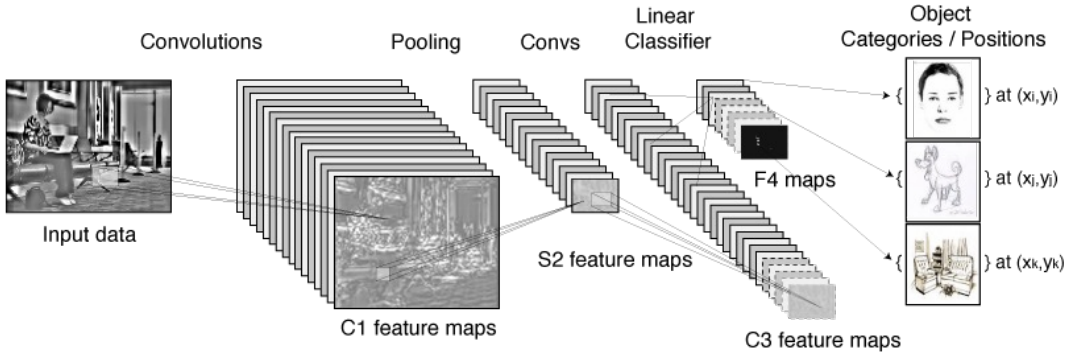
Research Overview of Energy-Efficient Multimedia Systems Group

Vivienne Sze

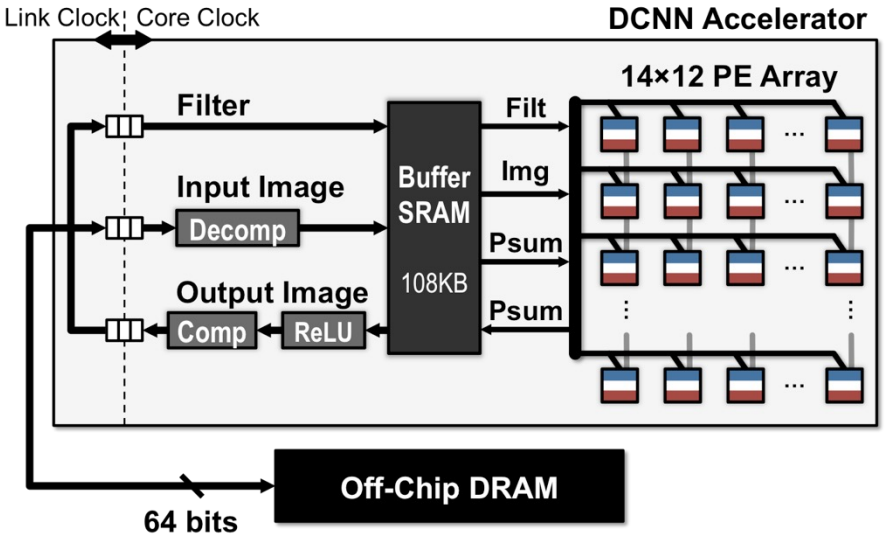
Efficient Computing with Cross-Layer Design

Algorithms

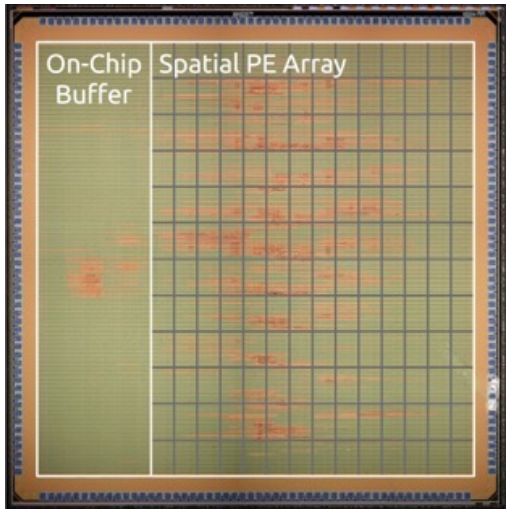


Systems

Architectures



Circuits



Accelerate the processing of sparse tensor workloads

- Sparse tensors used for tasks like deep neural networks and graph analytics

- Exploit sparsity to reduced amount of compute
  - e.g., Anything multiplied by **zero** is **zero**

- Exploit sparsity to reduced amount of data to be stored
  - Zeros don't need to be stored – apply compression

Challenge: Efficiently handle varying sparsity both *across* and *within* tensors

Papers to Appear at MICRO 23 (This Week)

- **HighLight: Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity**
  - Addresses varying sparsity *across* tensors
  - Project Website: https://emze.csail.mit.edu/highlight

- **Tailors: Accelerating Sparse Tensor Algebra by Overbooking Buffer Occupancy**
  - Addresses varying sparsity *within* tensors
  - Project Website: https://emze.csail.mit.edu/tailors