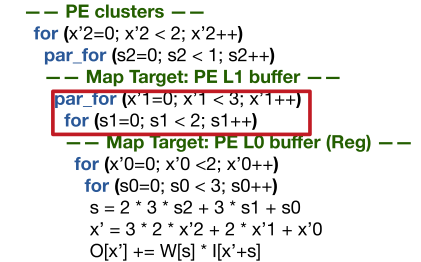
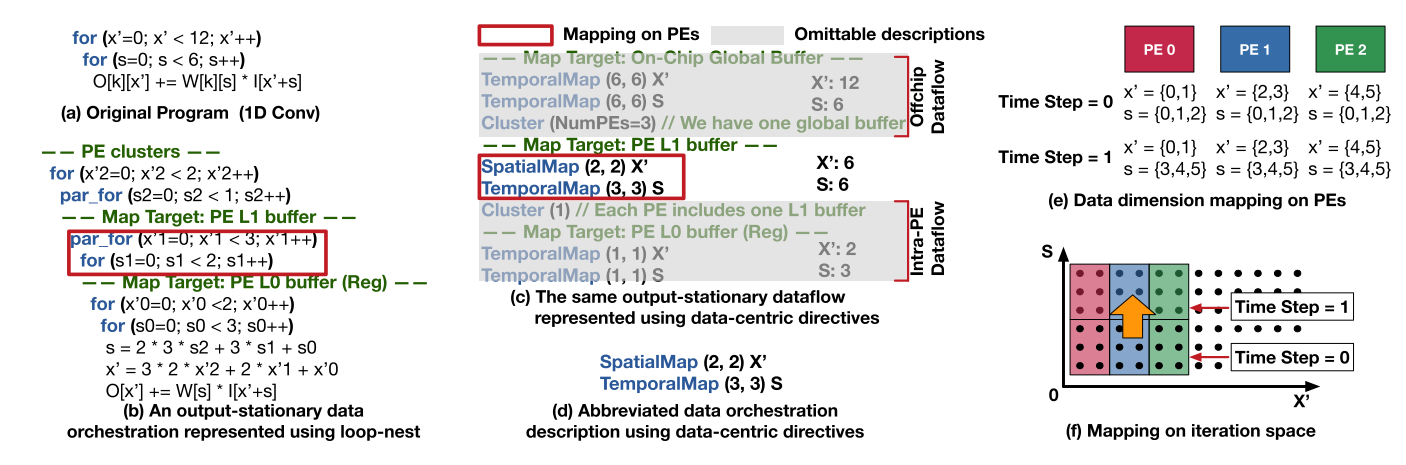
**Understanding Reuse, Performance, and Hardware Cost of DNN Dataflows: A Data-Centric Approach**

**——MICRO 2019 NVDIA，Georgia　Institute　of　Technology**

1. Contributions
   1. Discuss data reuse in DNN
   2. Propose a new method to represent dataflow
   3. Propose a analysis framework named MAESTRO to quantitatively estimating runtime and energy efficiency of dataflows on a target DNN model and hardware configuration
2. Data reuse
   1. Data reuse originates from two behaviors of DNN accelerators over time and space🡪multicasting and reduction
   2. Multicasting: deliver the data to multiple PEs, multiple temporal destinations
   3. Reduction: accumulates multiple data spatially or temporaly
3. Data-Centric representation
   1. Existing expressions: use output-stationary for an example

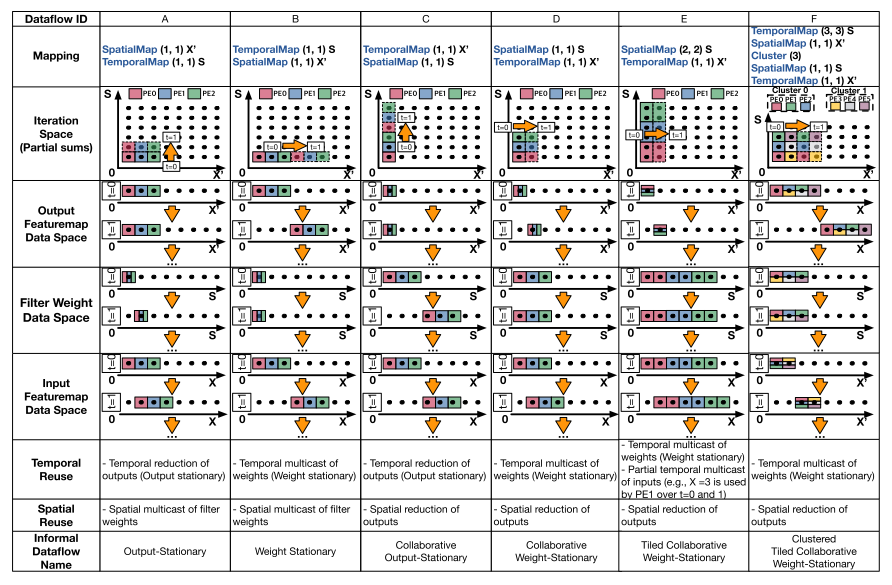


* 1. Spatial Map: size, refers to the number of indices mapped in the dimension to each PE，offset, refers to the shift in the starting indices of selected dimension across consecutive PEs; selected dimension
  2. Temporal Map: size, the same as spatial map; offset, describes the shift in the starting indices of selected dimension across consecutive time steps.
  3. Data movement order & directives order: the sequence of spatial and temporal maps in the dataflow specification dictates the order of data movement.
  4. Clusters: All the mapping directives specified above a cluster directive perform the mapping across logical clusters created by the cluster directive; All the mapping directives specified below a cluster directive perform the mapping across PEs or lower level logical clusters inside a logical cluster created by the cluster directive.
  5. Comparison with previous representation:



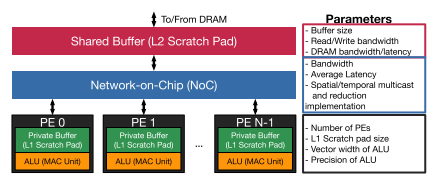
Example: 1D convolution(kernel size: 6, output pixels: 12)

* 1. More examples:

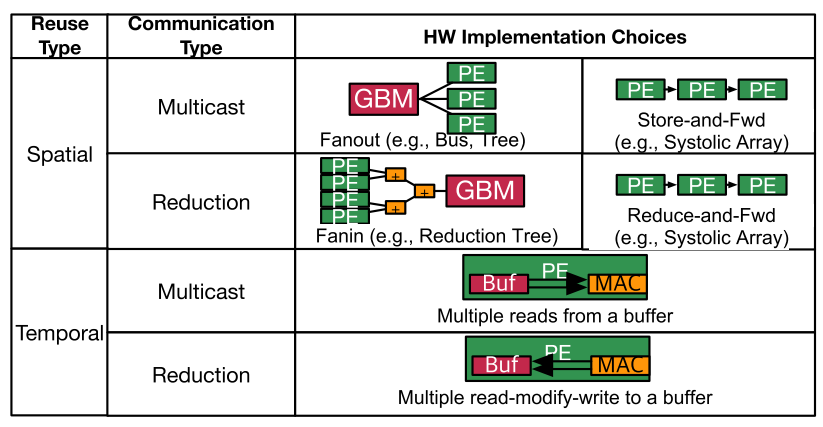


Different choices of data-centric representation

1. Hardware implications of reuse



Abstract hardware model



Hardware Implementation Choices for supporting spatial and temporal reuse

1. MAESTRO Analysis Framework

