Thinking Outside the Die - Summary

The Grand ML Challenge

- Cerebras' goal: Accelerating AI with chip, system and software
- Exponential growth of memory requirements for NN over the years
- (Micro)architecture has improved over the years, however not in the same scale needed
    - Cluster scale-up as an intermediate solution, but not future proof
- Massive models require massive memory, compute and communication
- New ways needed to scale up all of them

Amplifying Moore's Law

- Extending beyond a single die in the industry
    - At Cerebras: Wafer-Scale Engine
        - Largest chip in the world: Cluster scale acceleration on a single chip
    - Problems: Solved with co-design at TSMC etc.
        - Yield problem: Defects in the wafers
            - Redundant cores to replace defective cores
        - Lithography limitations
            - Using cross-die wires
        - Power and cooling
            - 3rd dimension with water cooling and power delivery perpendicular to wafer

Designing From Ground Up for Neural Networks

- NN expressed as GEMMs
- NN are naturally sparse or can be made sparse
    - Up to 10x gains
- Existing GEMM architectures are dense-only
    - Memory bandwidth limitations and are structured in the computation
- Memory bandwidth limitations
    - Central shared memory is slow and far away
    - Need caching and date reuse
    - Better way: fully distributed memory with cores
    - Full memory bandwidth helps performance a lot
- Architecture for unstructured sparsity
    - Dataflow scheduling in hardware
    - Filter out 0s at hardware to skip unnecessary processing
- Near-linear sparsity acceleration

Inherently Scalable Clustering

- Challenges to existing scale-out solutions
    - Data parallel, pipelined model parallel, tensor model parallel
        - Usually, high communication overhead or memory limits
        - Same limitations: memory tied to compute
- Cluster is the ML accelerator
    - Disaggregation of memory and compute
    - 850k compute cores in a single chip
    - MemoryX technology, 120 trillion params
    - SwarmX interconnect technology for near-linear performance scaling up to 192 CS-2s
- Solving latency problems
    - Removing latency-sensitive communication
    - Coarse-grained pipelining: no inter layer dependencies
    - Fine-grained pipelining
- MemoryX: scalable to extreme model sizes
    - Independent from compute
- SwarmX: Weights are broadcast to all CS-2s and reduced on the way back
    - Independent from capacity