

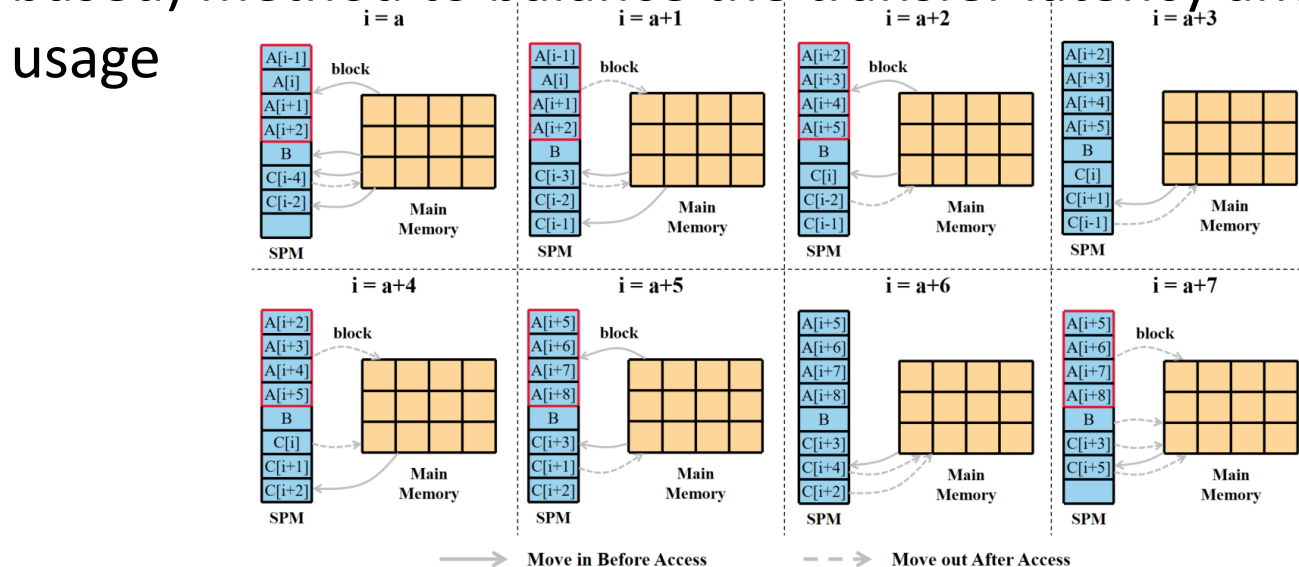
# MuDP: Multi-Granularity Data Placement for Uniform Loops on SPM-DRAM Architectures to Minimize Latency

**Yixuan DU, Edwin Hsing-Mean SHA, Yuhong SONG,  
Yibo GUO, Longshan XU, Qingfeng ZHUGE**

Frontiers of Computer Science, DOI: [10.1007/s11704-023-3566-y](https://doi.org/10.1007/s11704-023-3566-y)

# Problems & Ideas

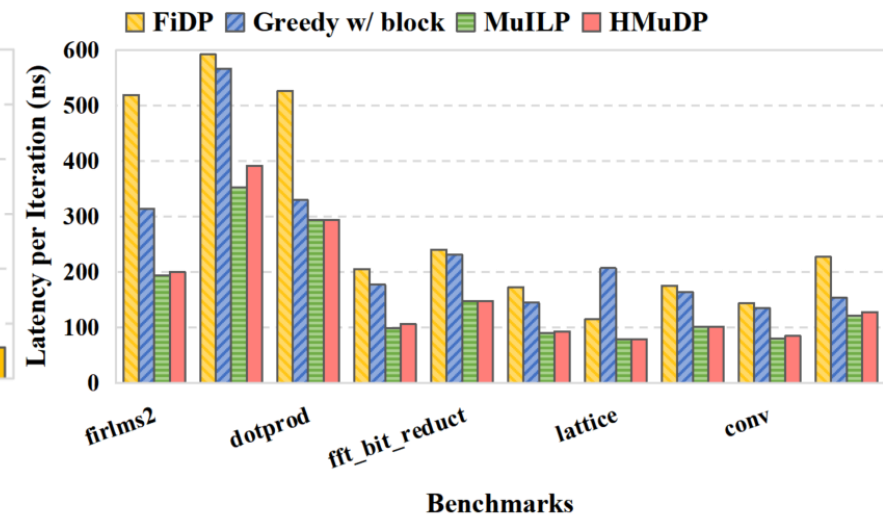
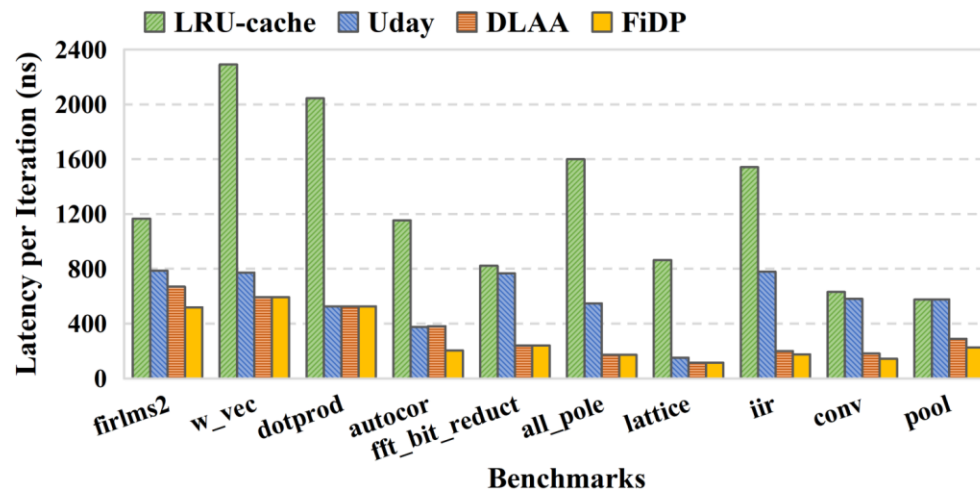
- Problems of conventional SPM data placement methods:
  - Coarse-grained data placement method causes high space usage. Fine-grained data placement method consumes excessive access setup time.
  - Existing fine-grained data placement method tend to fail in solving the case of inconsecutive array access in loops.
- Ideas: use both fine-grained and medium-grained (block-based) method to balance the transfer latency and space usage



The optimal solution to place variables in 8 iterations using our MullP method, where  $a$  is a constant value representing a start of a placement cycle. The solution demonstrates the movement of variables in each iteration. For example, when  $i=a+3$ ,  $C[i+1]$  is moved in to SPM before access and  $C[i-1]$  is move back to main memory after access.

# Main Contributions

- Contributions:
  - A fine-grained dynamic programming-based algorithm, called FiDP, which yields the minimal latency by addressing the insufficiency space and solves the case of inconsecutive array access;
  - A multi-granularity data placement algorithm with a medium-grained block access mode to mitigate the activation overhead, called MuILP, which generates an optimal solution;
  - A heuristic multi-granularity data placement method, called HMuDP, which achieves near-optimal results and polynomial time complexity.



Per iteration latency. Left: comparison among LRU-cache, Uday, DLAA and FiDP on ten benchmarks. Right: comparison among FiDP, Greedy w/ block, MuILP and HMuDP on ten benchmarks.