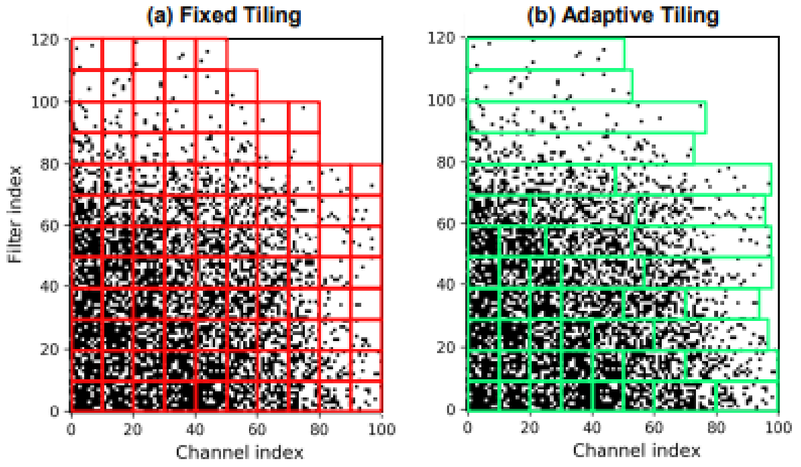
ADAPTIVE TILING: APPLYING FIXED-SIZE SYSTOLIC ARRAYS TO SPARSE CONVOLUTIONAL NEURAL NETWORKS

FIXED VS. ADAPTIVE TILING

- Fixed Tiling: A conventional 10x10 systolic array can only cover 10x10 dense tiles, resulting in many tiles required to cover the entire matrix.
- Adaptive Tiling: A 10x10 systolic array with adaptive tiling can cover sparse tiles of larger sizes by increasing the tile width, resulting in fewer total tiles and reduced runtime.

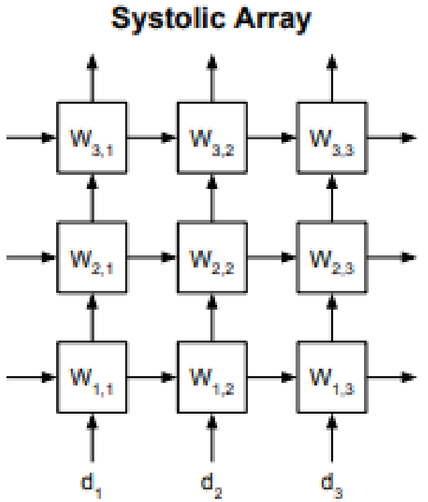


INTRODUCTION

- Matrix Multiplication in CNNs: CNN computations largely involve matrix multiplications. Systolic arrays are used to speed up these operations by minimizing I/O costs.
- Systolic Arrays and Sparsity: Traditional systolic arrays are designed for dense matrices, but many efficient CNN architectures have sparse filters with a high percentage of zero weights. This sparsity leads to inefficiencies when using systolic arrays designed for dense matrices because zero weights still occupy space in the array.
- Tiling Challenge: Efficiently partitioning CNN layers into blocks (tiles) that fit into a fixed-size systolic array is crucial. Minimizing the number of these tiles reduces the number of calls to the systolic array, thereby optimizing performance.

ADAPTIVE TILING

- Systolic arrays are traditionally used for dense matrix multiplications. However, CNNs, especially those optimized through pruning techniques, often have sparse filters where many weights are zero. Conventional systolic arrays do not handle sparsity efficiently, as zero weights still occupy cells in the array.
- Adaptive tiling addresses the challenge of implementing sparse CNN layers with a fixed-size systolic array by:
- Tile Adaptation: Allowing tiles to cover larger areas with high sparsity, thereby reducing the total number of tiles needed and the number of systolic array calls.
- Multiple Columns per Systolic Cell: Adding multiple input data columns per cell in the systolic array allows it to cover wider sparse areas with minor additional hardware cost.



IMPLEMENTATION

- The key idea is to allow each cell in the systolic array to handle multiple input data columns.
- By combining several columns into a single cell, the array can span wider areas for sparse regions.
- This reduces the number of calls to the systolic array.

- Hardware Considerations:
- Adding multiple columns per cell incurs minor hardware costs.
- These costs are negligible compared to the benefits of reduced Multiply-Accumulator (MAC) operations.