Course Wrap-Up

Jin-Soo Kim
(jinsoo.kim@snu.ac.kr)

Systems Software &
Architecture Lab.

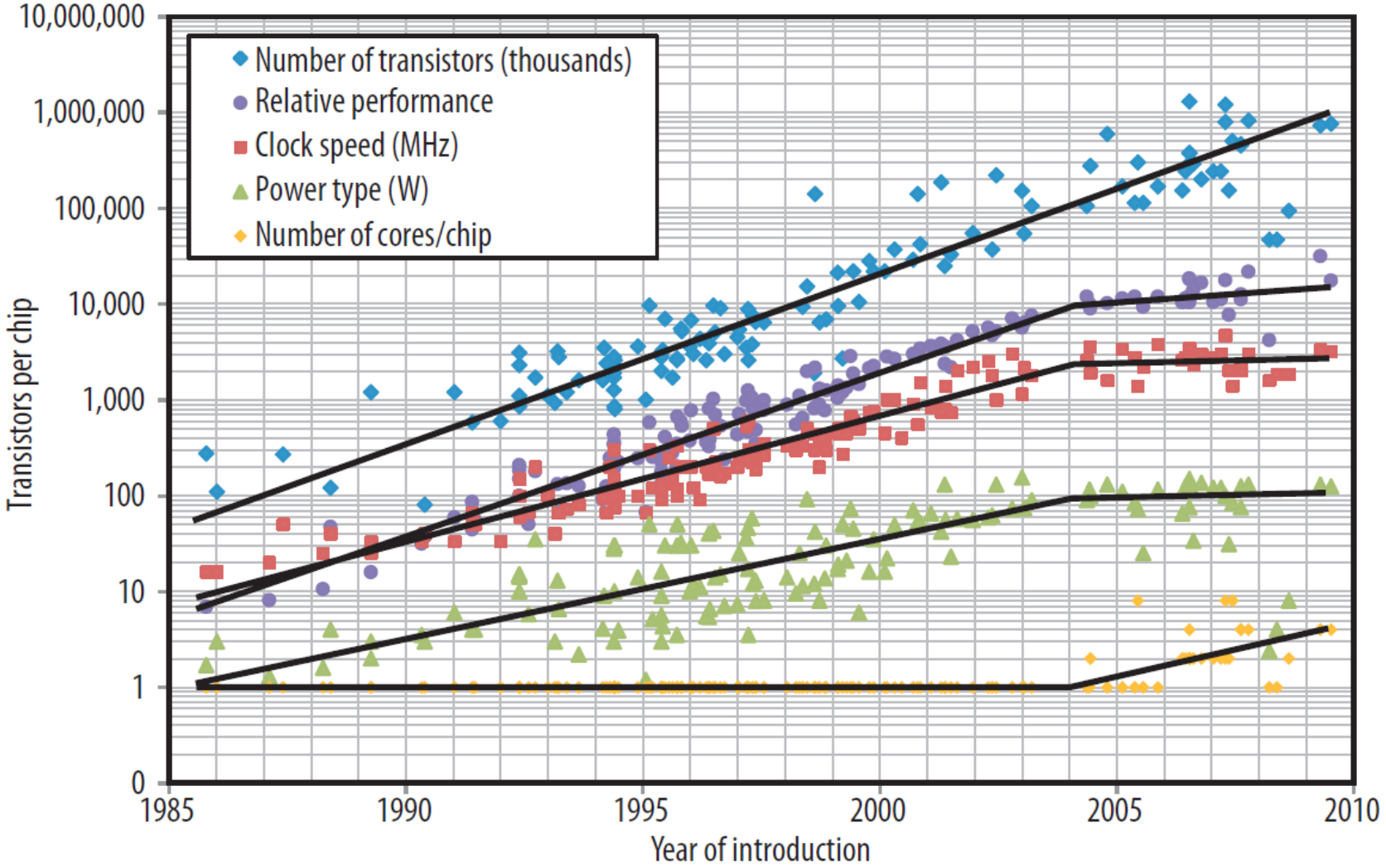
Seoul National University

Fall 2019

What We Have Learned

- Instruction Set Architecture (i.e., abstraction of hardware)
- Representing numbers: integer and floating-point
- RISC-V assembly and how to translate C program into it
- Basic processor organization
- Pipelining
- Branch prediction
- Locality and memory hierarchy
- Caches
- Program optimization for caches
- Virtual memory
- Performance evaluation

Remember? – The End of Historic Scaling



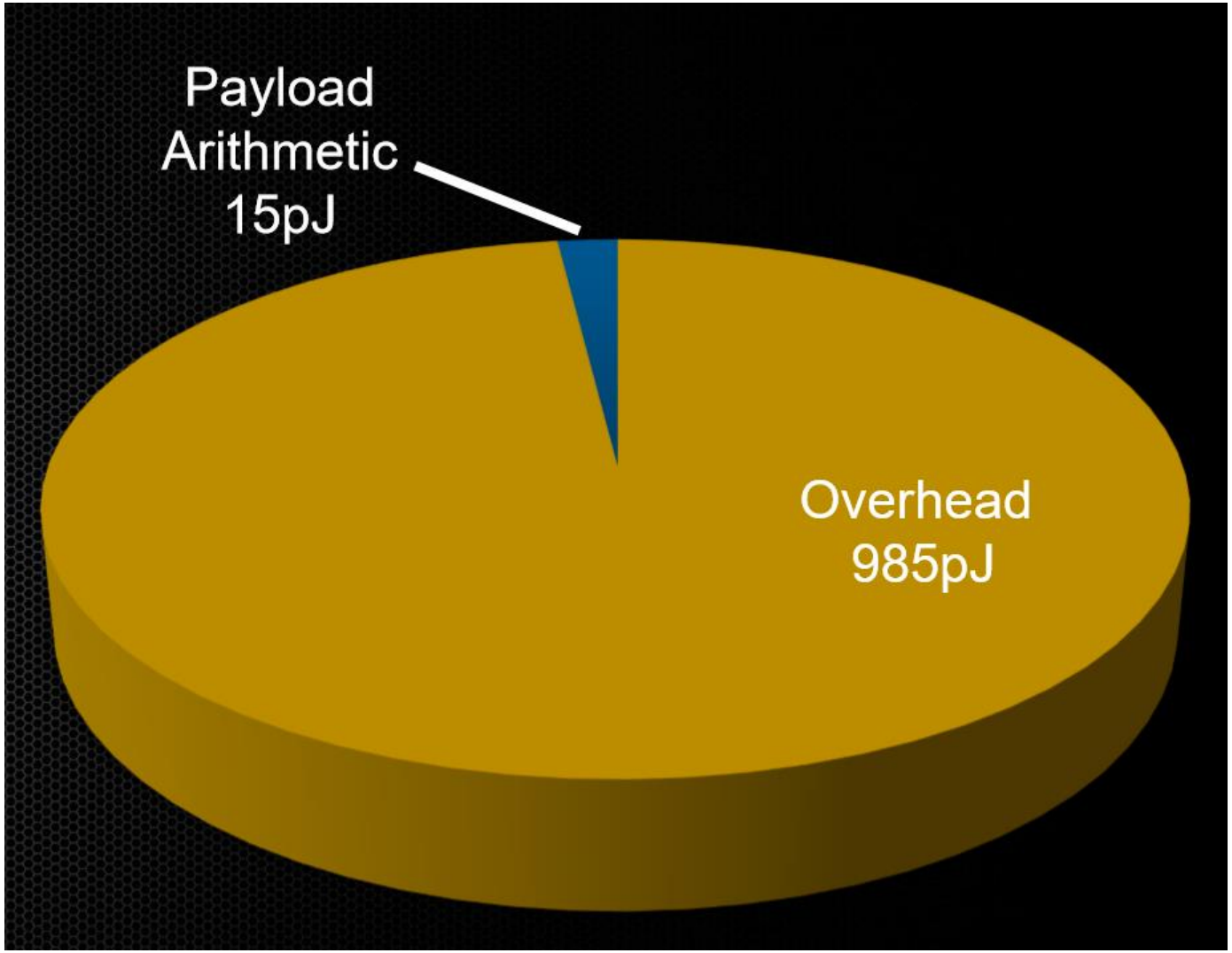
Source: IEEE Computer, 2011

Beyond Pipelining

- Hyperthreading
- SIMD (MMX, SSE)
- Vector processor (AVX-512)

- Manycore (Xeon Phi: Knights Ferry / Corner / Landing / ~~Hill~~)
  - Knights Landing (KNL): 72 cores, 4 threads / core
- GPUs
  - Nvidia Titan V (Volta): 5120 CUDA cores + 640 tensor cores
- Accelerators
  - Amazon EC2 F1 instances: up to 8 Xilinx FPGAs (each with 2.5M logic elements)

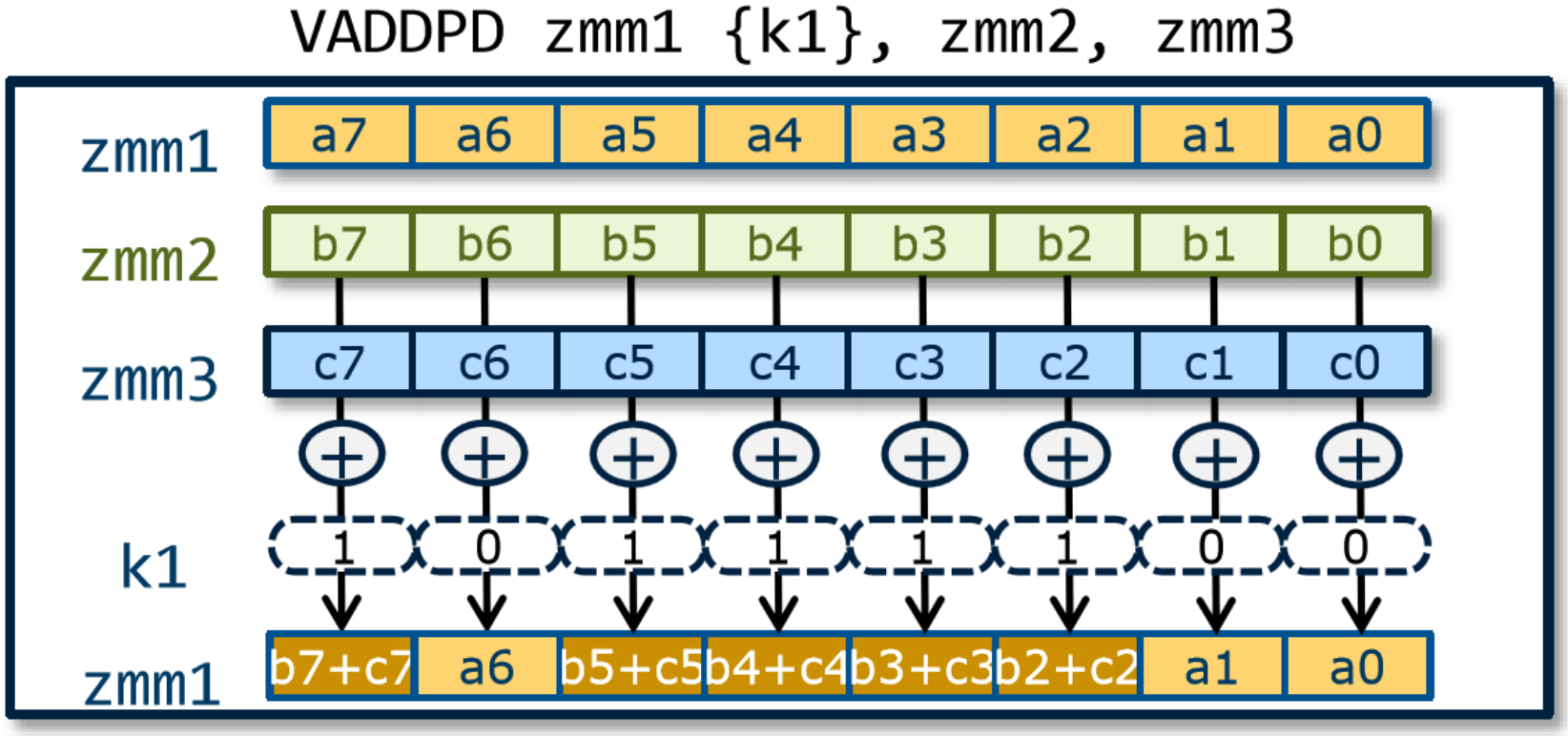
CPU Overhead



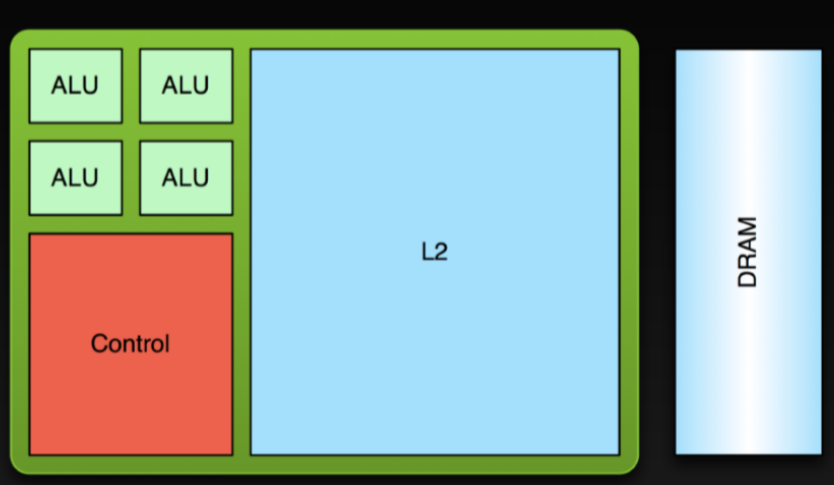
Source: Bill Dally, "Deep Learning and HPC," 2017.

Intel AVX-512

- 32 512-bit registers (%zmm0-31)
  - 64 bytes
  - 32 words
  - 16 doublewords or single-precision FPs
  - 8 quadwords or double-precision FPs
- 8 mask registers

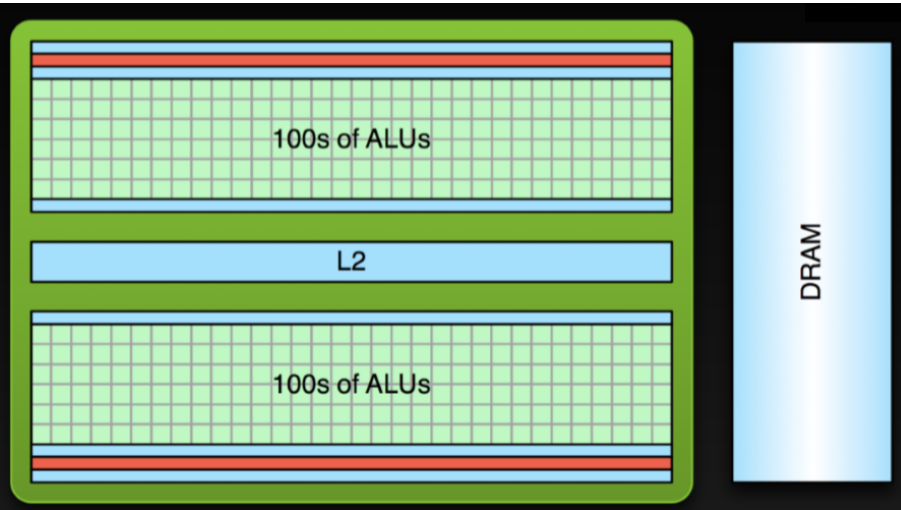


NVIDIA GPU



CPU

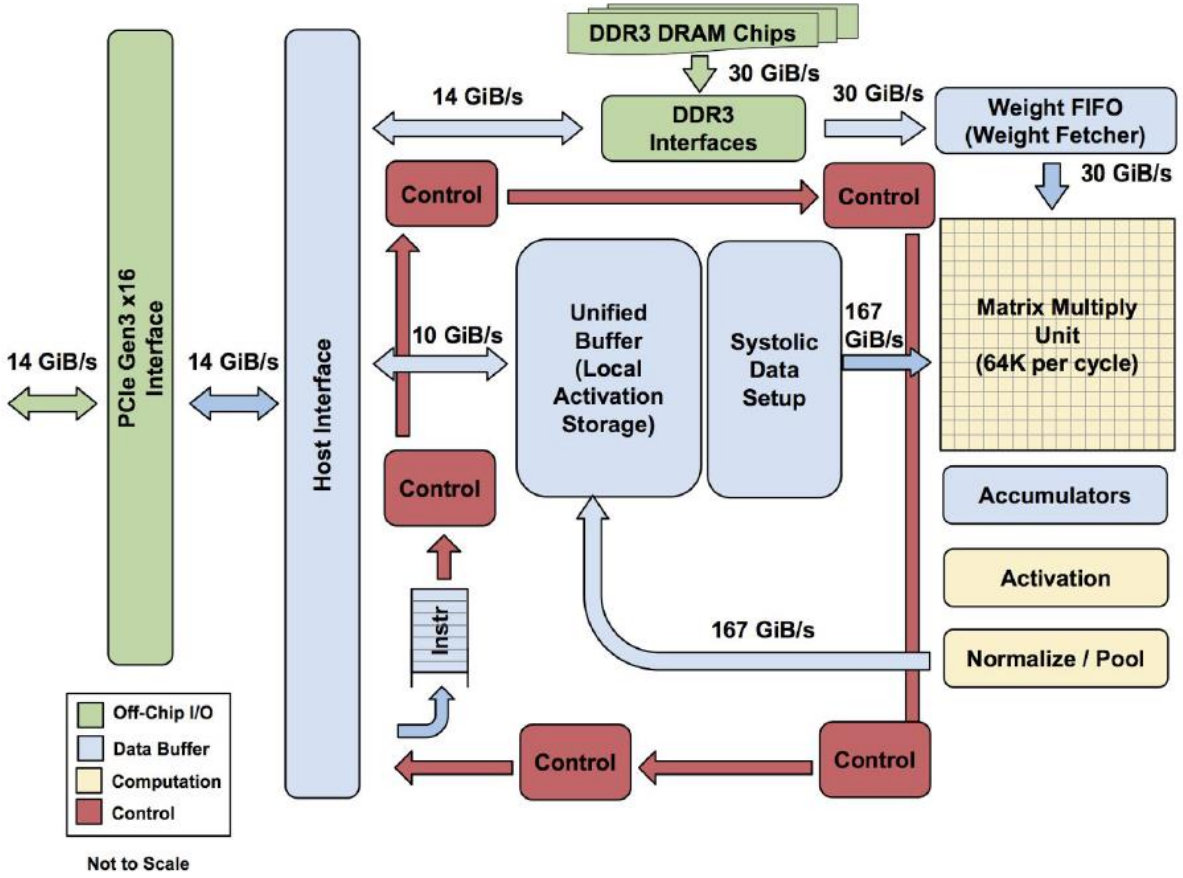
- Optimized for low-latency access to cached data sets
- Control logic for out-of-order and speculative execution

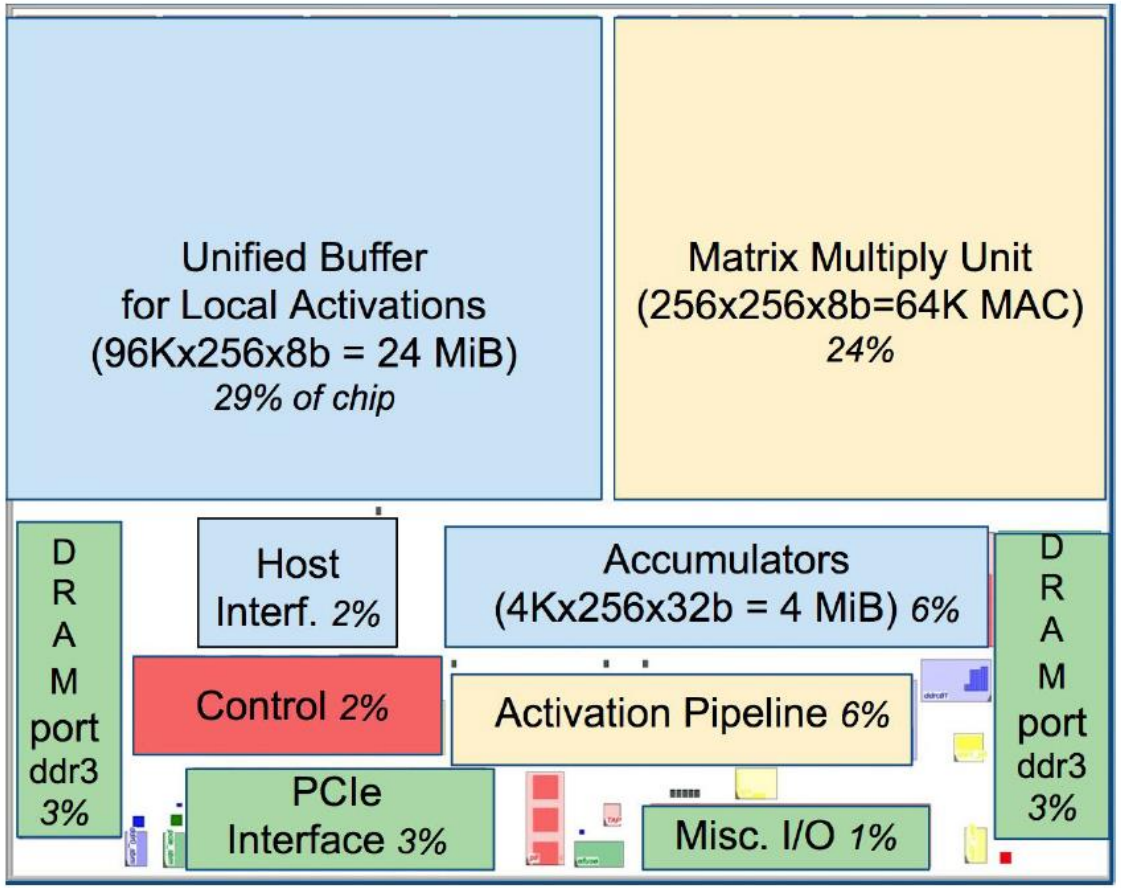


GPU

- Optimized for data-parallel, throughput computation
- Architecture tolerant of memory latency
- More transistors dedicated to computation

Google TPU (Tensor Processing Unit)





Computer Systems Research

"domain-specific computing"
or "vertical optimization"

- Architecture
- OS
- Compiler
- Network
- Database
- …
  +
- Security