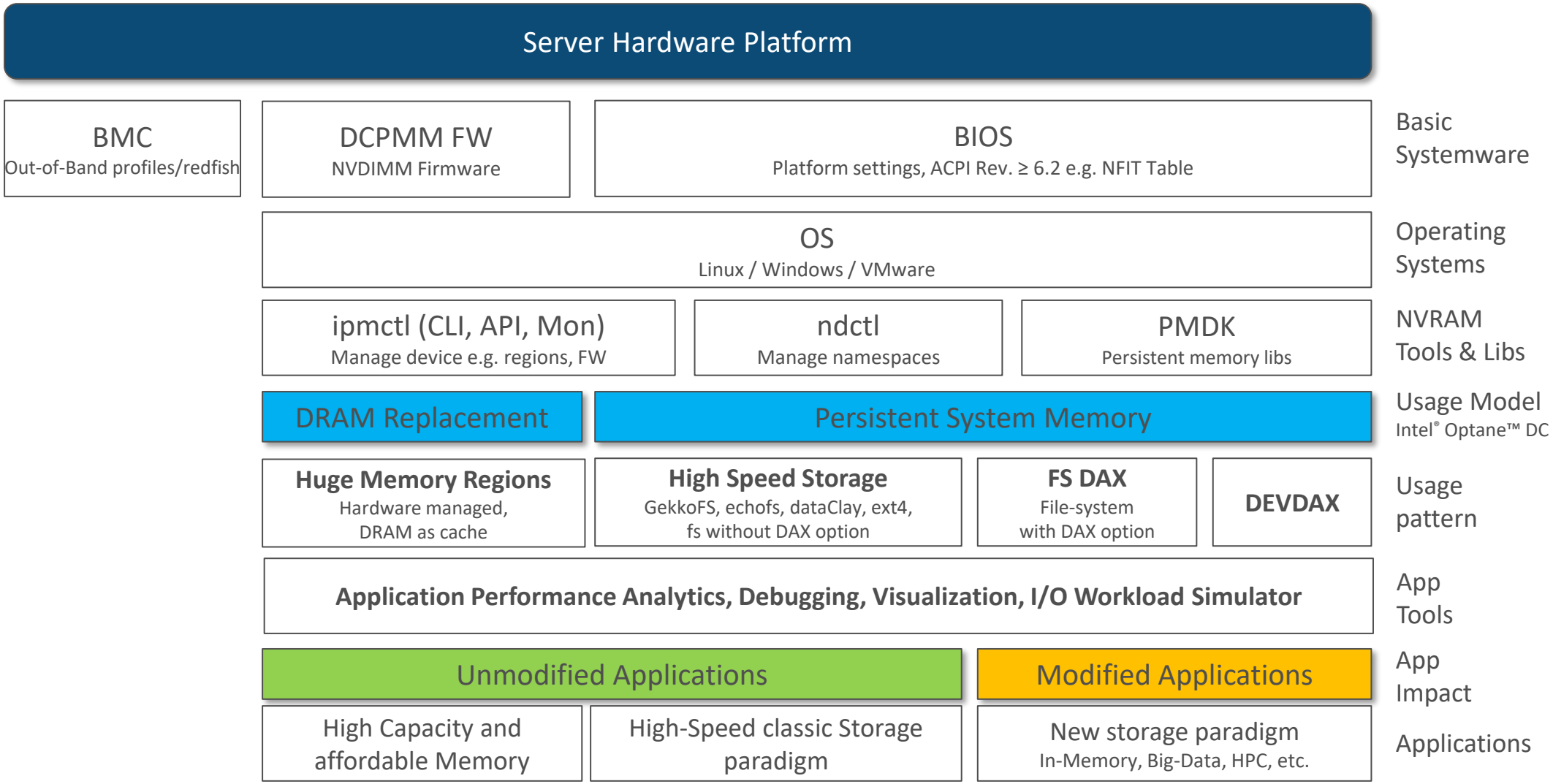


  - 3D-Xpoint developed by Intel and Micron
    Capacity up to **512GB** per DIMM, Speed up to **@2933MHz**
    max. read throughput **~8GB/s** per DIMM and write **~1.5GB/s**.
    (DRAM around 18,7GB/s read and ~8.9GB/s write), Latency around 176nsec
    About **60% of DRAM $/GB** depending on the DIMM capacity.

  - New Versions of DCPMM ahead
    Slightly higher usable throughput for **Icelake @3200MHz**.
    Optional **eADR** which improves app performance by avoiding CPU Cache Flush operations.
    For Sapphire Rapids 2022 DDR5 compliant**, ≥1TB** per DIMM
    Supports more flexible assignment of DRAM as cache for DCPMM



/ PERSISTENT MEMORY SOLUTION STACK



WHAT'S NEXT

SVA

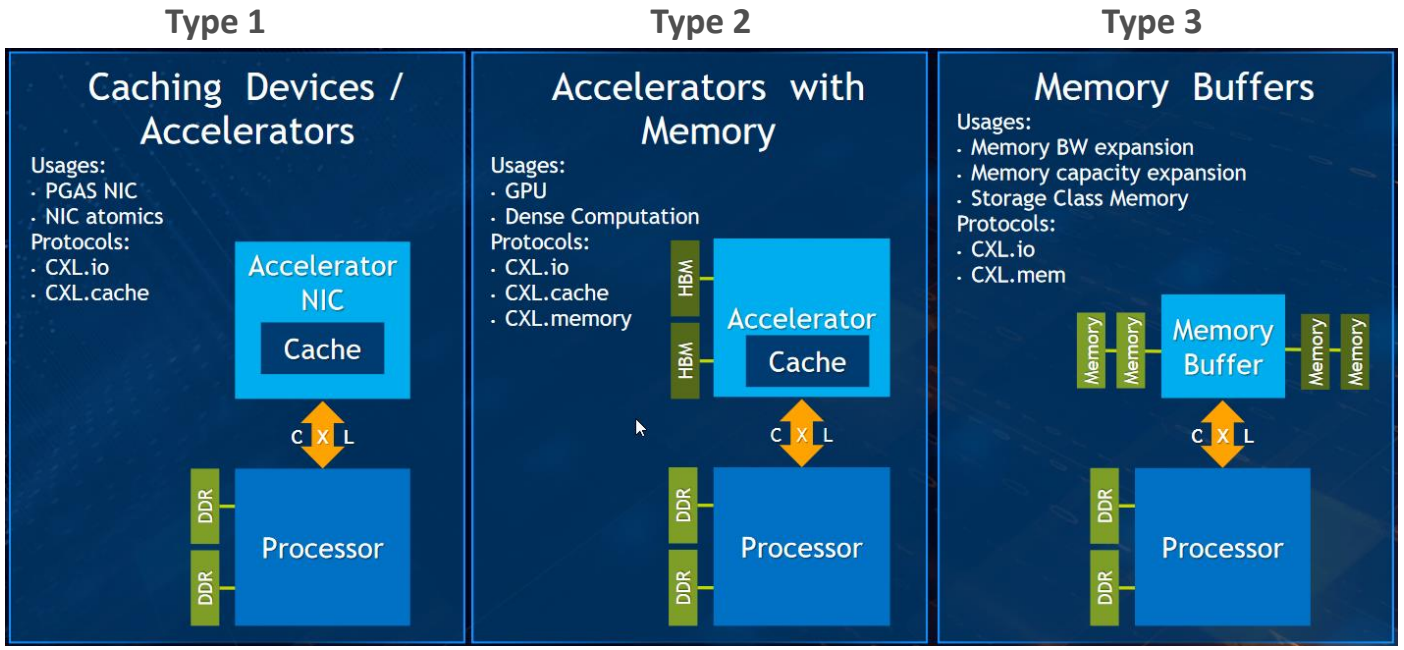
SOME MORE DISRUPTIVE THOUGHTS

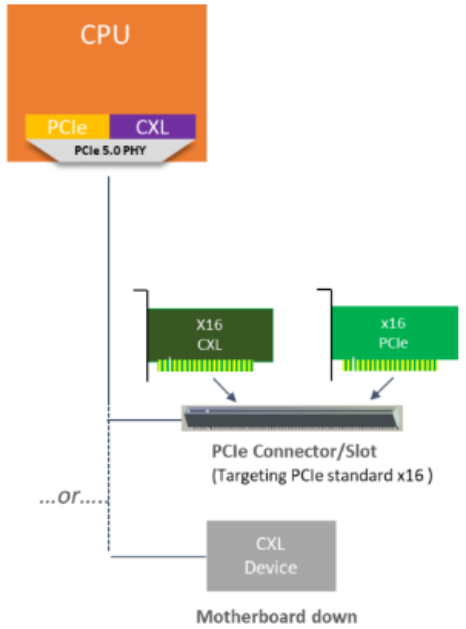
- Is it the right way to bring more and more memory to the core monsters or should we bring some computing power to the memory?
  - Dozens of cores are the de facto standard for CPUs, thousands of cores are standard for GPUs
  - Steady increasing Instructions per cycle (IPC) plus vector and tensor extensions allow TFLOPS of operations.
  - Data movement becomes an even bigger issue. How to solve it?
  - Who is feeding the core beast with data, to get the performance out?

- Are these questions of interconnection or questions of system design or application architecture?
  - Connection of CPU to different memory pools like Gen-Z imagines.
  - Using Cache Coherent local I/O for memory like CXL propagates
- Is the von Neumann architecture the right approach or have we to reconsider how information are computed and stored?
  - Computational memory start adding processing near memory. Where the data rests.
  - New memories such as ReRAM, PCM, neuristors can be used in the non Von Neumann architecture, which at the same time serve as computer and memory

CXL COMPUTE EXPRESS LINK.

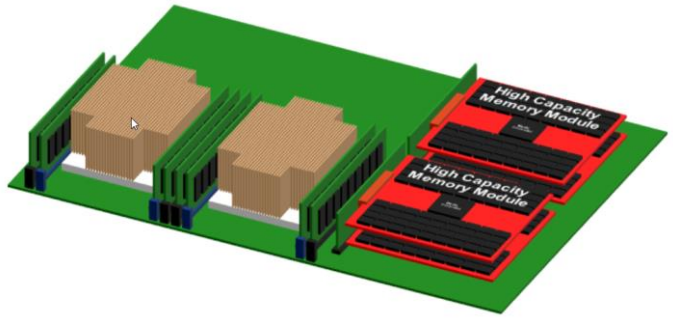
- CXL a new interface starts in 2021/22
  - Based on PCIe 5.0 physical layer using its own protocol supporting cache coherent transfers
  - Up to 64GB/s per direction with x16 PCIe 5.0. That memory bandwidth comes in addition to DDR memory bandwidth.
  - Low Latency .Cache and .Mem targeted at near CPU cache coherent latency (CPU to CPU latency)
  - Allows new and bigger main memory form factors beside DIMM such as Ruler or PCIe AICs or….
    (New FF support higher power envelope than DIMM slots and support higher capacity.)

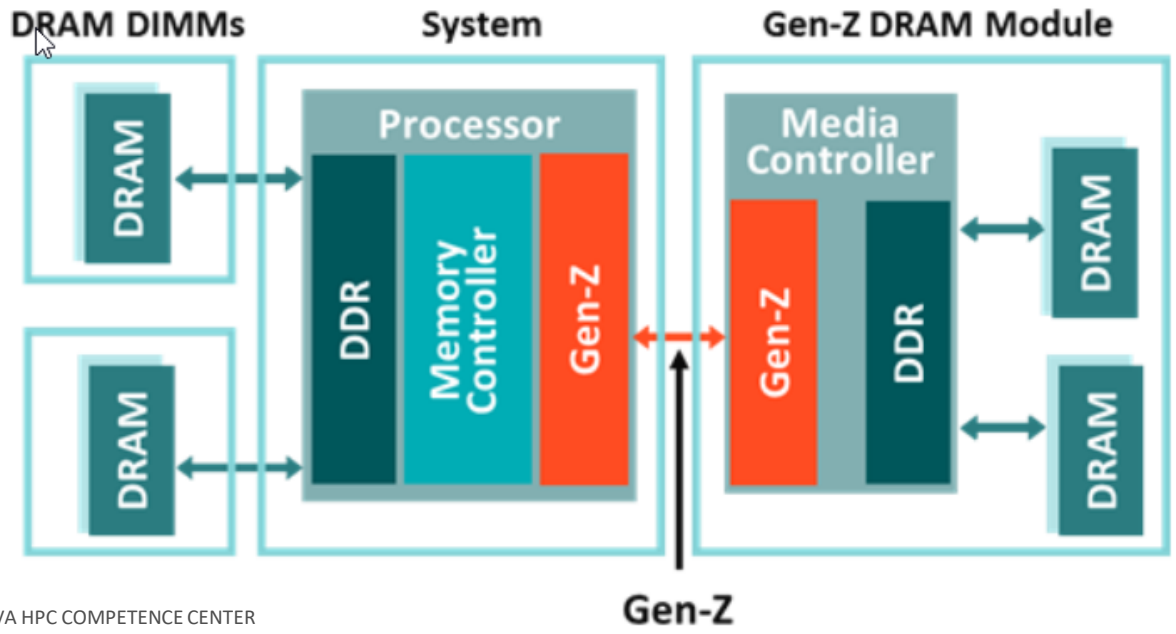


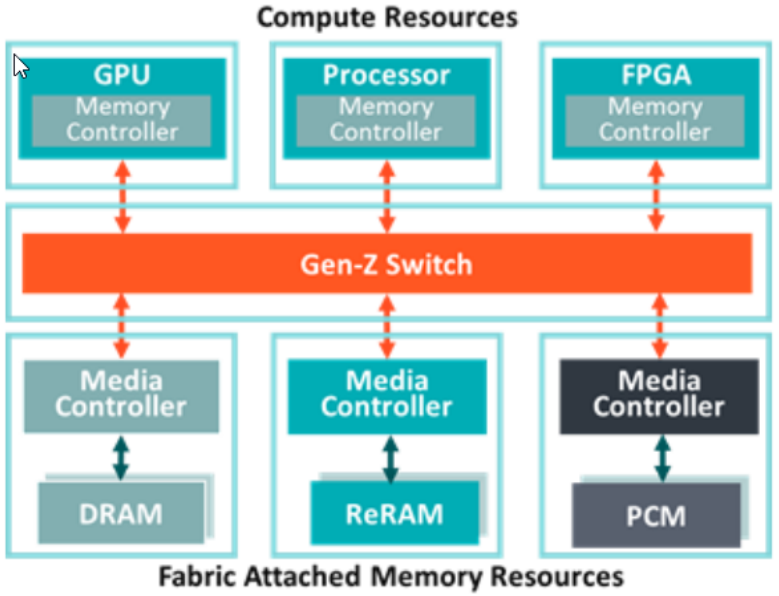


GEN-Z COMPUTE EXPRESS LINK.

- Gen-Z a new system interface
  - Fabric Architecture connecting high-speed memory sematic devices
  - Uses CPU native memory semantics such as load/store operation to talk to remote devices .
  - Supports a disaggregated architecture e.g. flexible assignment of CPU to memory resources.
  - Supports symmetric and asymmetric interfaces e.g. more read than write.
  - Memorandum of Understanding (MOU) recently agreed between CLX and Gen-Z, a path forward for interoperability.







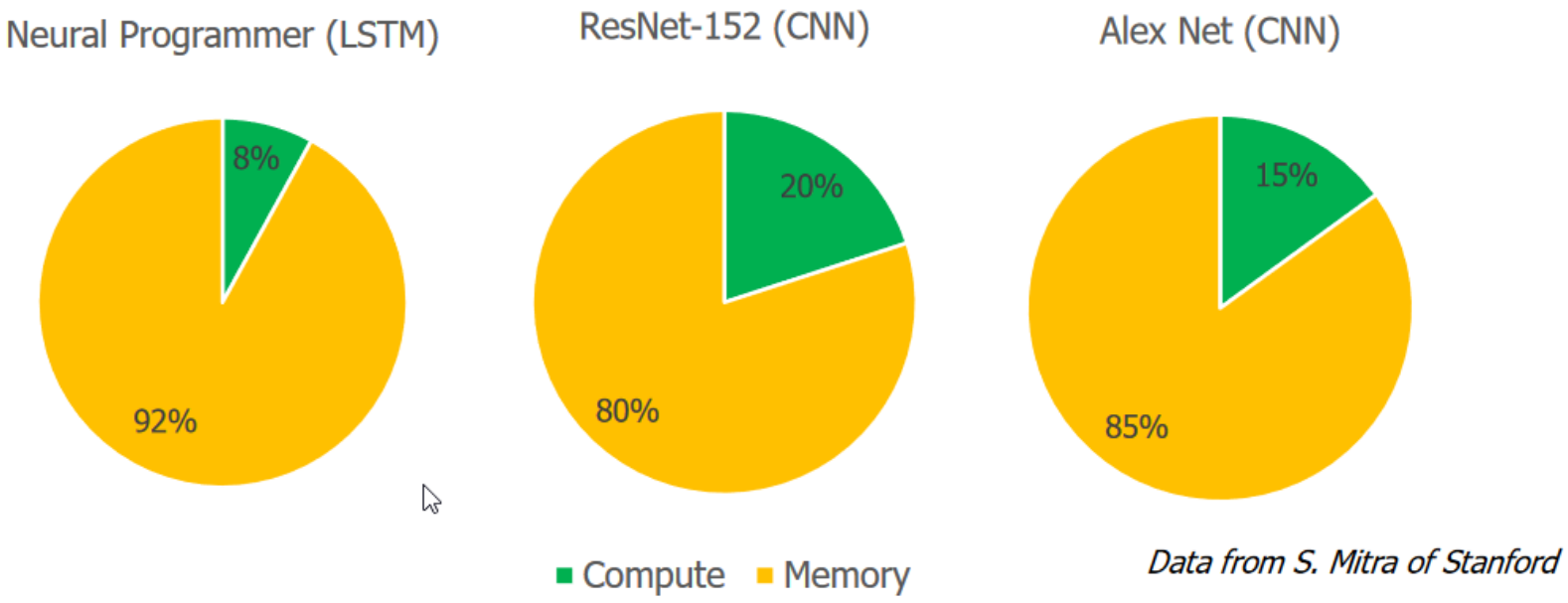
INTENTION OF CXL AND GEN-Z

CXL ON MOTHERBOARD AND GEN-Z ON RACK LEVEL

- Break down the barrier between classic main memory and I/O operations.
  - Expand the coherency domain towards former I/O spheres. Mix Block and Byte access.
  - Bring modern point-to-point interfaces to memory devices. E.g. use PCIe Gen 5 with 64GB/s per direction)
- Enables new non-DIMM form factors to better leverage new technologies in system designs.
  - Space is a critical resource in system designs. The old fashion DIMM FF is not the best answer for cooling and density.
- Add more memory bandwidth or capacity as needed (e.g. mesh calculations and visualization)
  - It allows late decisions on memory bandwidth and capacity to be made, not during development, but during the purchase process or later. Add memory pools like today PCIe cards or connect a memory pool over a network.
- Allow disaggregation of Compute, Memory and Storage.
  - Think more in Racks rather in Motherboards.
  - Optimize your investment by better utilizing the infrastructure (shared usage models).
  - Flexible connect, disconnect, reconnect Compute to memory/storage

The Challenge: Beyond von Neumann Computing

Current von Neumann architecture spends more time moving data than processing it



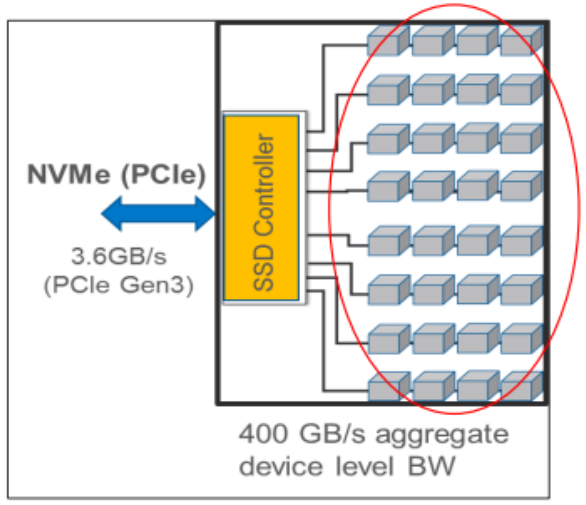
Data from S. Mitra of Stanford

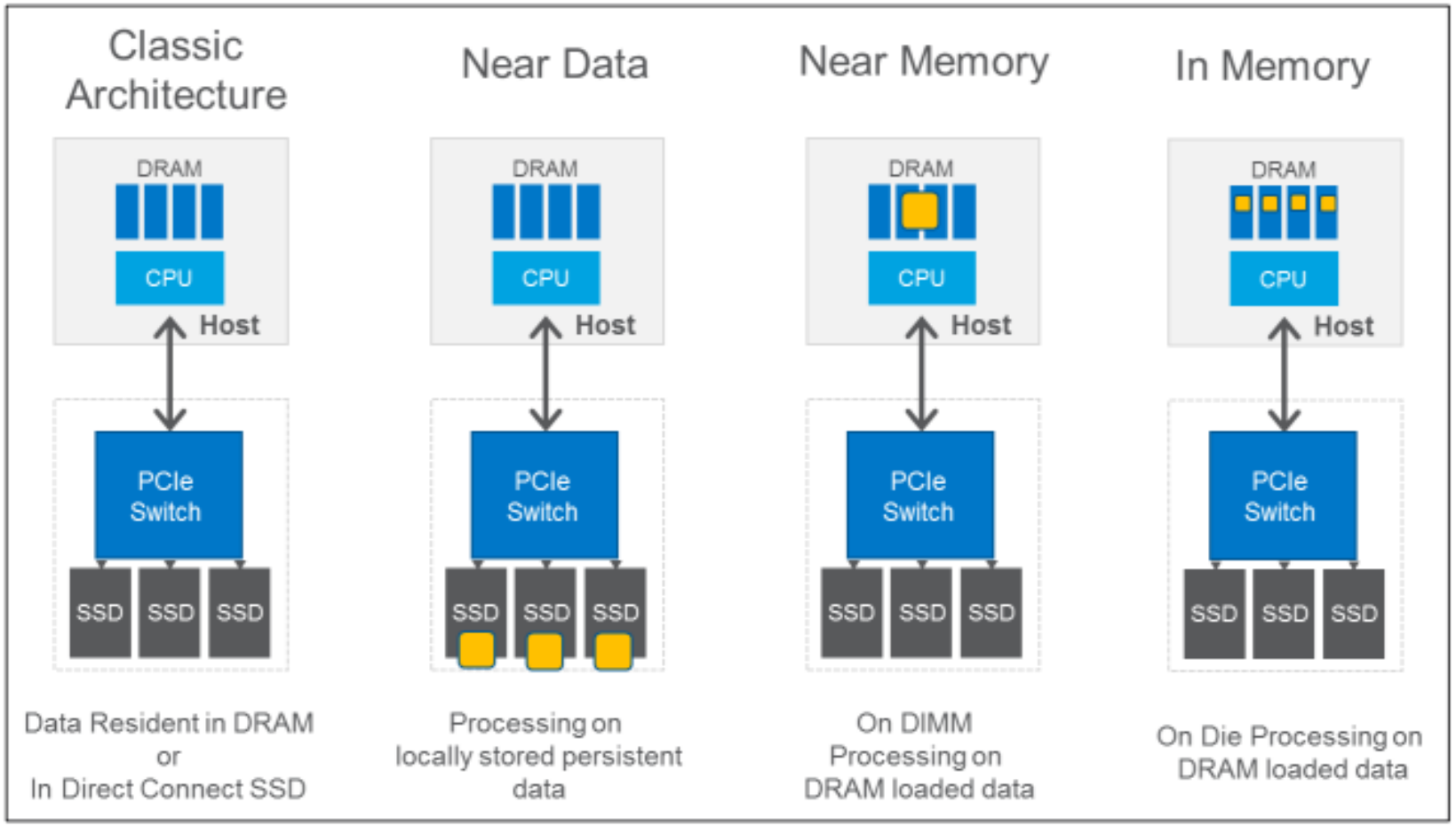
Data for 7nm instantiation of a state-of-the-art Machine Learning accelerator

<u>Accelerators don't help (enough) if using the same architecture</u>

COMPUTATIONAL STORAGE AND MEMORY SOME APPROACHES

- Goal to reduce data movement. Benefit high embedded bandwidth and massive parallelism.
- Enables (pre-)processing on the devices which holds the data.
- Software ECO system critical.



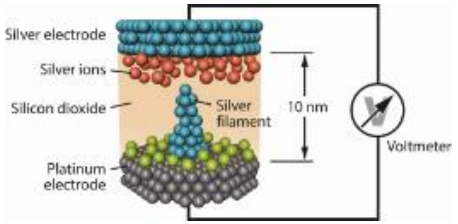


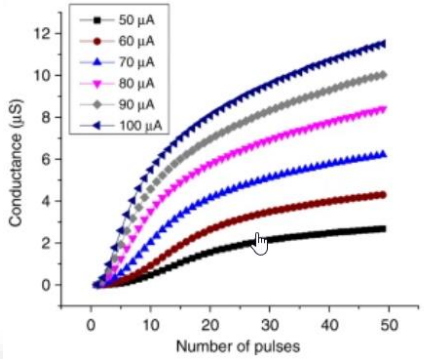
*source Micron

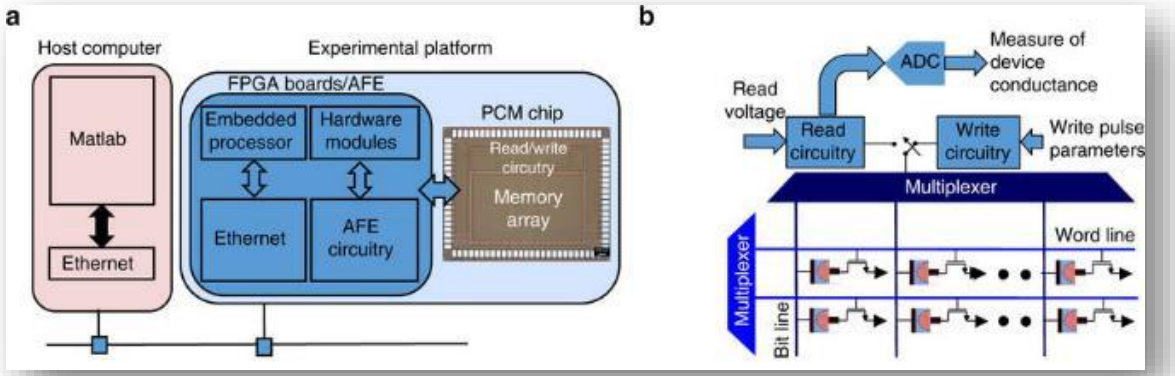
USE THE MEMORY FOR COMPUTING/COMPUTATIONAL MEMORY.

SEE ALSO NATURE ARTICLE

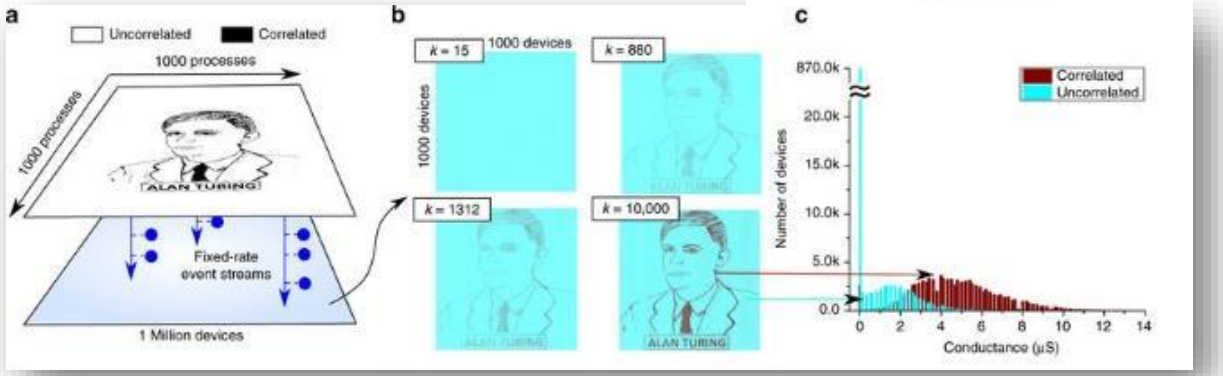
- If a memory cell can store not just “1” and “0” but also all values between it can be used e.g. weight value for machine learning
- If a memory cell changes the stored value (resistance/conductance states) depending on the energy it got in the past it behaves similar to human neuron (often used paths grow, less used paths shrink).
- This behavior can be used to perform certain computational tasks within the memory unit in a non-von Neumann manner.
- Data are no longer moved between CPU and Memory. Instead the Memory itself is doing the computation and storing the result.







Prototype



Example Image Recognition

BACK TO THE PROBLEM STATEMENT

- Reality Tomorrow
  - Memory most likely still the buffer/cache for data to be computed.
    Computational memory can be in some case an option. Software ECO system is critical.
  - New high bandwidth solutions can somewhat reduce the pain.
    DDR Improvement in the bandwidth are just evolutionary but will still be the main memory interface.
    Latest HBM will improve for arithmetic intensive workloads such as Mesh creation and visualization.
    New interfaces such as CXL and Gen-Z will help to enable additional bandwidth sources.
  - Speed of memory falls behind CPU core frequency and the more and more cores.
    The core race hasn't finished yet. No carte blanche for meshing (sorry) but improvements.
  - Persistence memory will more and more make its way into the servers and system concepts.
    Reduce data movement due to convergences of memory and storage. Also cost-effective TB's of memory possible.
  - Memory in $/GB is getting cheaper.
    Persistent memories like PCM can be offered at lower costs due to the simpler manufacturing process.
  - Memory power(W) is coming down for data especially at rest.
    New system designs such as disaggregation can help. Only currently used devices burn energy

Always question where and how to compute ….

Danke