DAC YF

Reducing Bit Writes in Non-volatile Main Memory by Similarity-aware Compression*

Zhangyu Chen, Yu Hua, Pengfei Zuo, Yuanyuan Sun, Yuncheng Guo
Huazhong University of Science and Technology

Background

- Image bitmaps are widely used in memory for fast accesses
  - Application: image processing and machine learning
  - High space and energy consumption in DRAM
- Non-volatile main memory (NVMM) is ideal for bitmaps
  - **Pros**: high density, near-zero stand-by power
  - **Cons**: high latency and energy for writes

- Bit-write reduction in NVMM is cost-inefficient
  - Bitmaps are hard to match the general-purpose data patterns
  - Different patterns due to various bitmap formats

Partial patterns in FPC [NANOARCH'14]

Prefix	Pattern	Example
000	Zero run	0x00000000 => 0x0
001	4-bit sign extended	0x00000002 => 0x12
010	1-byte sign extended	0xFFFFFFCC => 0x2CC
011	Halfword Sign Extended	0x00001C23 => 0x31C23

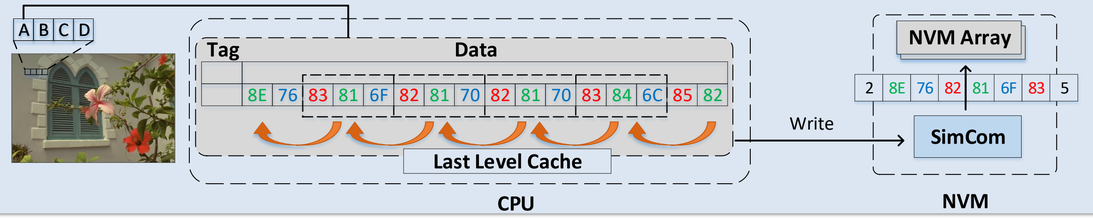
RGB bitmap (three channels): 8E 76 83 81 6F 82 81 70 82 81 70 83 84 6C 85 82

Gray bitmap (one channel): 7E 7B 7B 7B 7B 7C

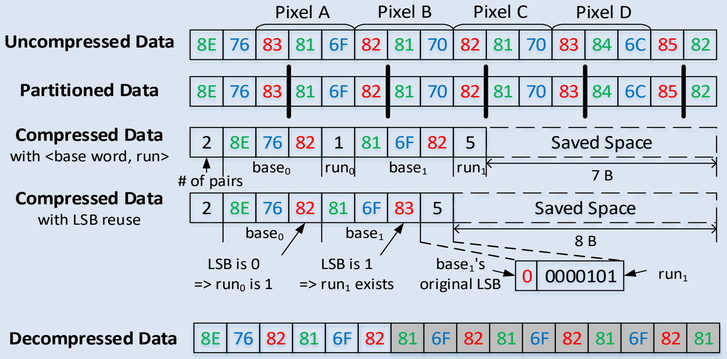
Problem: inefficient bit-write reduction for bitmaps in NVMM (e.g., 94.2% compression ratio for FPC)

The SimCom Design

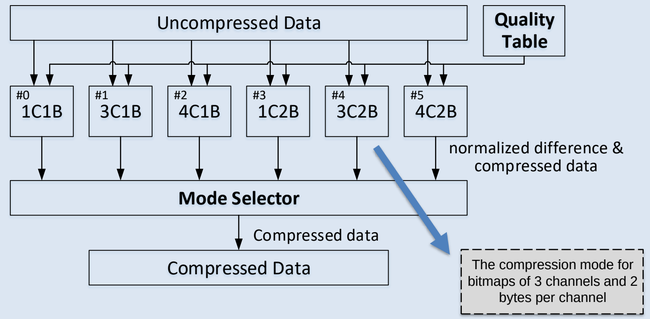
- Idea: leverage the *pixel-level similarity* and *error-tolerance* for approximate compression



- Similarity-aware data compression

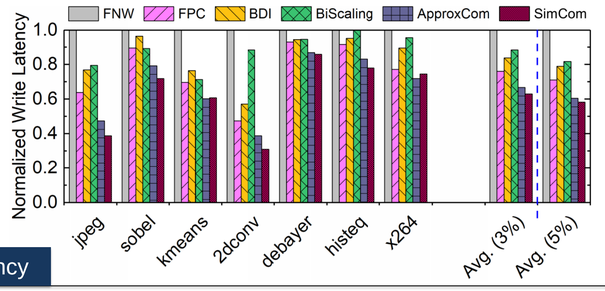


- Adaptive compression scheme

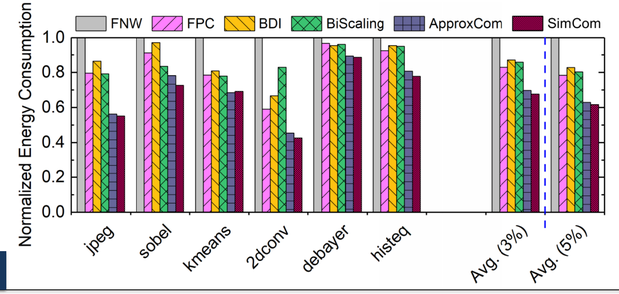


Evaluation

- Comparisons: FNW [MICRO'09], FPC [NANOARCH'14], BDI [PACT'12], BiScaling [ISLPED'17]



Write latency



Energy

* Zhangyu Chen, Yu Hua, Pengfei Zuo, Yuanyuan Sun, Yuncheng Guo, "Reducing Bit Writes in Non-volatile Main Memory by Similarity-aware Compression", Proceedings of the 57th Design Automation Conference (DAC), 2020.