ICS

KU-The Future

SparsePIM:

An Efficient HBM-Based PIM Architecture for Sparse Matrix-Vector Multiplications

Taewoon Kang, Geonwoo Choi, Taeweon Suh, Gunjae Koo

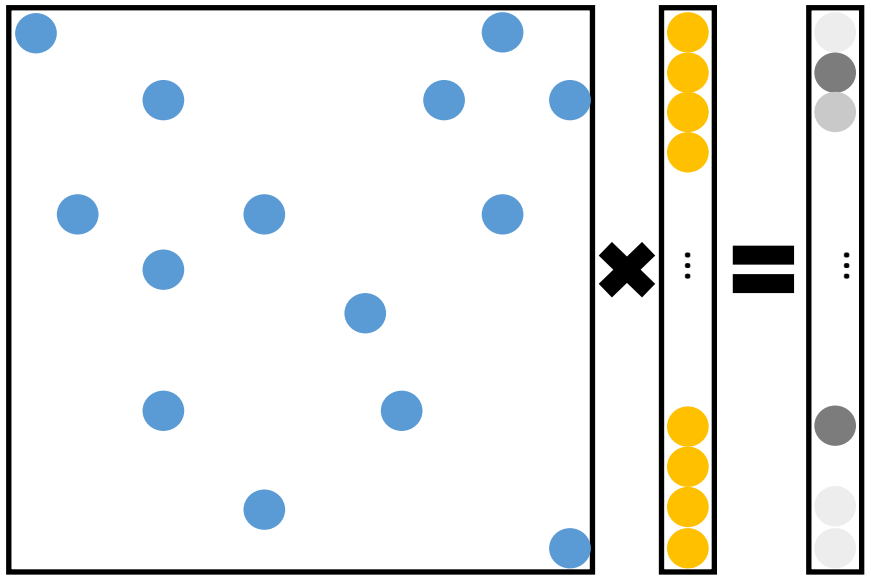
Korea University

ACM International Conference on Supercomputing 2025

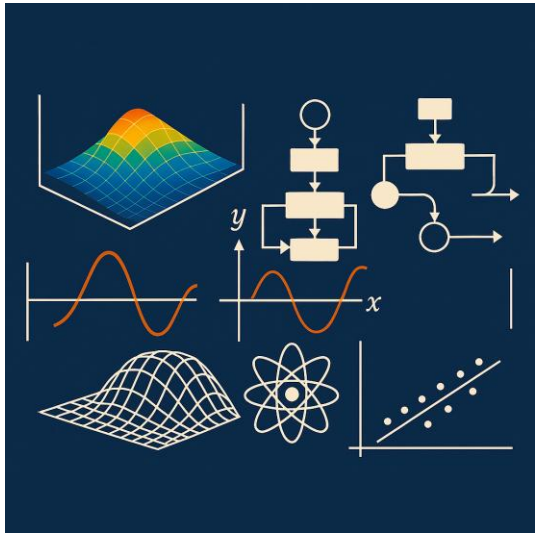
KOREA UNIVERSITY

Sparse Matrix-Vector Multiplication

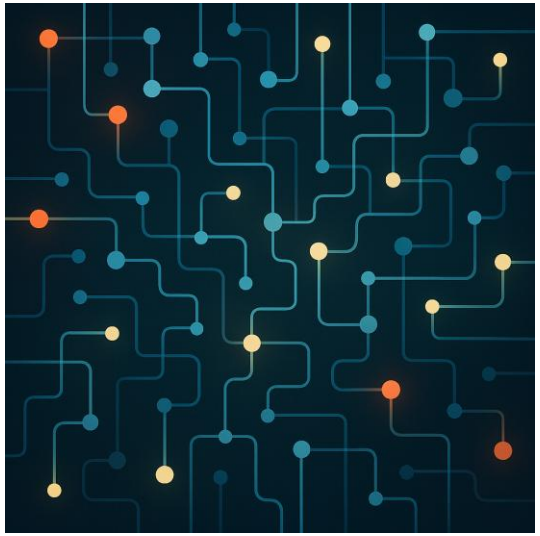
SpMV is widely used in



Sparse Matrix-Vector multiplication

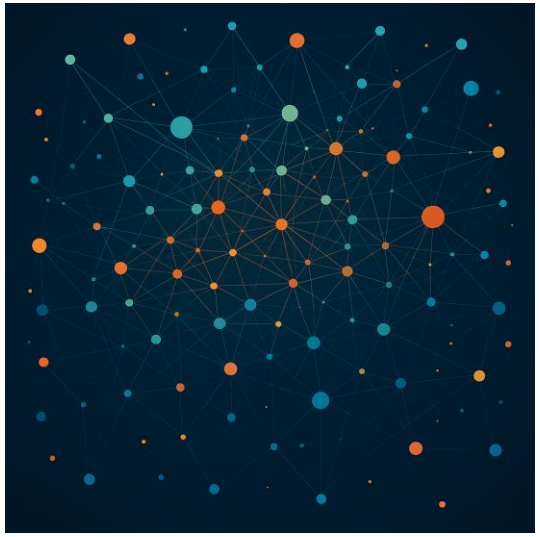


Scientific computing



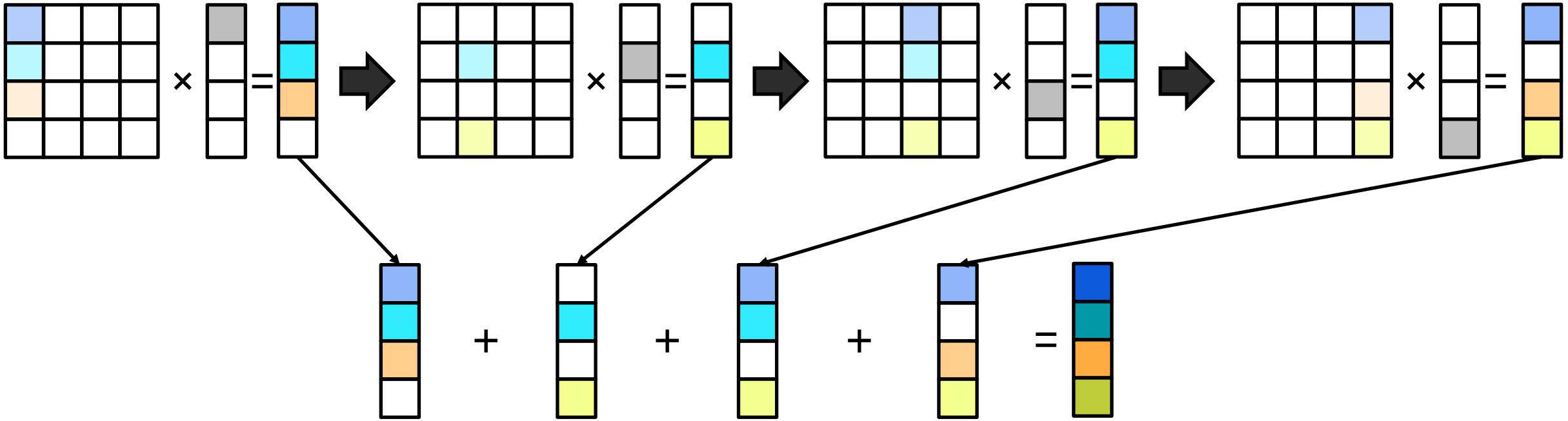
Circuit simulation

AI workload



Graph analysis

Outer Product SpMV



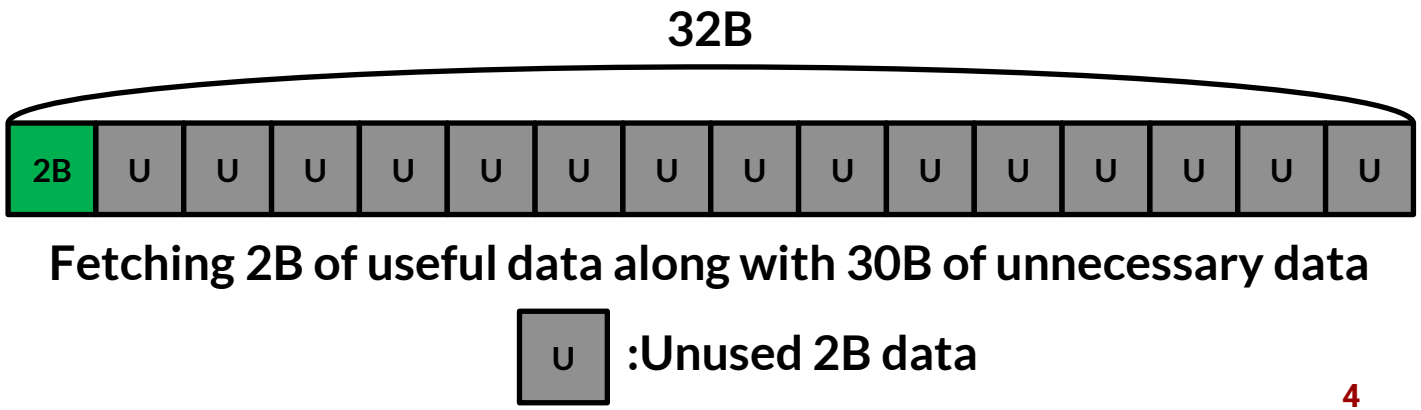
Conventional Compression Format

- SpMV using conventional format
  - Causes indirect access
  - HBM-PIM can't support indirect access

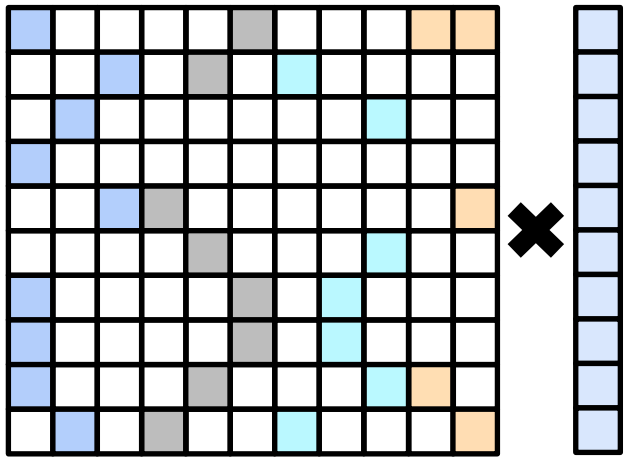
  + Fetched data wasted

```
for (i = 0; i < N_ROWS; i++) {
  sum = 0;
  for (j = row_ptr[i]; j < row_ptr[i+1]; j++) {
    sum += val[j] * vec[cols[j]];
  }
  (*out)[i] += sum;
}
```

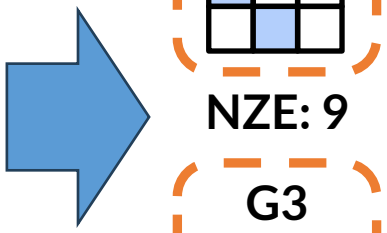
<SpMV using CSR format>



Without Software Optimization

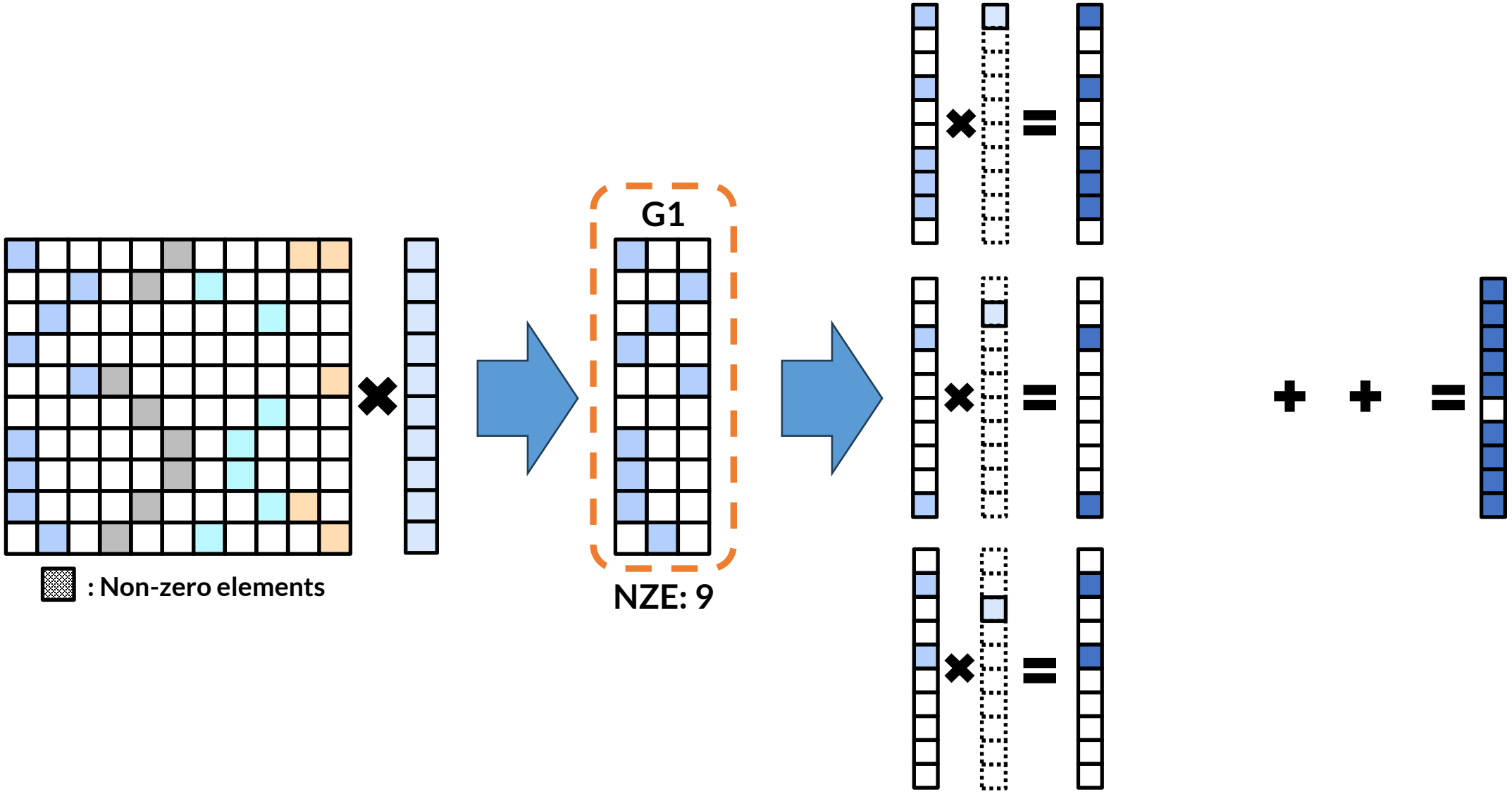


: Non-zero elements

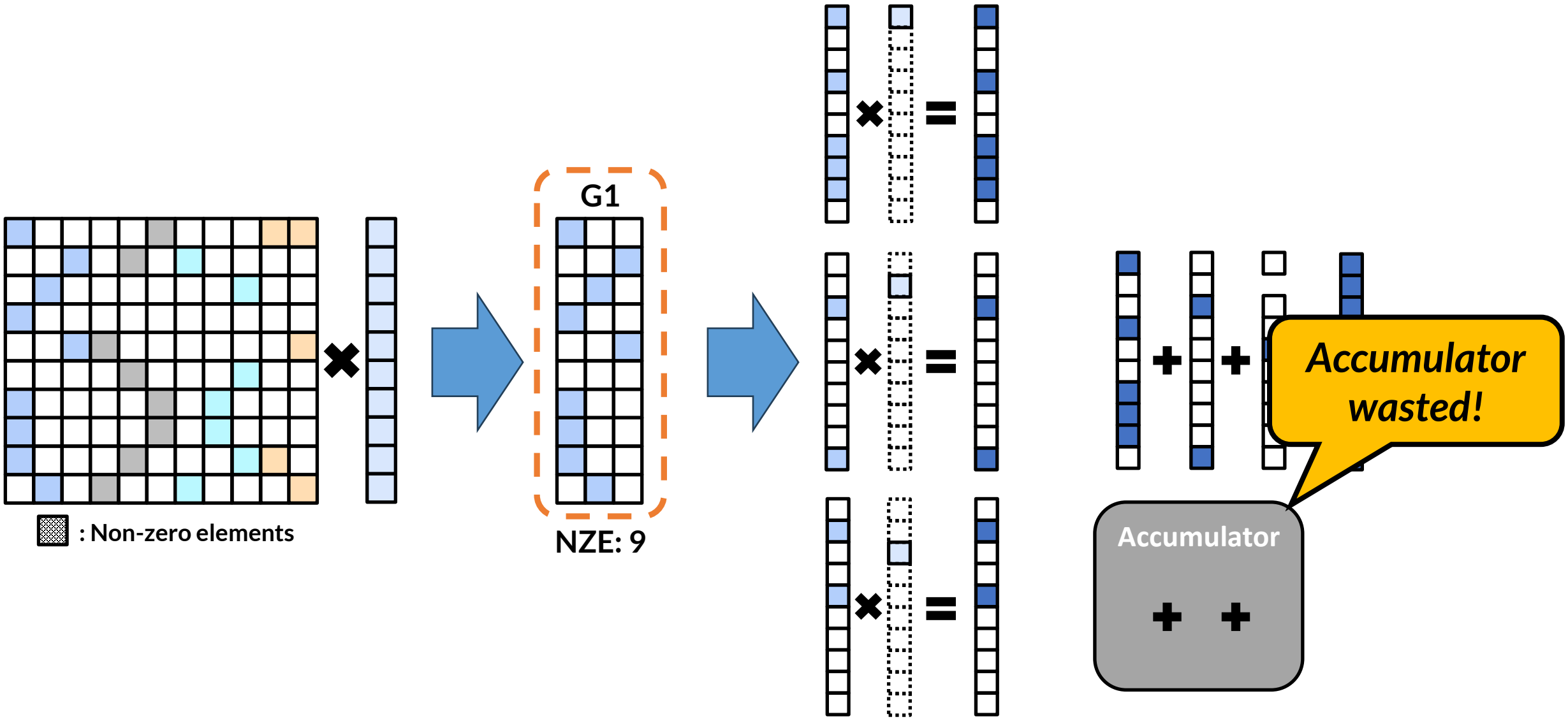


Load imbalance

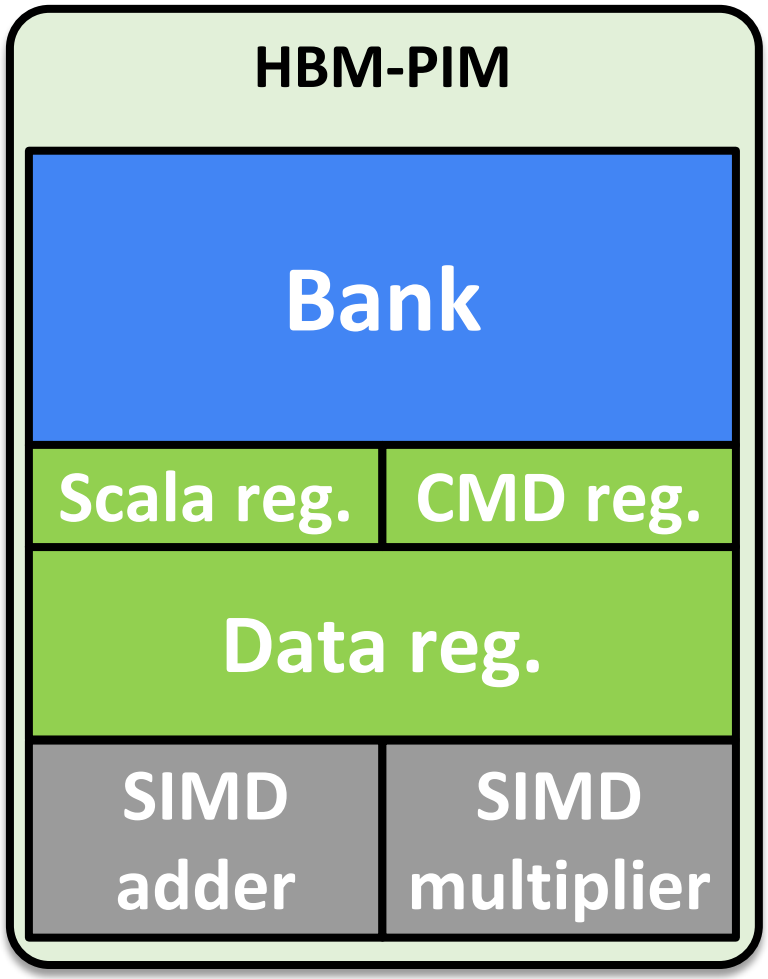
Without Software Optimization

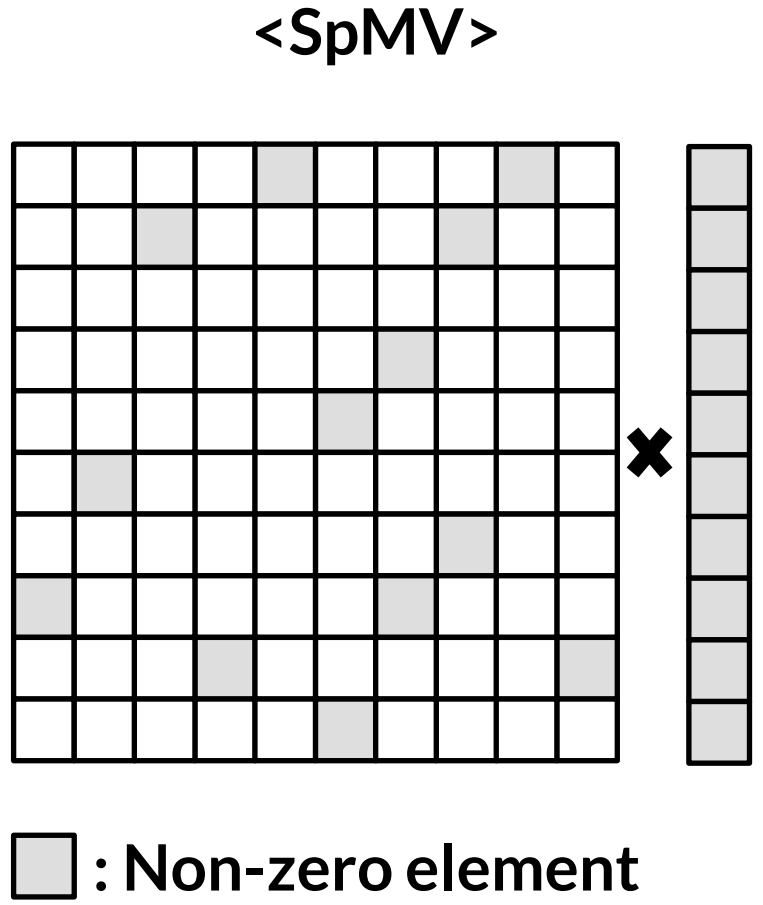


Without Software Optimization



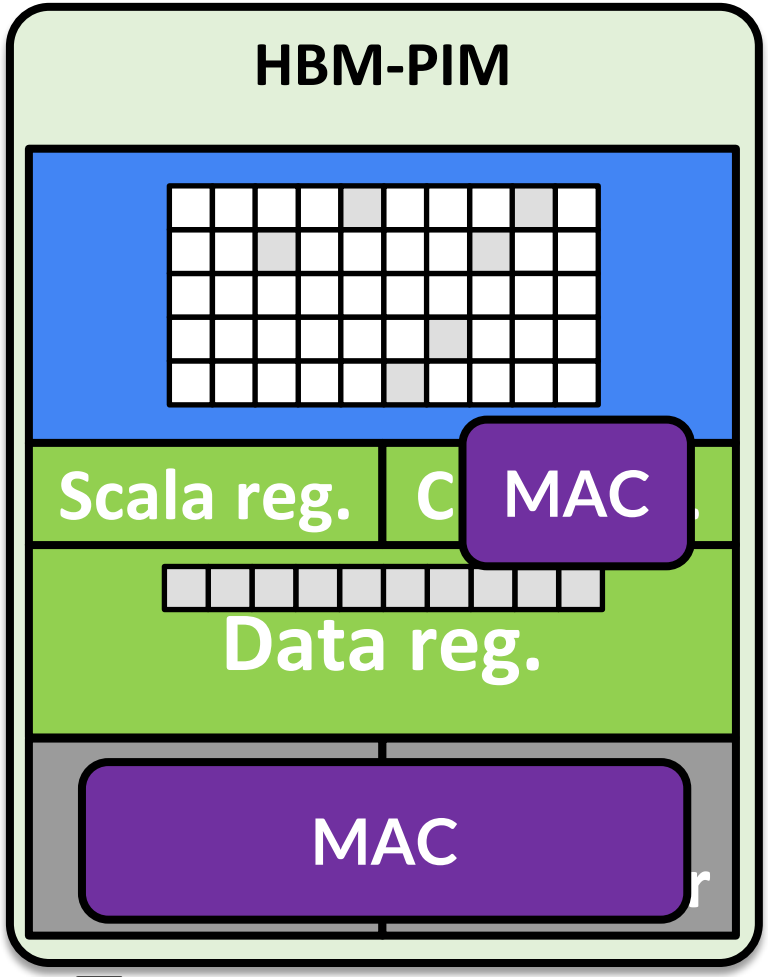
Problems Of Previous Work





Lee, Sukhan, et al. "Hardware architecture and software stack for PIM based on commercial DRAM technology: Industrial product." *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021.

Problems Of Previous Work



: Matrix-vector multiplication result

: Non-zero element

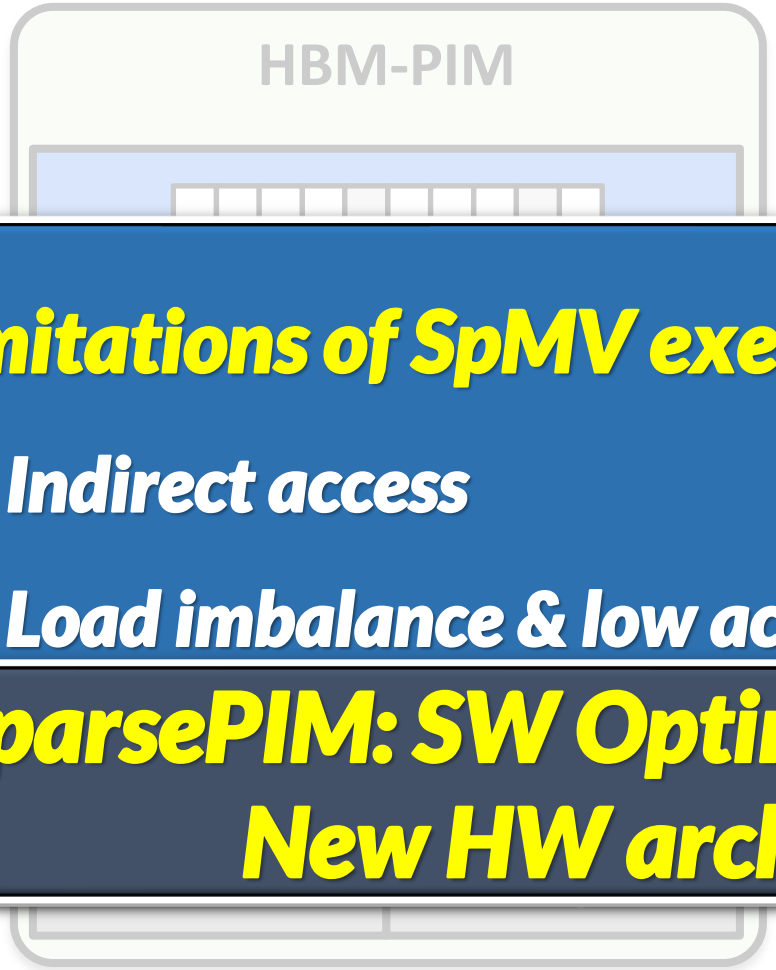
- Webbase-1M
  - Sparsity over 99.99%
  - Size of matrix ≈ over 1,000,000

Compute unit wasted

Matrix row used in SpMV

Lee, Sukhan, et al. "Hardware architecture and software stack for PIM based on commercial DRAM technology: Industrial product." *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021.

Tackling SpMV Challenges



- Sparse matrix
  - Sparsity ≈ 99.9999%

Limitations of SpMV execution

- ***Indirect access***
- ***Load imbalance & low accumulation ratio***

SparsePIM: SW Optimization + Compression format + New HW architecture to handle SpMV

: Matrix-vector multiplication result

: Non-zero element

Lee, Sukhan, et al. "Hardware architecture and software stack for PIM based on commercial DRAM technology: Industrial product." *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021.

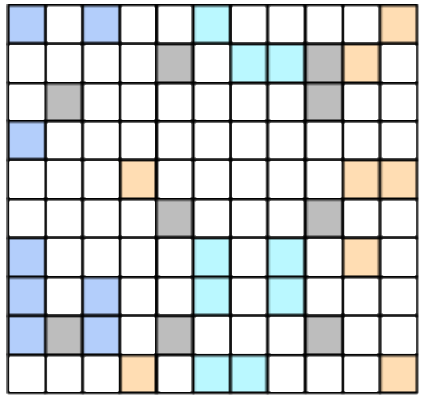
Overview

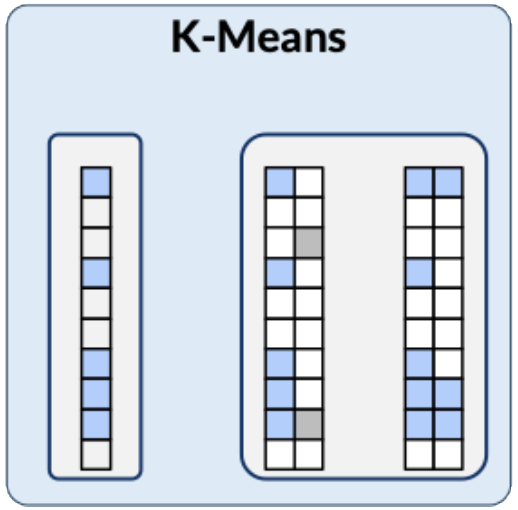
DRAF

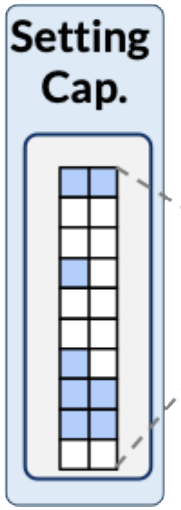
1KB

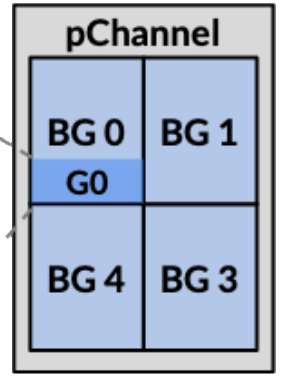
Column Idx.	Rsvd	Values	Row Idx.	Partial result	Vec.	Rsvd
28B	4B	224B	448B	288B	14B	18B

SW optimization

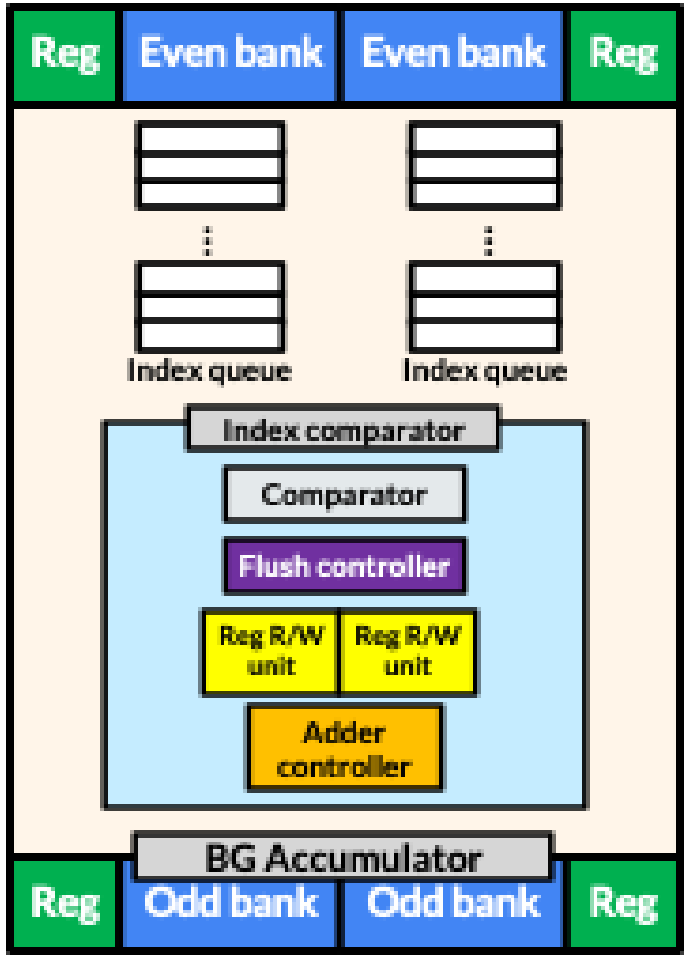




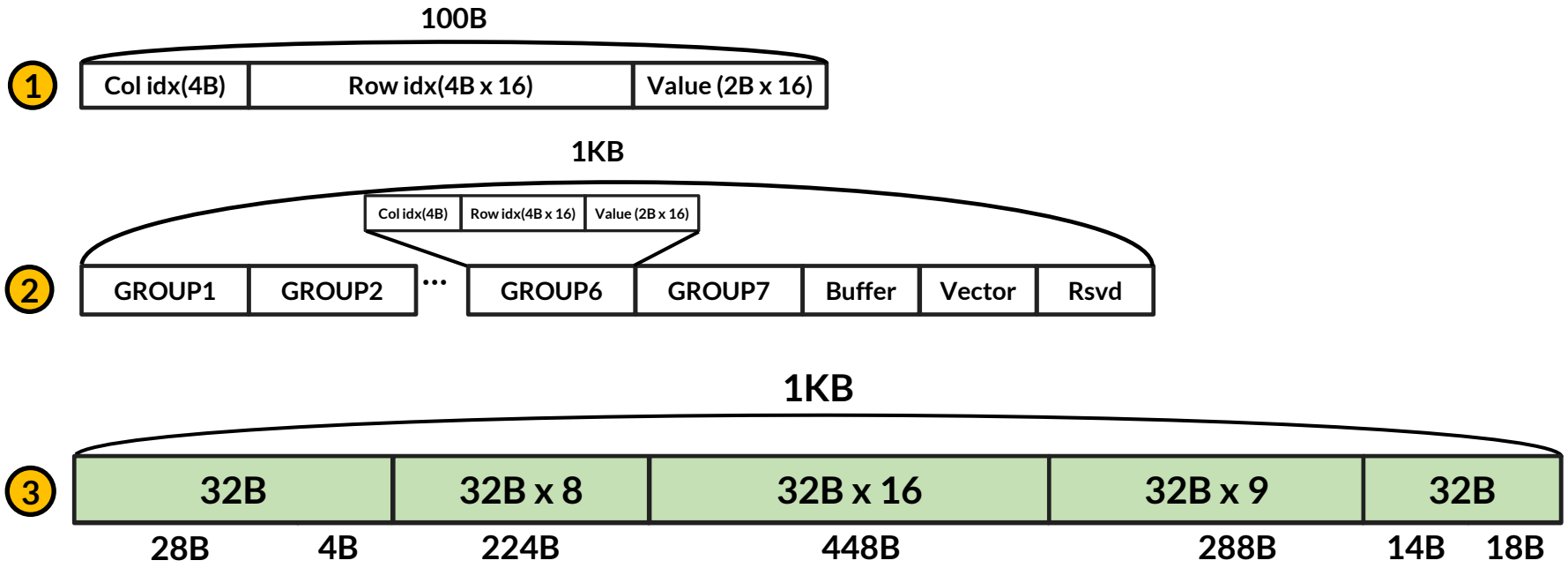




BG accumulator



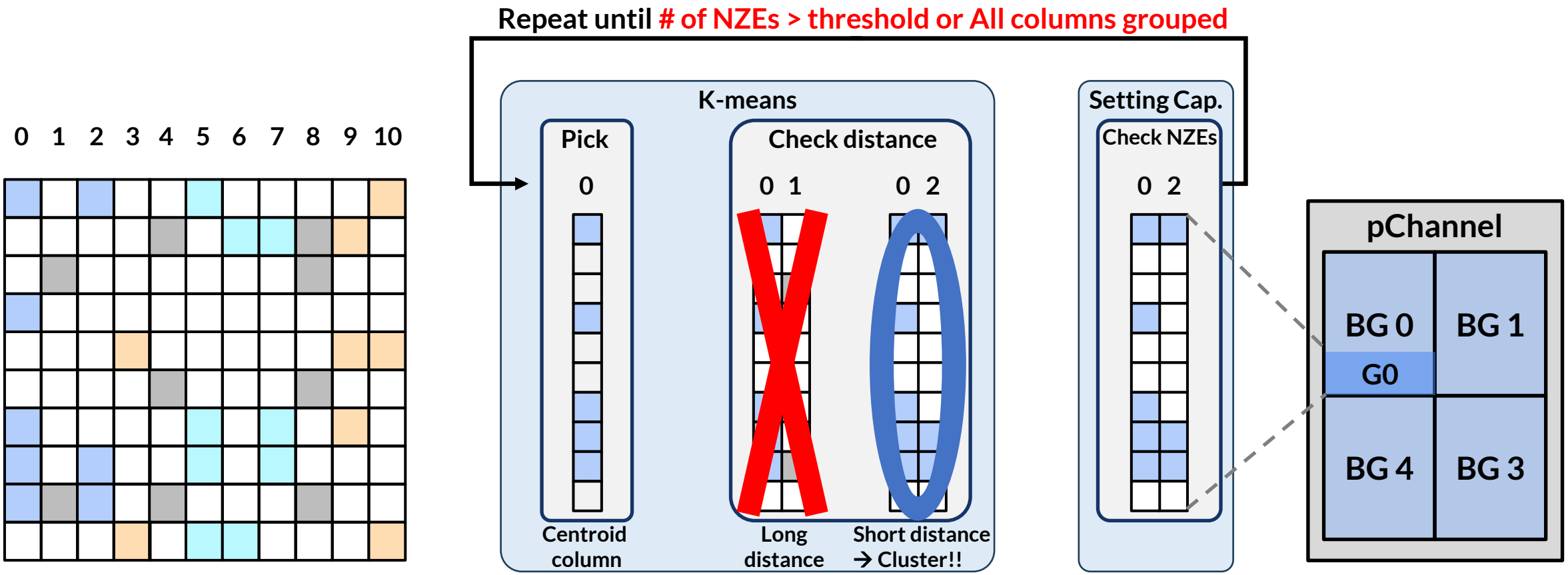
DRAF: DRAM Row Aligned Format



DRAM row aligned format

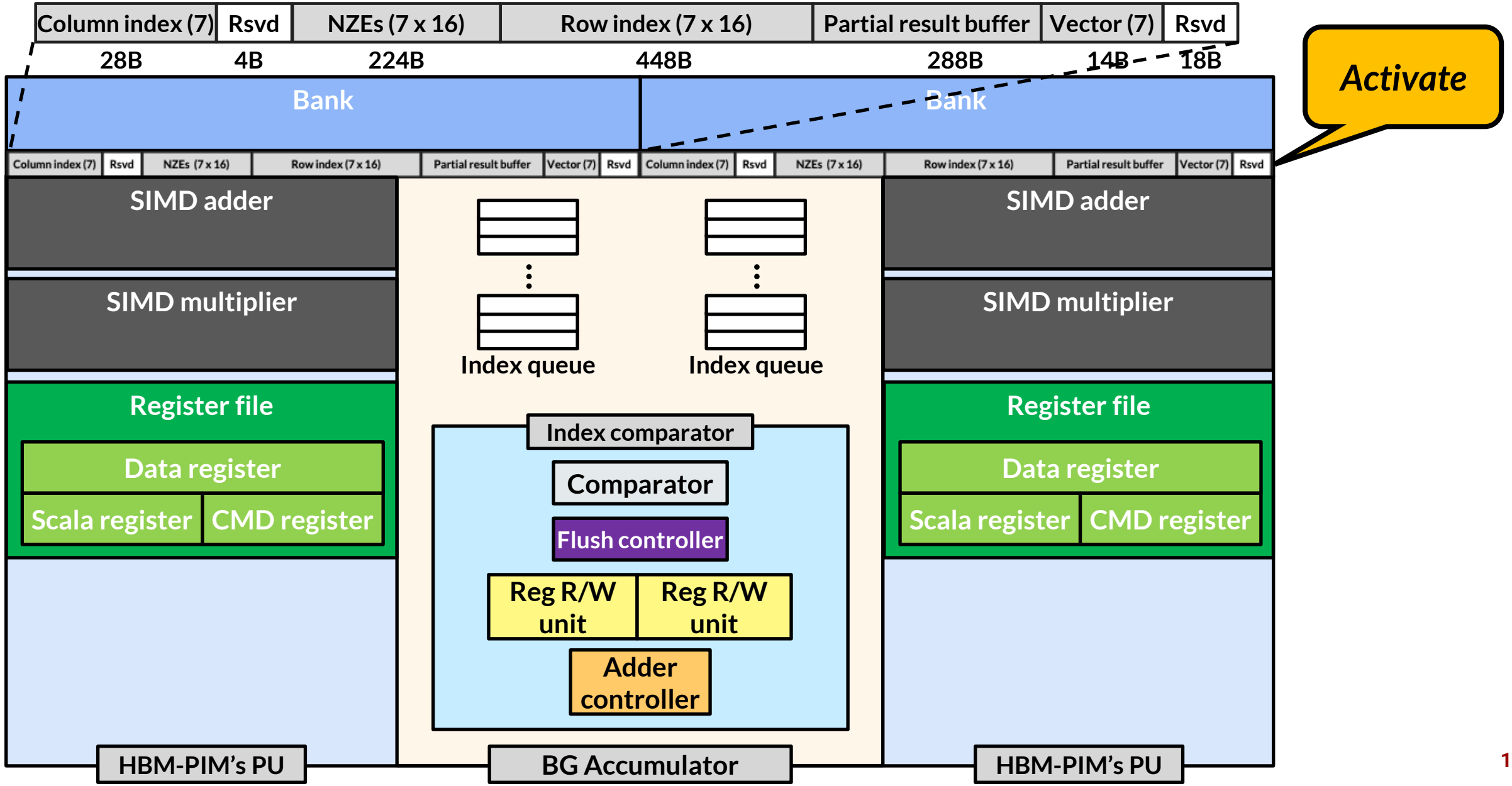
- **Single index represents 16 values**
- **Include reserved area for partial results**
- **Fit column access granularity**

SW Optimization: Bounded-cap K-means

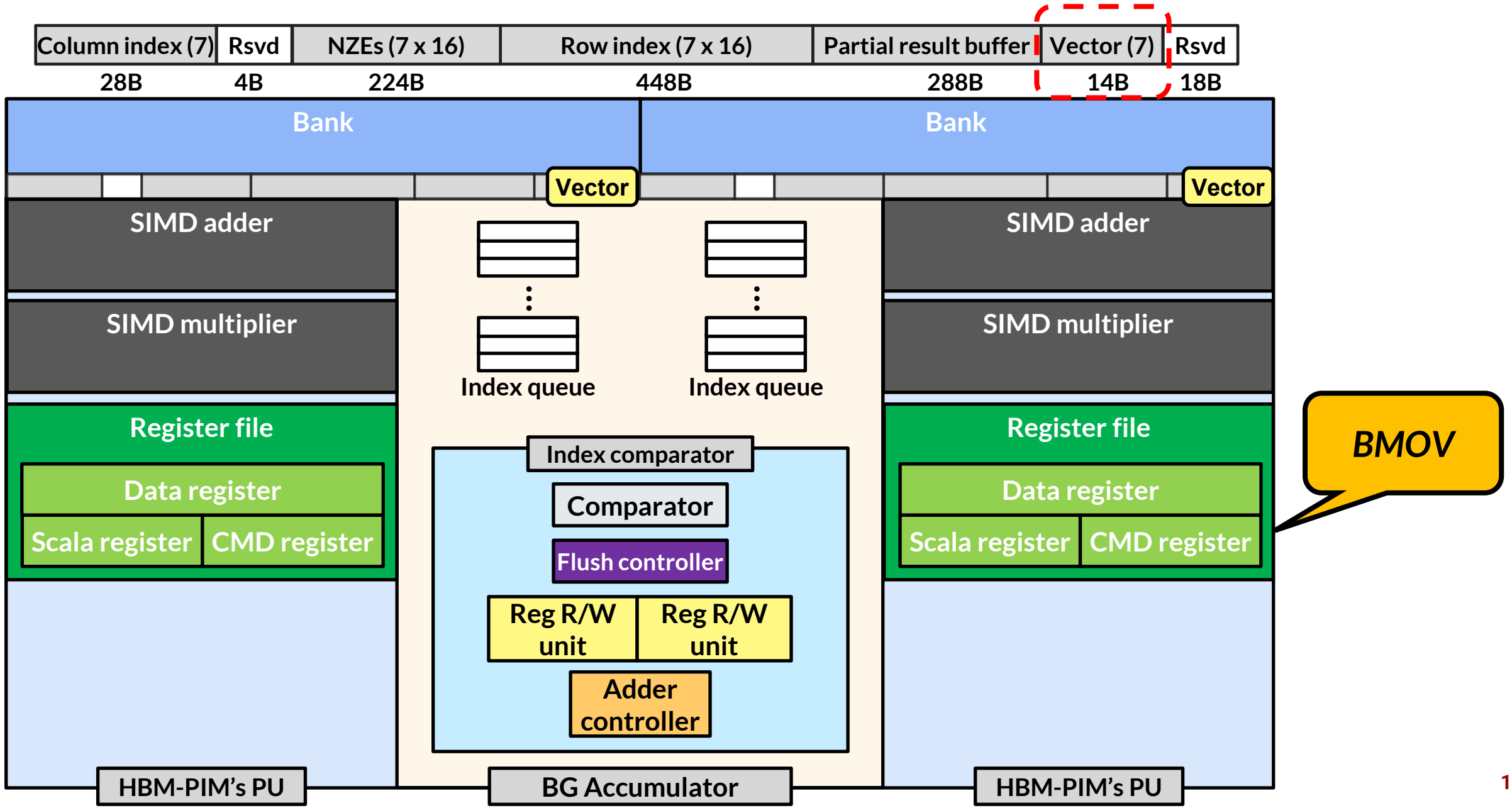


See paper for additional refinement step

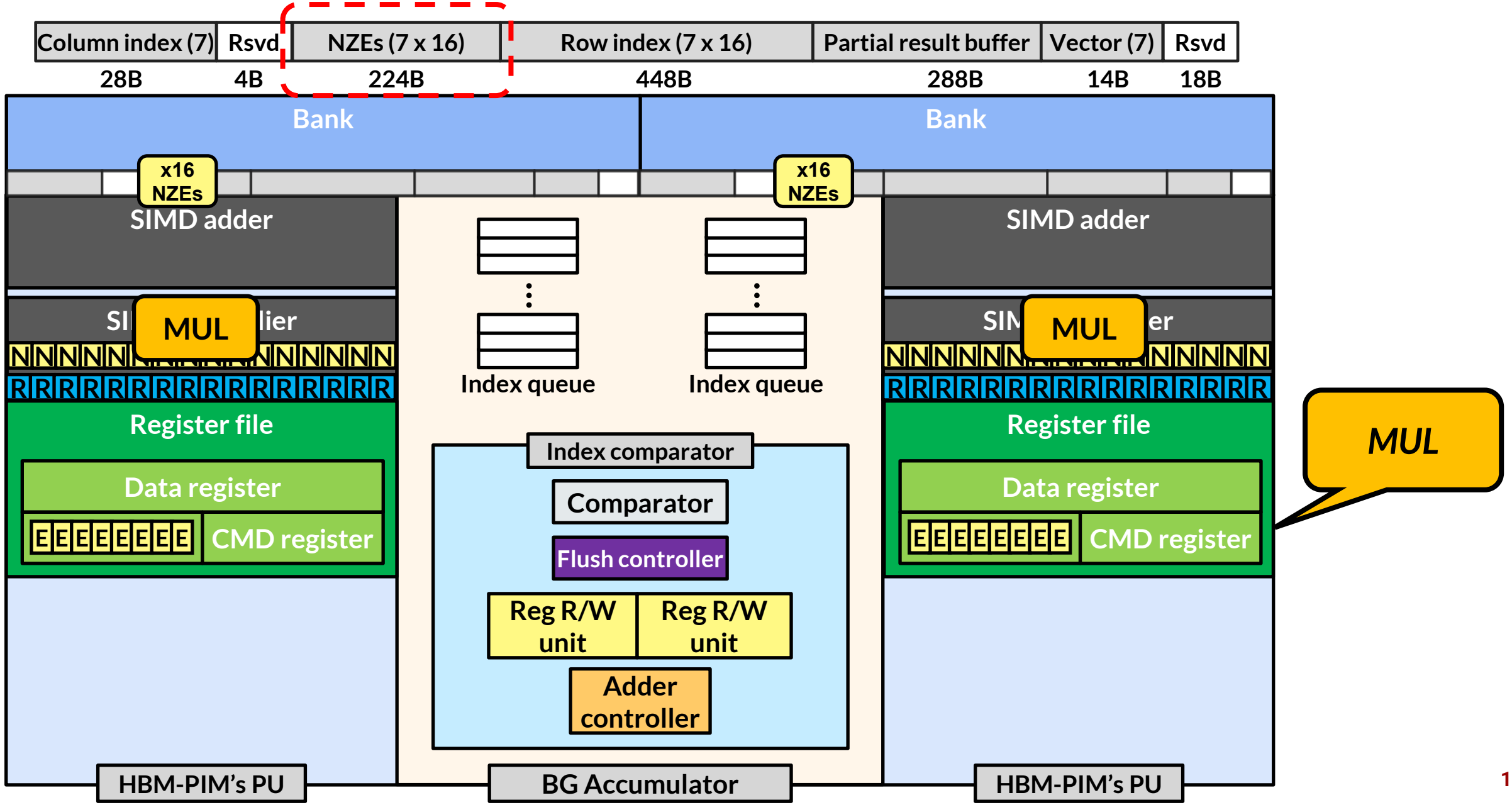
Execution Flow & Architecture



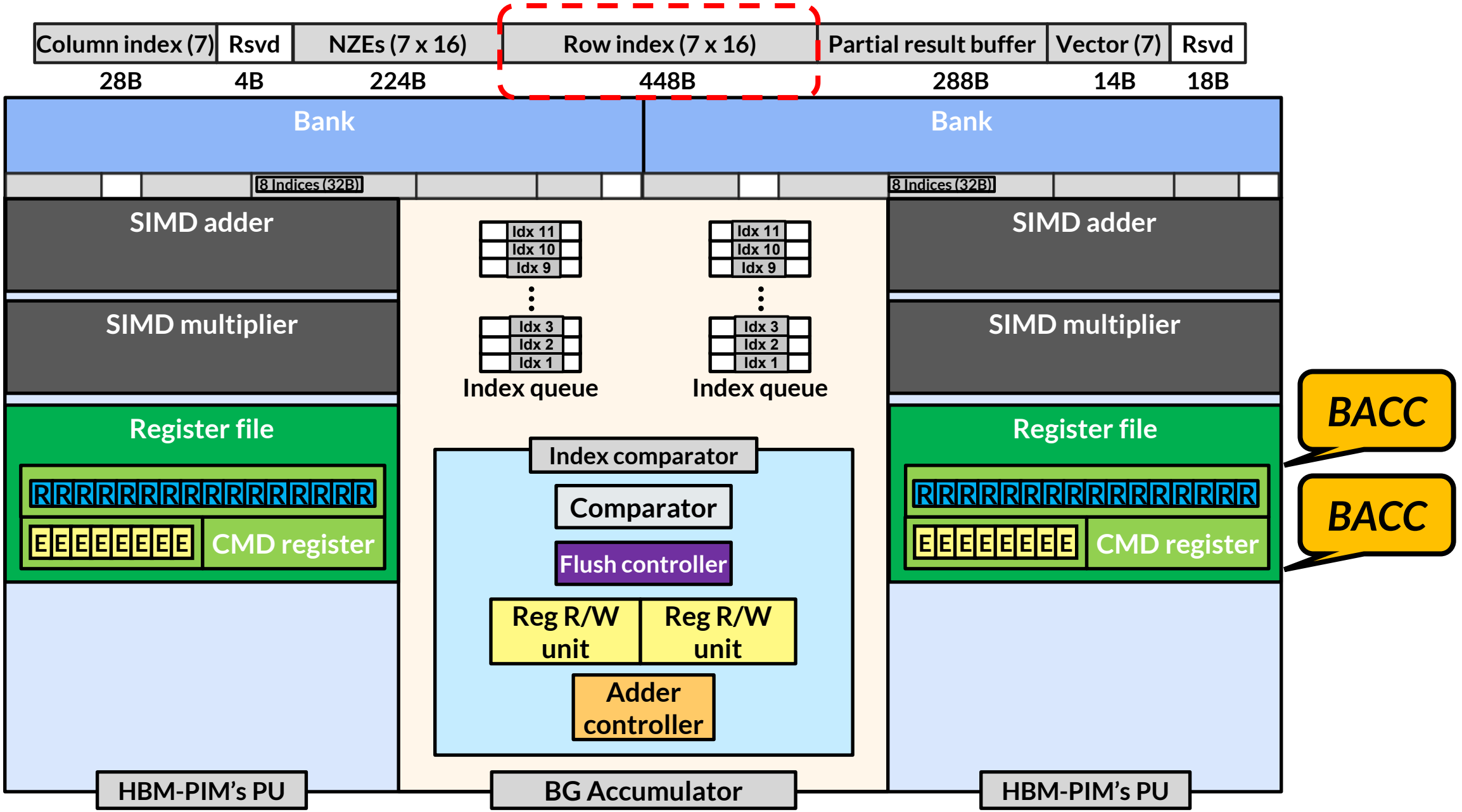
Execution Flow & Architecture



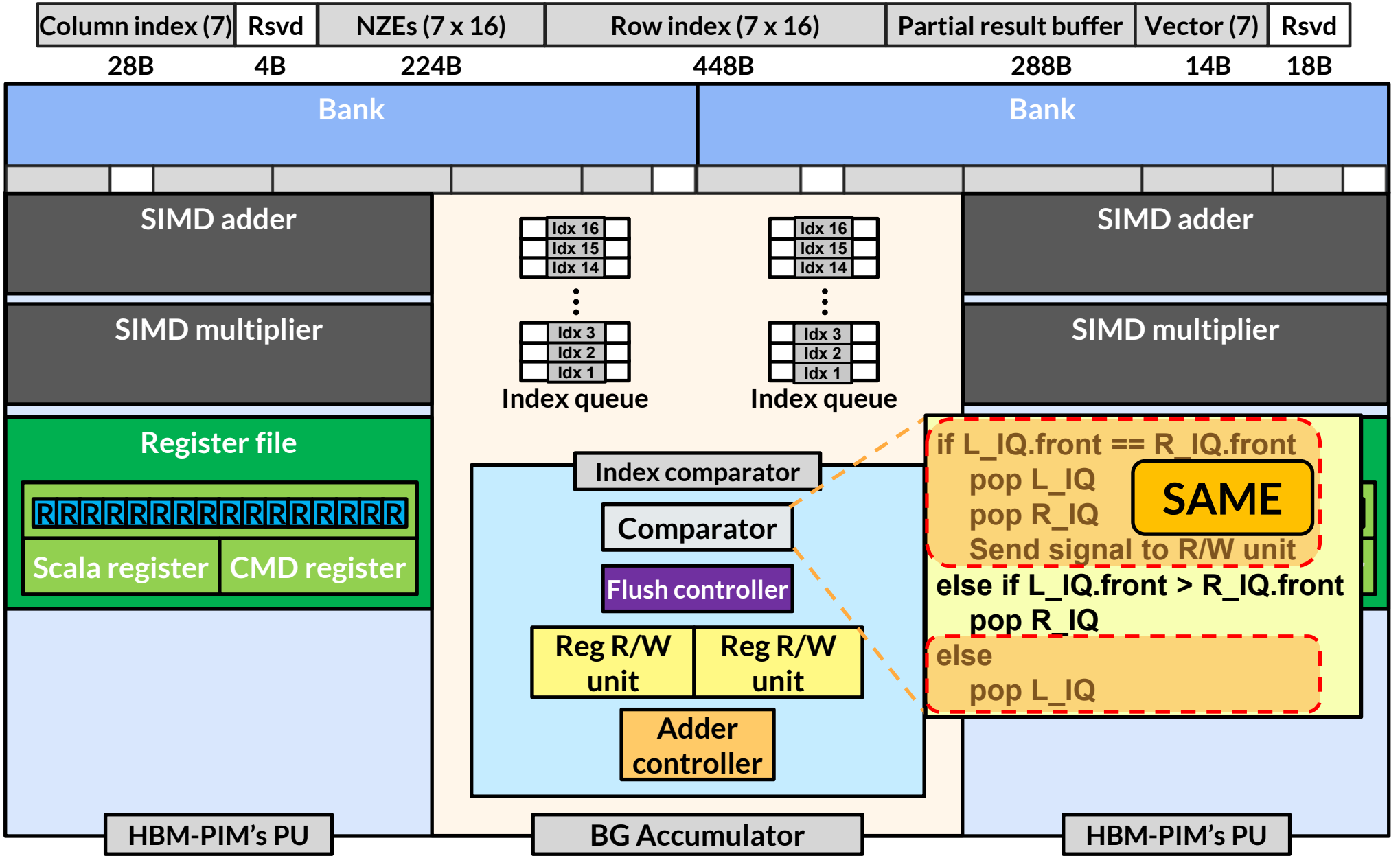
Execution Flow & Architecture



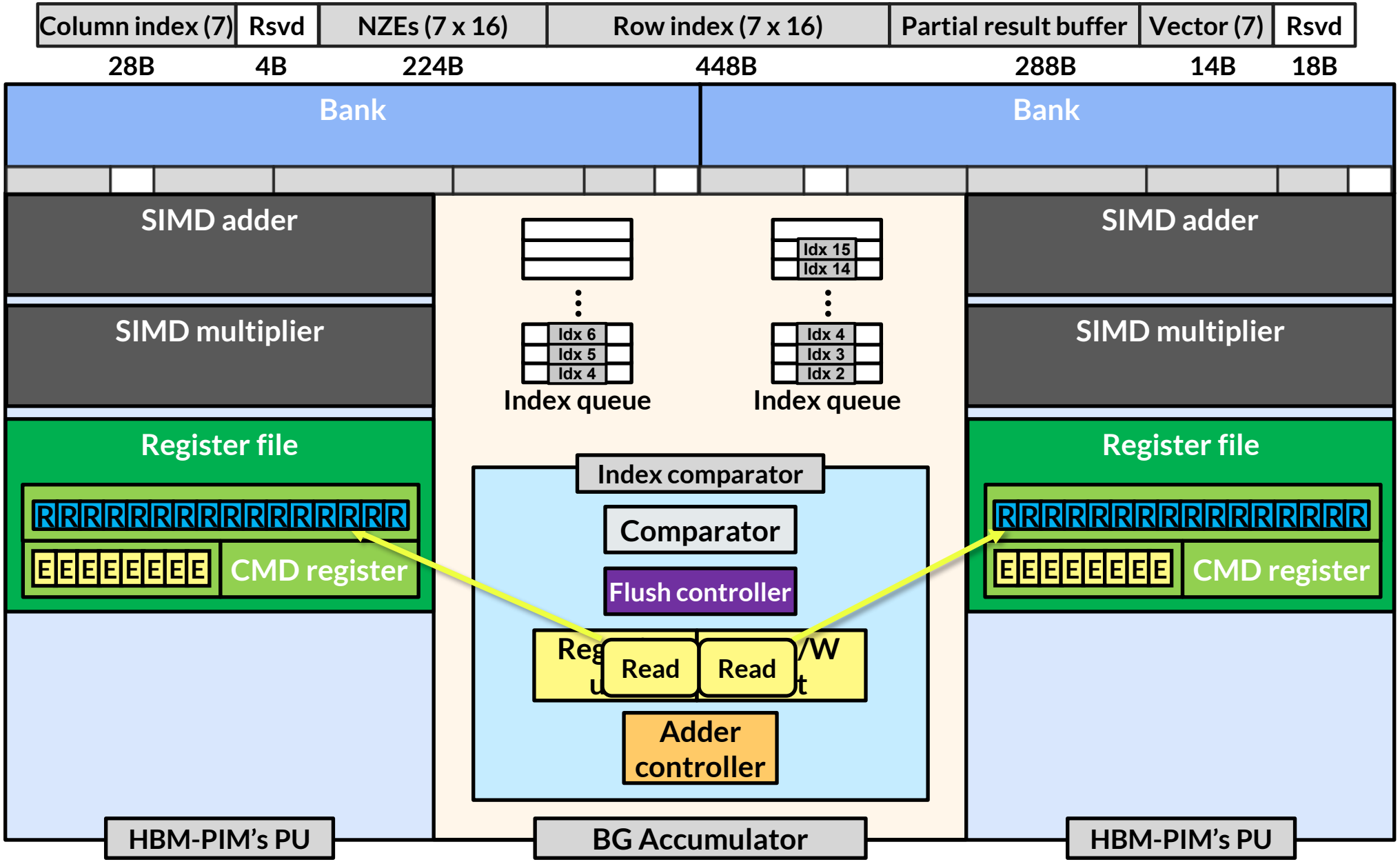
Execution Flow & Architecture



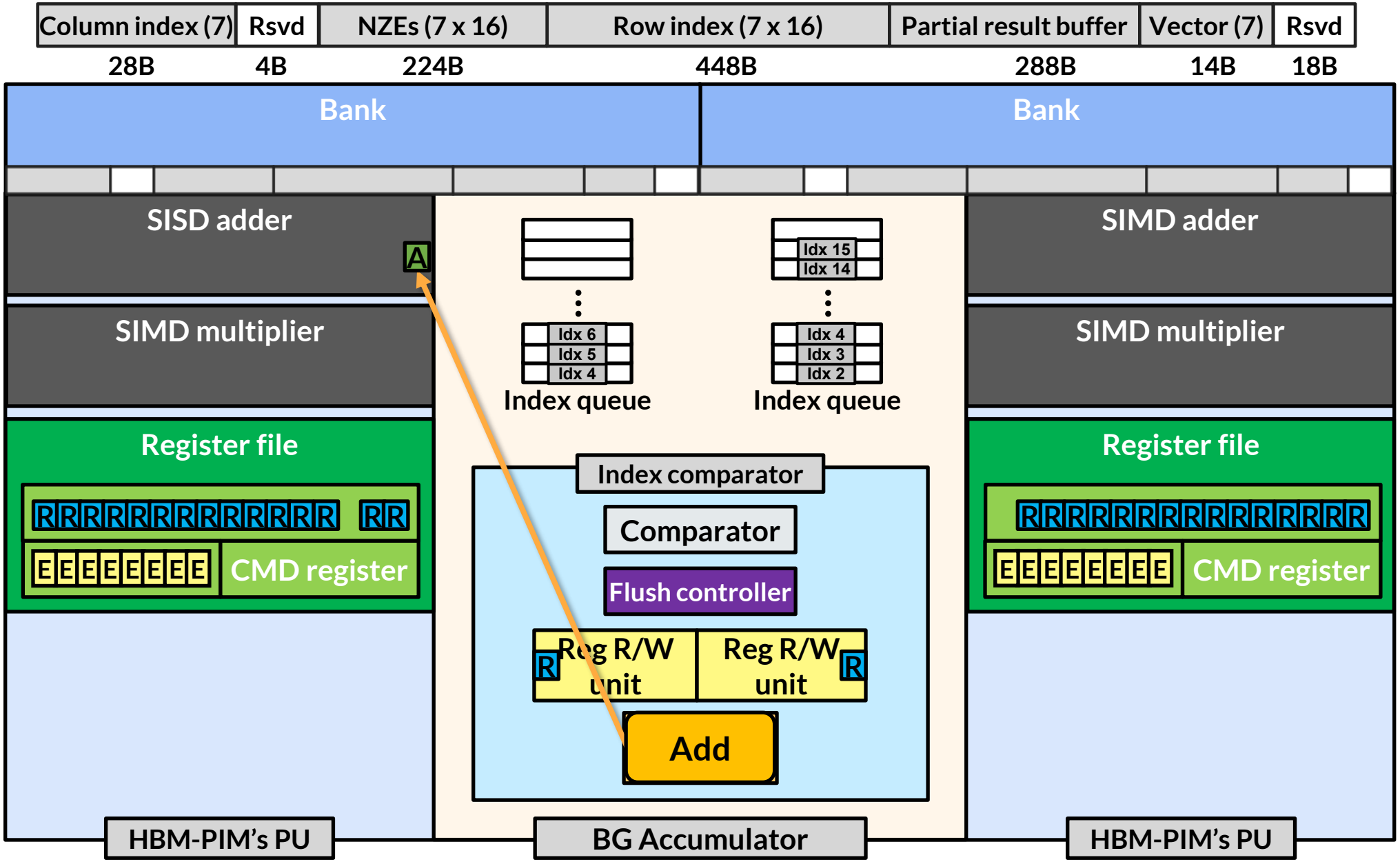
Execution Flow & Architecture



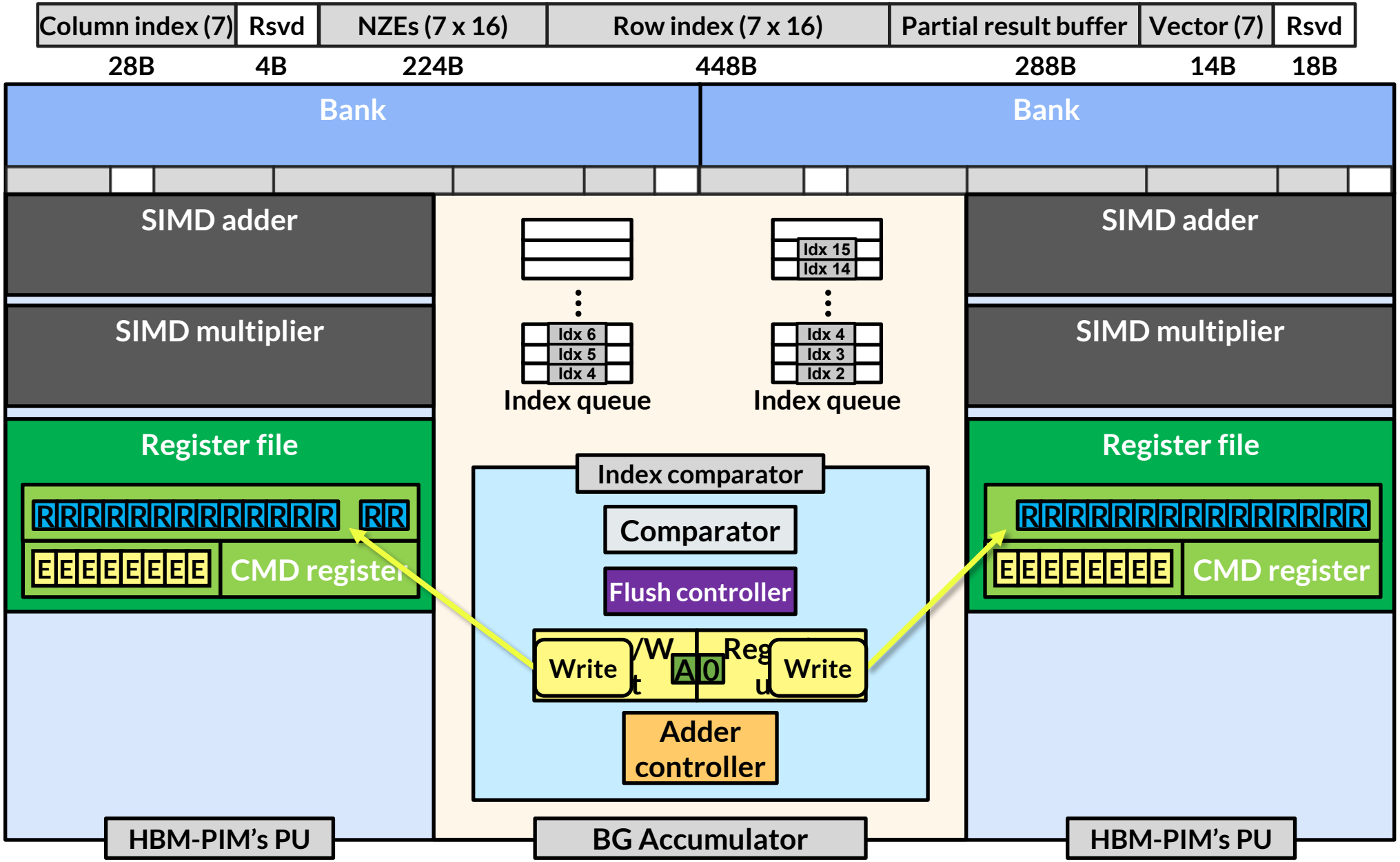
Execution Flow & Architecture



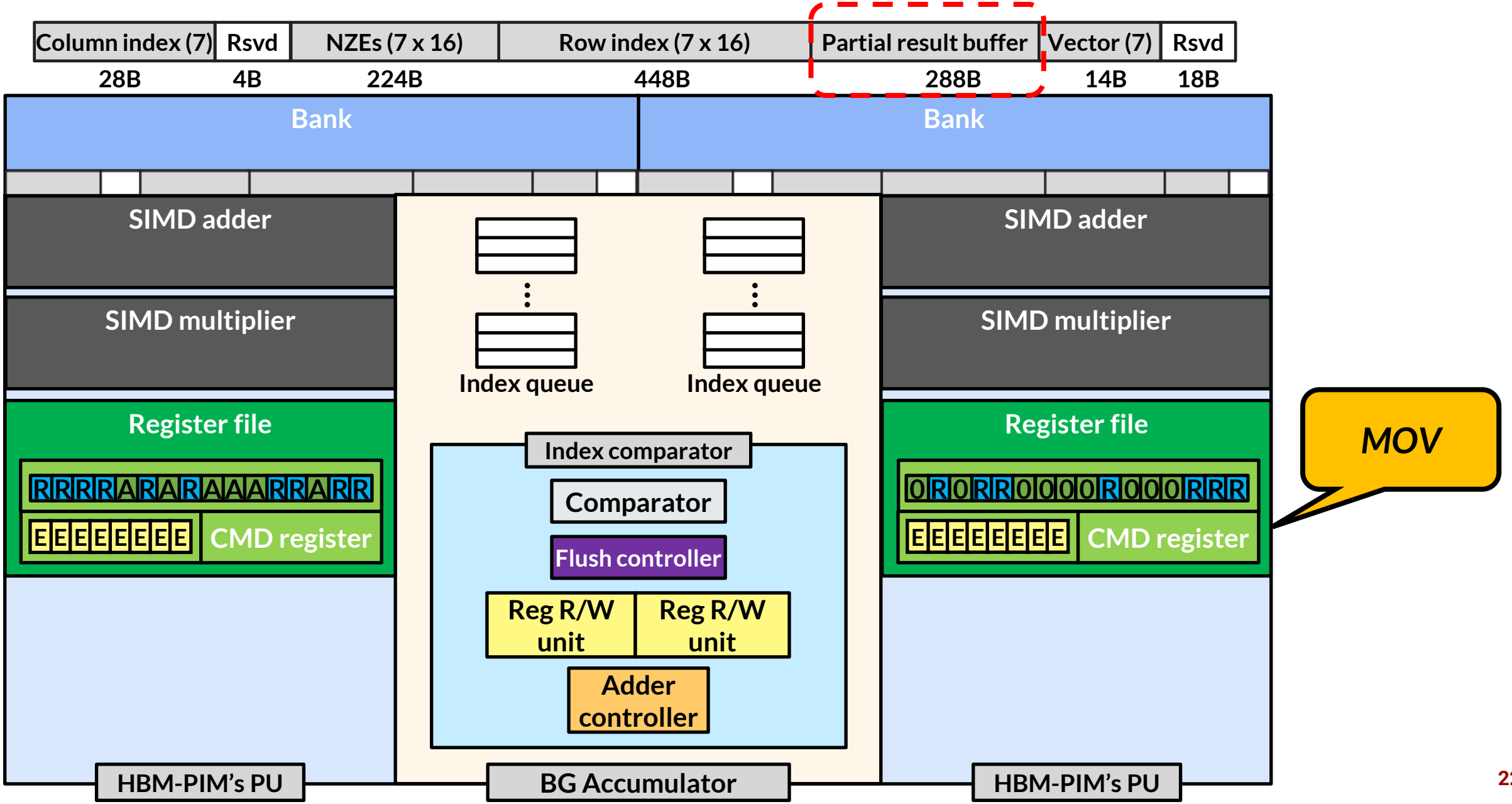
Execution Flow & Architecture



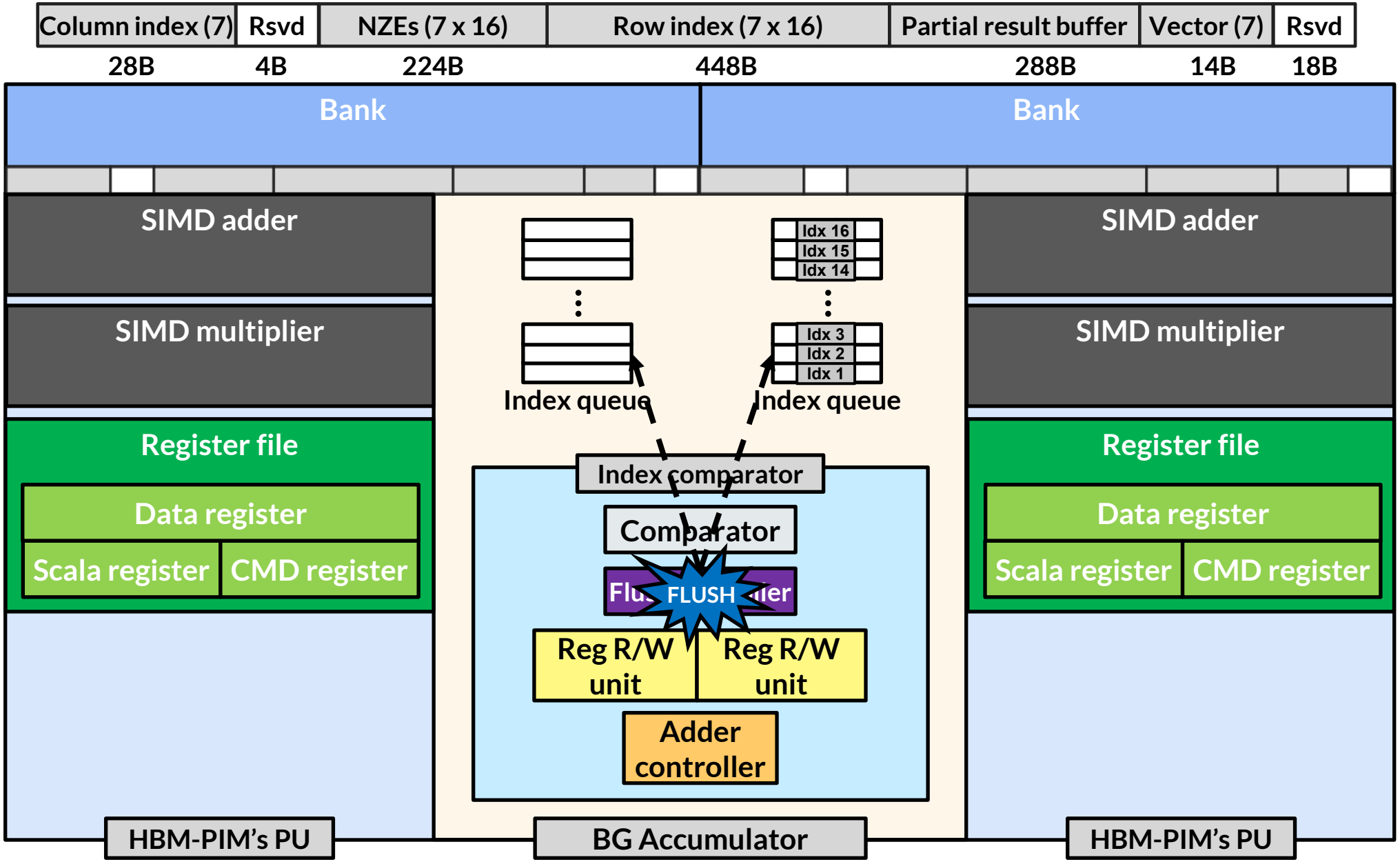
Execution Flow & Architecture



Execution Flow & Architecture



Execution Flow & Architecture



Evaluation Methodology

Evaluation methodology
- Modified DRAMsim3 to support SparsePIM

Baseline
- NVIDIA GeForce RTX 3090 + cuSPARSE

Benchmark
- 16 Sparse matrix workloads from SuiteSparse Matrix Collection

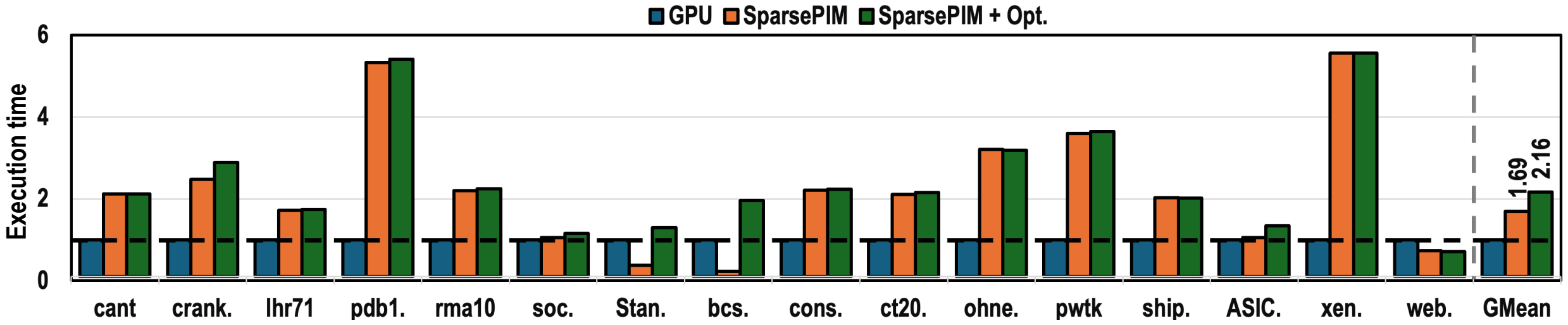
Field	Config
Memory type	HBM2
No. of pseudo-channels	16
No. of BG	4
No. of banks per BG	4
No. of DRAM rows	16,384
No. of DRAM columns	128
Clock frequency	1 GHz
Timing parameter	Default DRAMsim3 HBM2 timing

Simulator configuration

Field	Config
Data register / 2 banks	512 B
Scalar register / 2 banks	32 B
Command register / 2 banks	128 B
Index queue	4 B x 16 entry
Total No. of SIMD FPU	128
Total No. of BG accumulator	64

Hardware configuration

Result: Performance



Performance

- Without optimization: average 1.69x faster than GPU
- With optimization: average 2.16x faster than GPU
- Up to 5.61x faster than GPU

* Benefits of DRAF & Software optimization are details in the paper

Result: Power & Area

Field	Result
Area	0.00275 mm^2
Gate count	8853
Compared to HBM-PIM's PIM logic	≈ 2.21% area overhead
Dynamic power	31.8530 uW (less 8% than SIMD multiplier)

Power & Area

- Design Compiler, SAED 14nm FinFET process
- 2.21% area overhead compared to the HBM-PIM design
- Power constraint compare with SIMD multiplier

Conclusion

SpMV challenges

- Irregular/Indirect memory accesses
- Conventional PIMs support

SparsePIM

- Bank group accumulator (BGA)
- DRAM row aligned sparse matrix compression format (DRAF)
- Bounded cap K-means clustering

Key result

- Up to 5.61x speedup than RTX3090

Thank you

SparsePIM:
An Efficient HBM-Based PIM Architecture for Sparse Matrix-Vector Multiplications

Taewoon Kang, Geonwoo Choi, Taeweon Suh, Gunjae Koo

✉ taewoon_kang@korea.ac.kr