An Efficient High-quality FHD Super-resolution Mobile Accelerator SoC with Hybrid-precision and Energy-efficient Cache

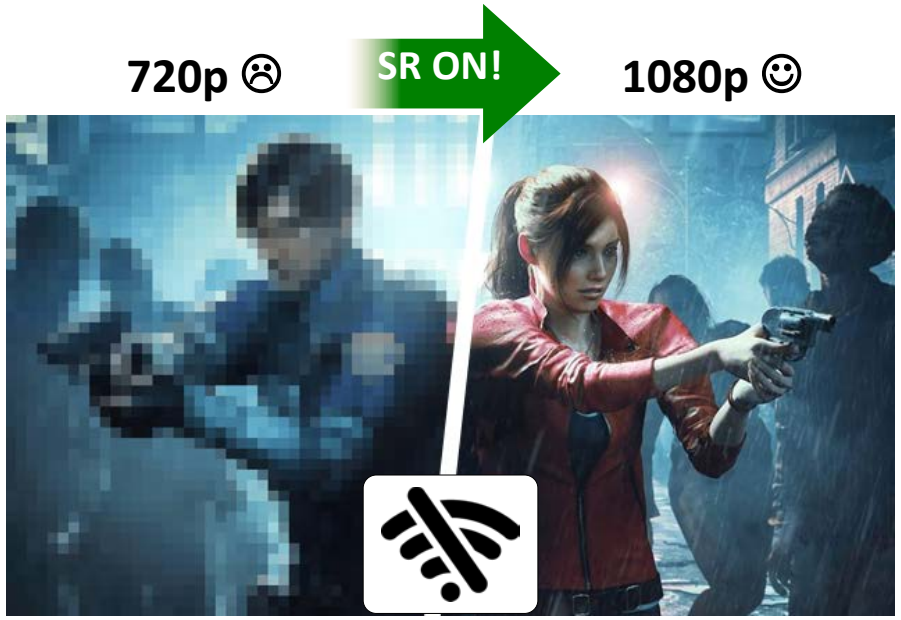
Zhiyong Li, Sangjin Kim,
Dongseok Im, Donghyeon Han, and Hoi-Jun Yoo

zhiyong_li@kaist.ac.kr

Semiconductor System Lab.
School of EE, KAIST

Super-Resolution on Mobile Platform

- **Improve User Experience with High Quality Images**
  - **Expansion of image feature channels and maintain higher image resolution**
  - **Enhance quality of streaming media/camera shot for better QoS***



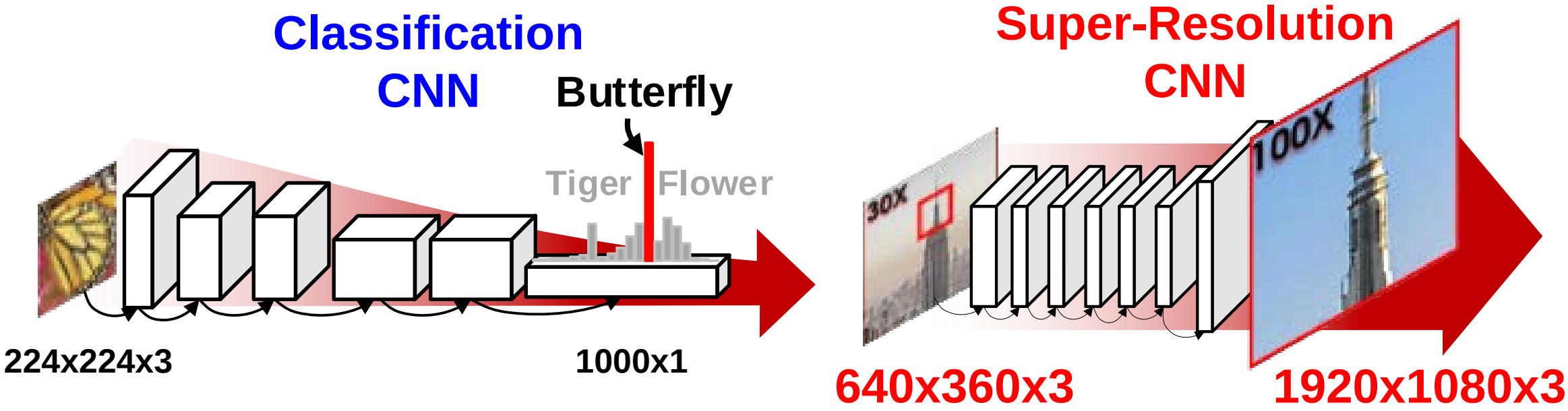
< Streaming/Video Quality Enhancement >

< Camera + SR Zoom Shot >

*QoS: Quality of Service

Characteristic of SR CNN

- **Large and Non-zero Feature Resolution**
  - **Maintaining or enlarging feature resolution**
  - **9x input image resolution**
  - **Non-ReLu act. func. remove sparsity→ zero skipping is impossible ☹**

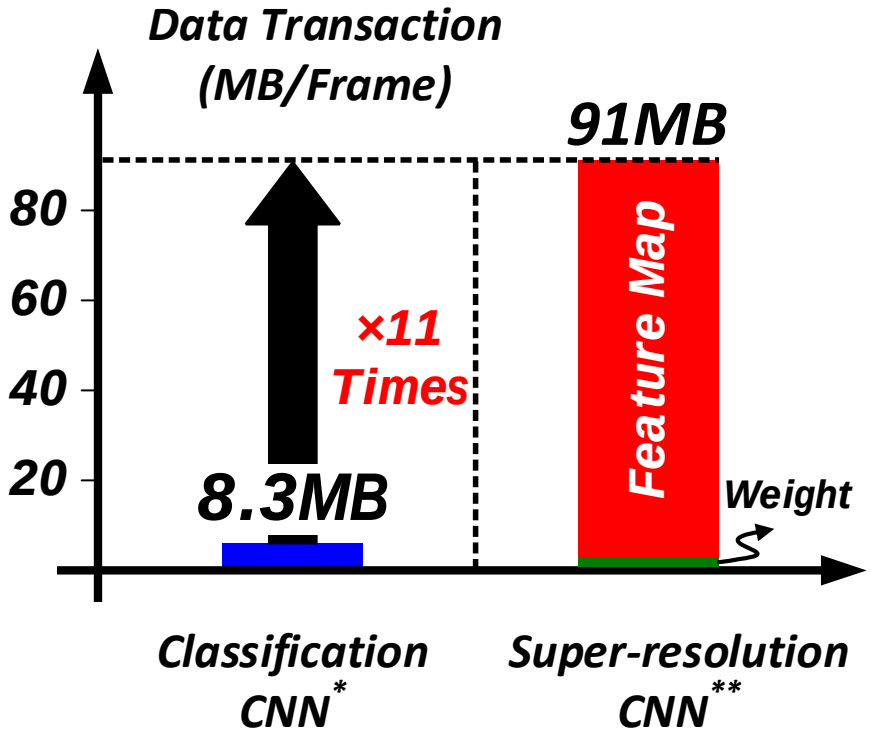


< Comparison of Classification* and Super-Resolution Networks >**

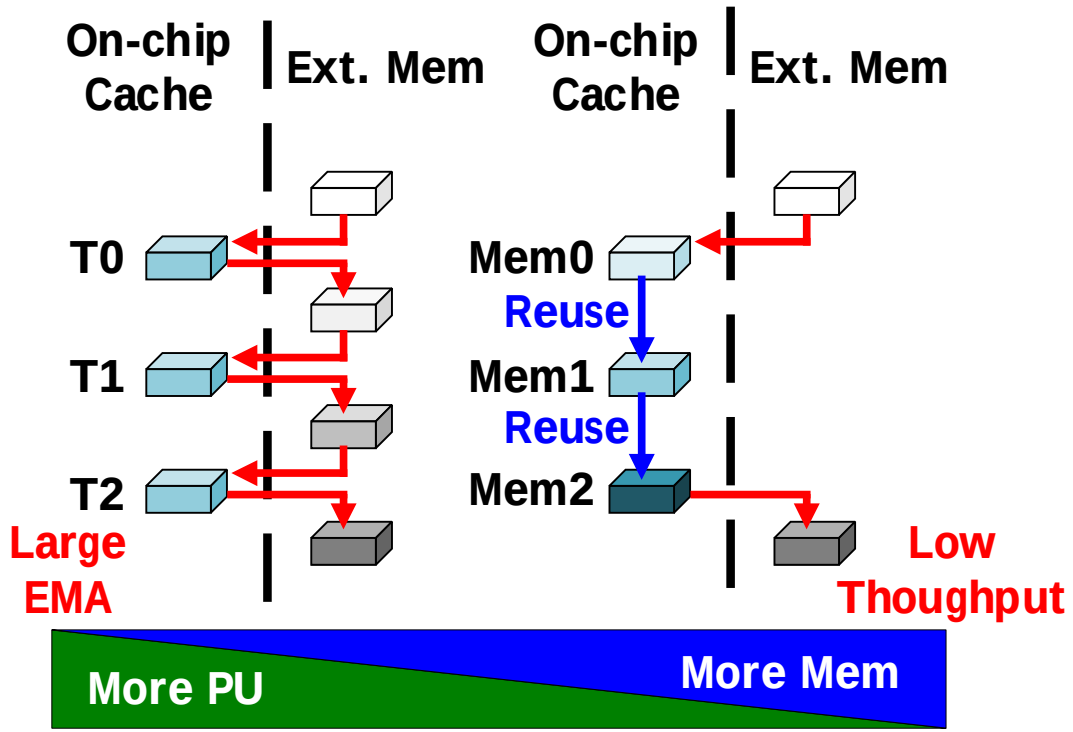
*Measured @ VGG-16, **Measured @ FSRCNN (Set 5)

High Memory Requirement

- **Large Data Transactions**
  - Total 11x data transaction than classification CNN
  - HW design issue of on-chip cache units and processing units



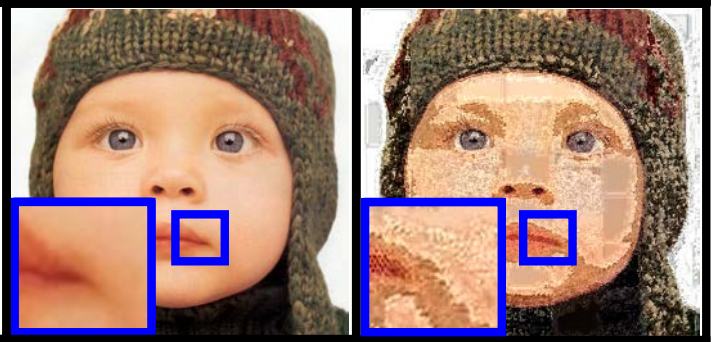
< Comparison of Total Data Transaction >



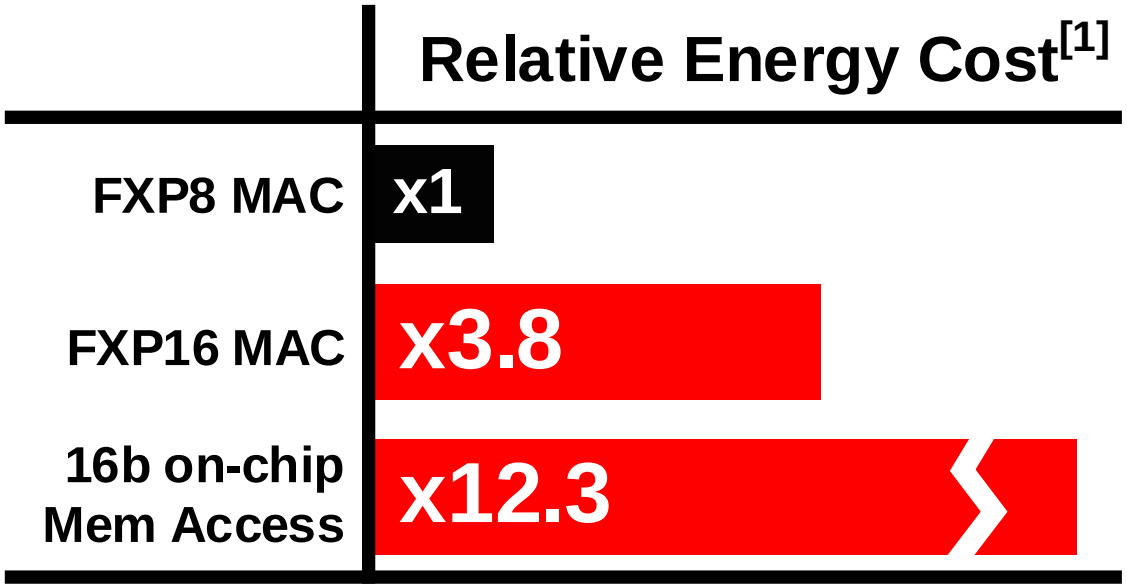
< HW Design Trade-off by Large Data Transaction >

Characteristic of SR CNN

- **Large Bit Precision for QoS**
  - Inefficient high bit data transactions and computations
  - **Efficient Processing algorithm & HW** is needed

Precision	FXP16	FXP8
Results on Set5		
PSNR	38.69 dB	18.85 dB

< SRCNN w/ Different Precision >



< Inefficient High Bit Precision >

[1]: A. Raha et al. VLSID 2021

Features of Proposed SR-SoC

- **2-Part Optimization with Heterogeneous Hybrid-precision**

SW-HW Co-design

Heterogeneous Hybrid-Precision
For Data Compression and Enhance Heterogeneity

Micro-Arch. Level

Process Unit

Heterogeneous Accel. Architecture
For Energy-Efficiency

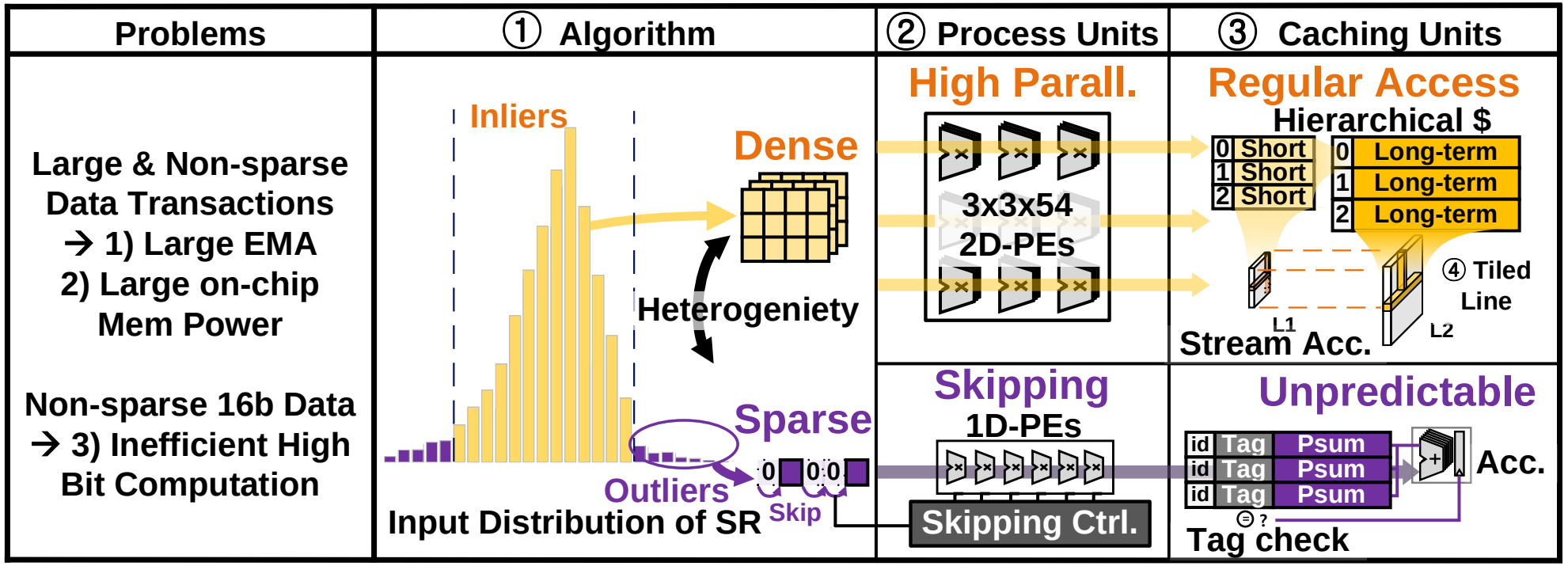
Cache Unit

Heterogeneous L1 Line Cache
For Smaller L1 Cache and low power

Hierarchical Line Cache Subsystem
For Smaller Cache Footprint and Less EMA

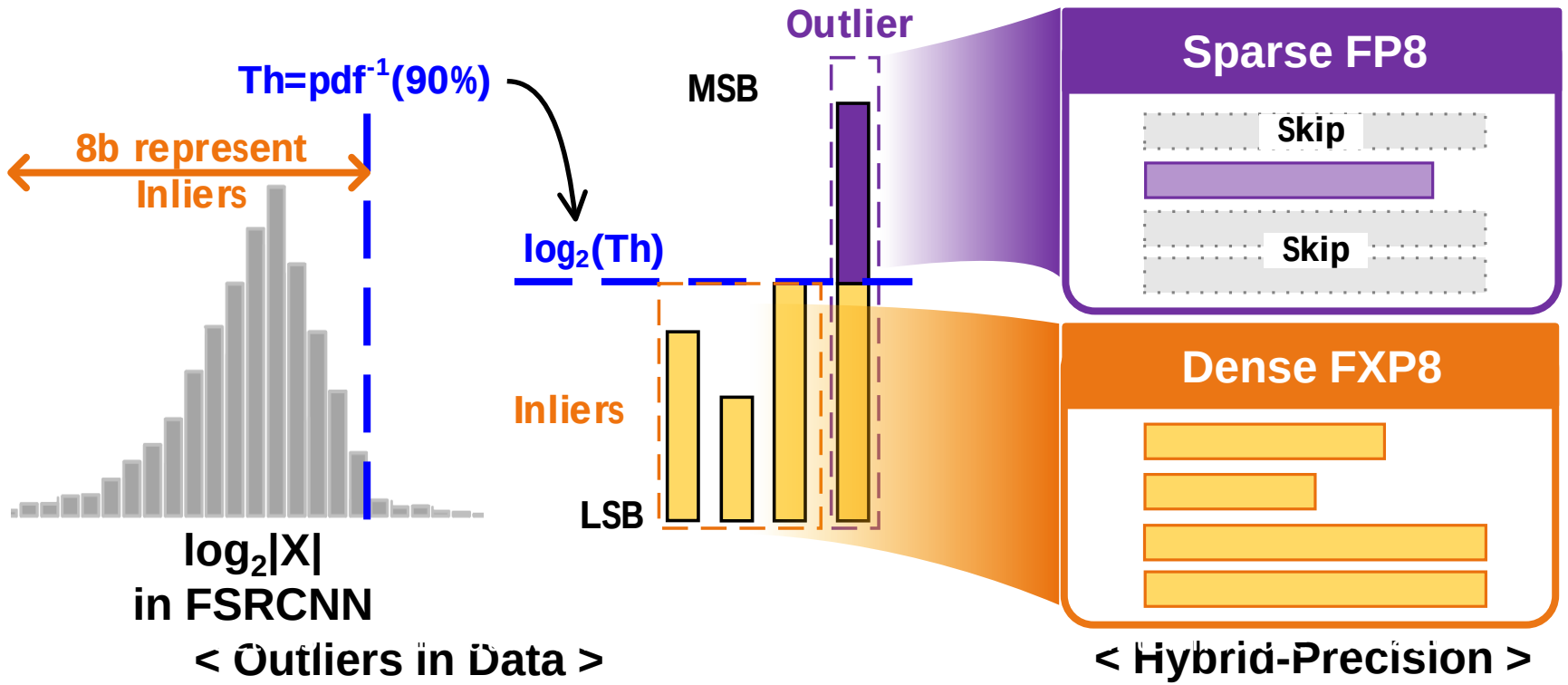
Heterogeneous Hybrid-Precision Architecture

- **SW-HW Co-designed Architecture**
  - ① Heterogeneous Hybrid Precision for less data transaction & sparsity
  - ② High parallelism & Skipping architecture for high computing efficiency
  - ③ Heterogeneous hierarchical cache w/ ④ Tiled exec. for high mem efficiency



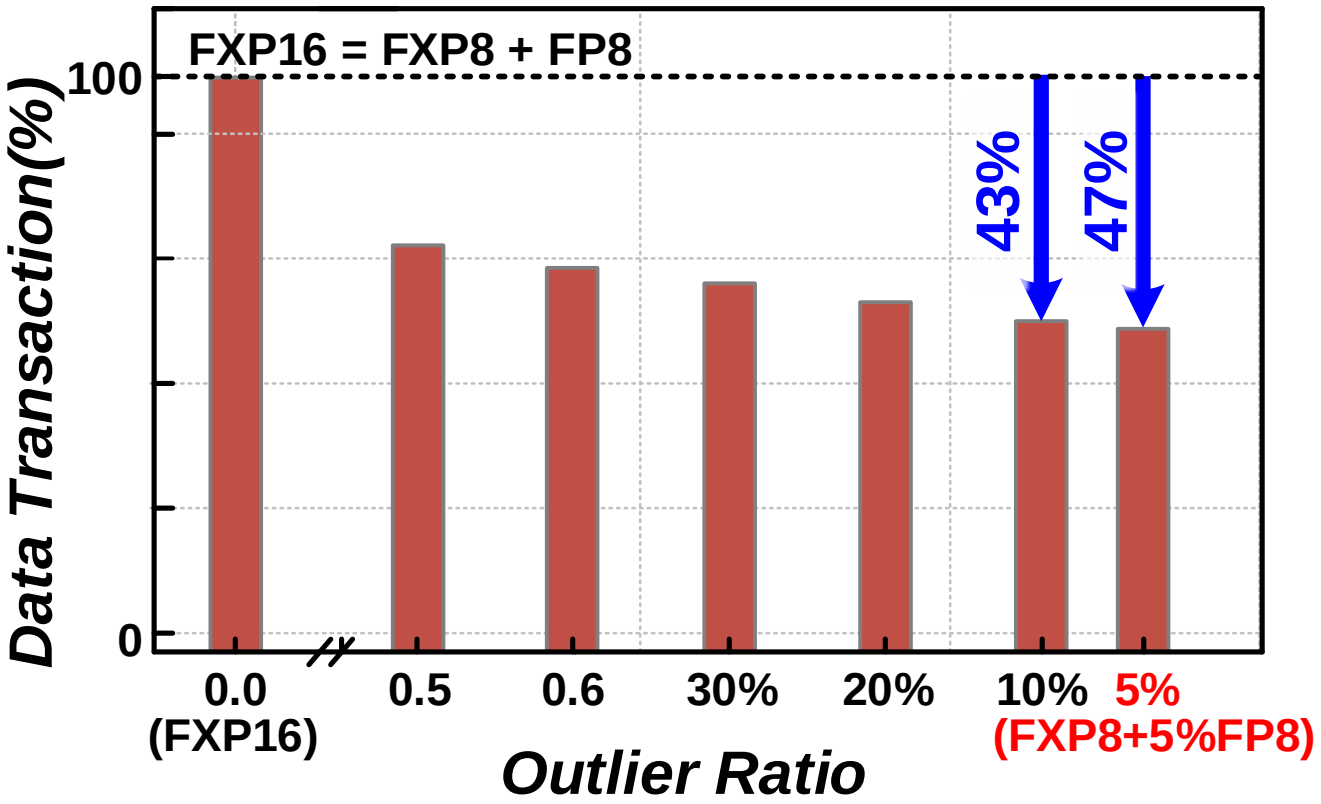
Proposed Heterogeneous Hybrid Precision

- **Divide Data into Sparsity-biased 2 Groups**
  - **Thresholding with Probability Density Function (p.d.f.) @ lower 90%**
  - **100% non-zero dense FXP8 group and 10% non-zero sparse FP8 group**

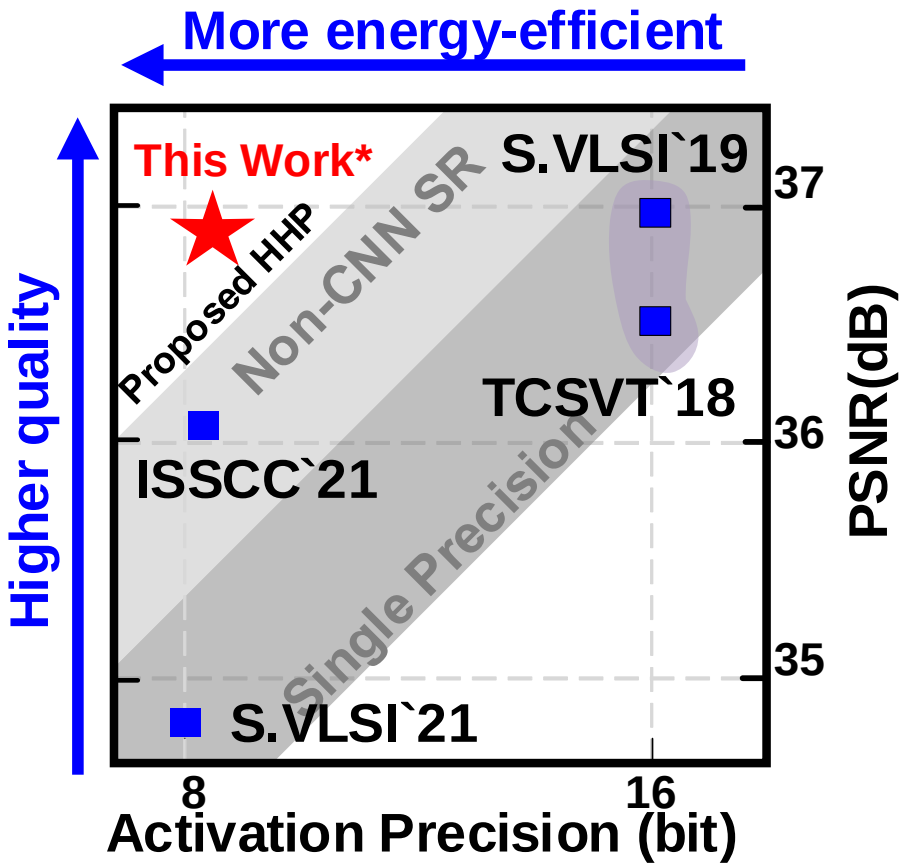


Result of Heterogeneous Hybrid-Precision

- **Maintaining Quality(<0.5dB loss) while Reducing 47% External Memory Access**



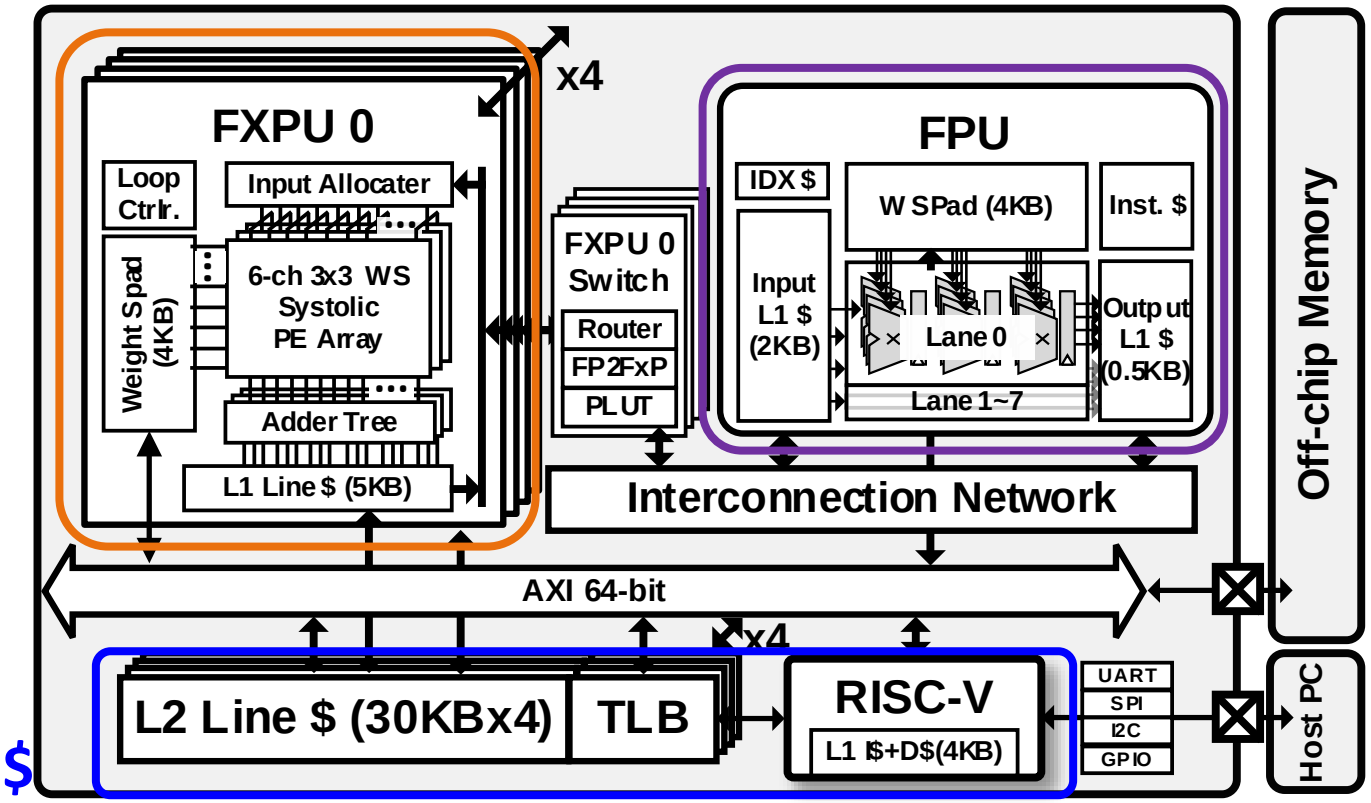
< Data Transitions* vs Grouping Threshold >



< Performance Summary of Precision >

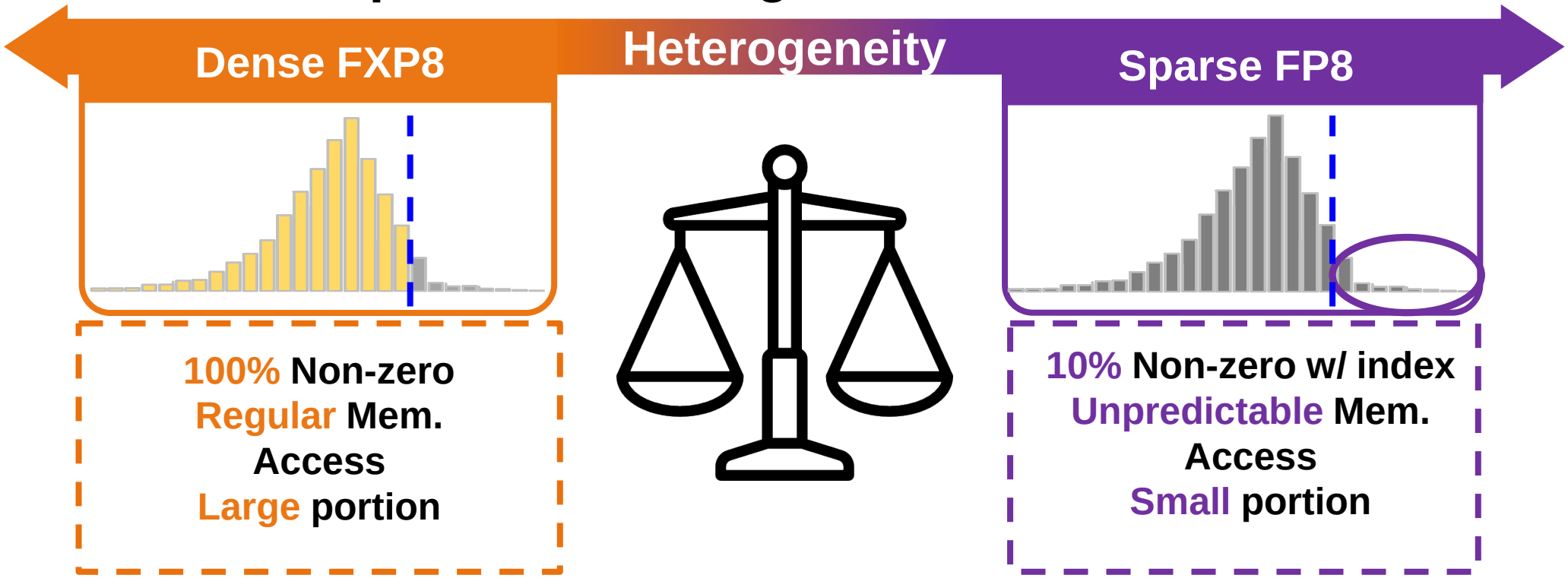
Proposed Super-resolution SoC Architecture

1. FiXed-point Processing Units
   - For Non-zero FXP8 Convolution
   - For Short-term Line $
2. Floating-point Processing Unit
   - For Sparse FP8 Convolution
   - For Short-term Tagged Line $
3. Global L2 Line Cache
   - For top-level threshold biasing
   - For Long-term Overlapped Line $



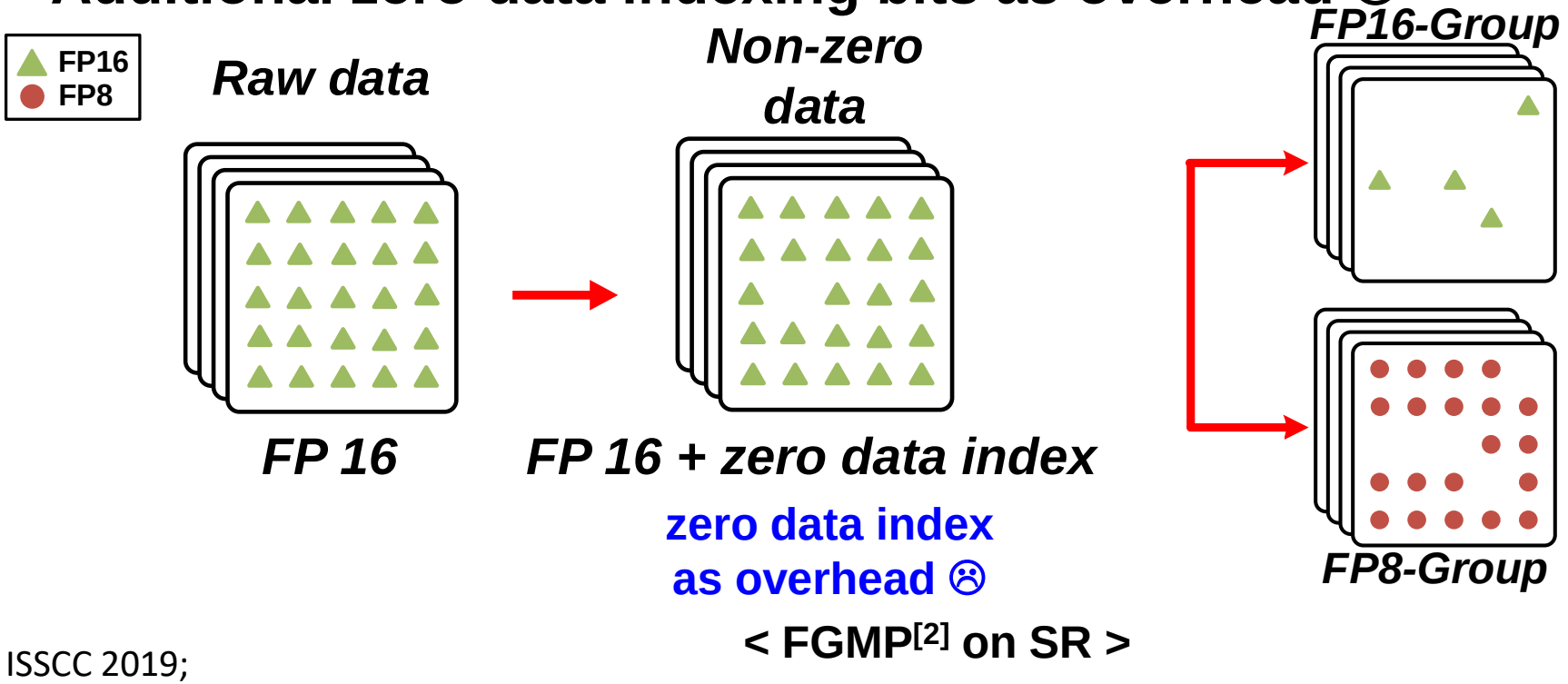
Motivations of Heterogeneous Arch.

- **Sparsity Heterogeneity of HHP Data**
  - **100 % non-zero FXP8 and 10% non-zero FP8 group**
  - **Inefficient in previous homogeneous architecture ☹**



Data Compression for Non-Sparse SR

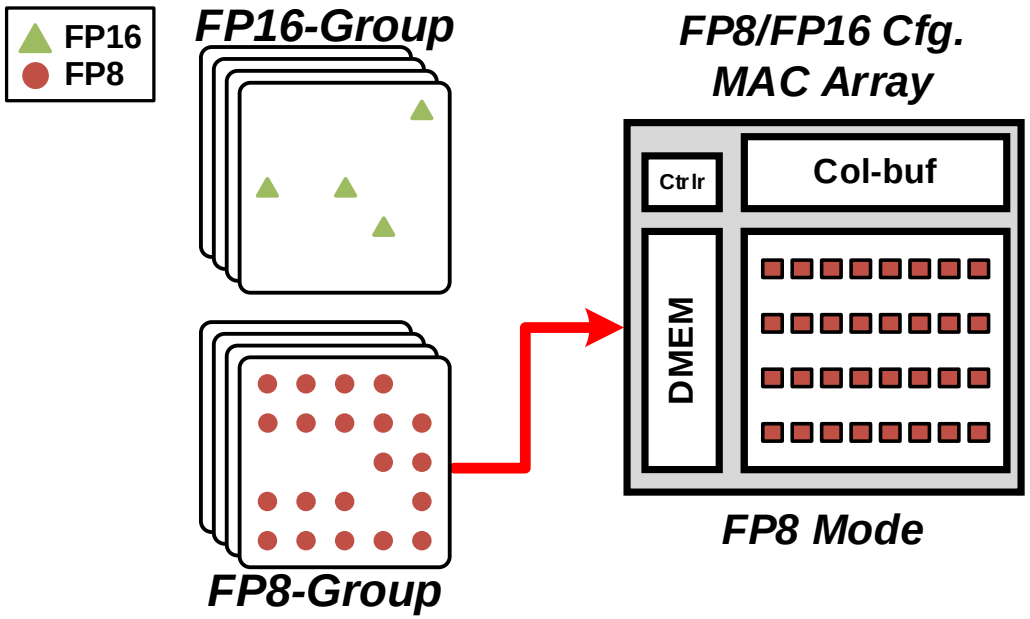
- **Previous Mixed-Precision Hardware**
  - Divide data into 16b large value group and 8b small value group
  - Low compression ratio with non-zero SR data ☹
  - Additional zero data indexing bits as overhead ☹



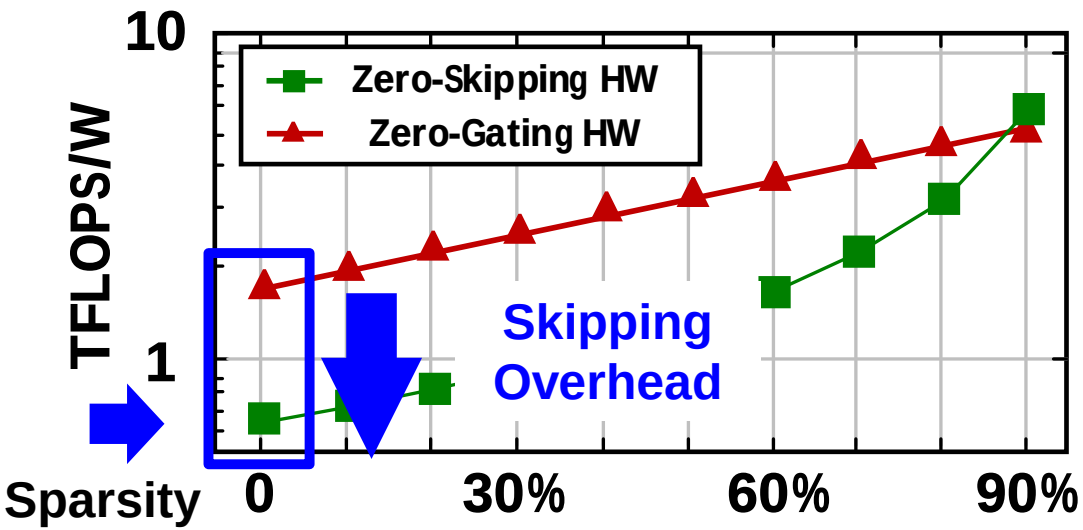
[2] Lee et al. ISSCC 2019;

Previous Mixed-Precision Architecture

- **Homogeneous Mixed-Precision Accelerator**
  - Homogeneous skipping arch. for higher energy-efficiency, BUT
  - Dense low precision group → Skipping Overhead dominant (>90% of SR data)
  - Sparse high precision group → Benefits from small portion (10%) of total data



< Previous Mixed-Precision Arch. >

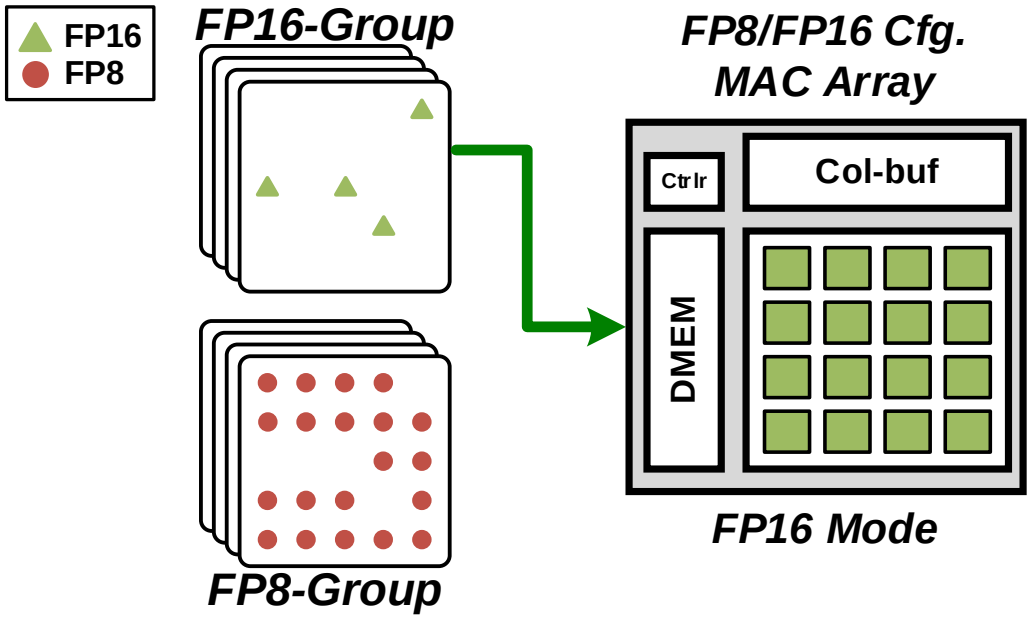


< Energy-efficiency on SR dataset >

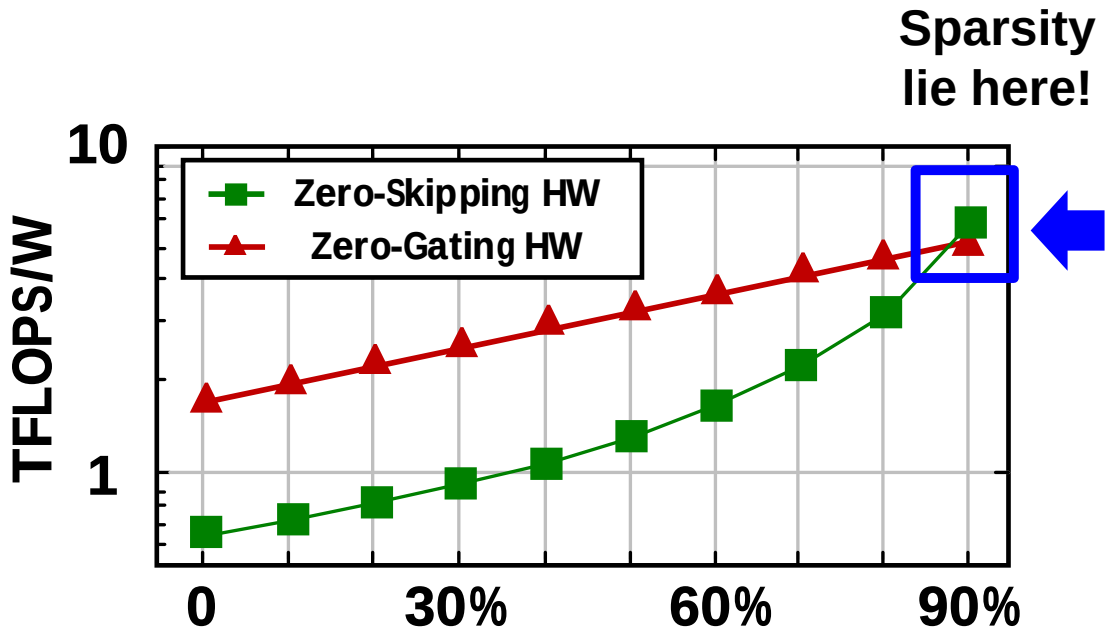
[2] Lee et al. ISSCC 2019;

Previous Mixed-Precision Architecture

- **Homogeneous Mixed-Precision Accelerator**
  - Homogeneous skipping arch. for higher energy-efficiency, BUT
  - Dense low precision group → Skipping Overhead dominant (>90% of SR data)
  - Sparse high precision group → Benefits from small portion (10%) of total data



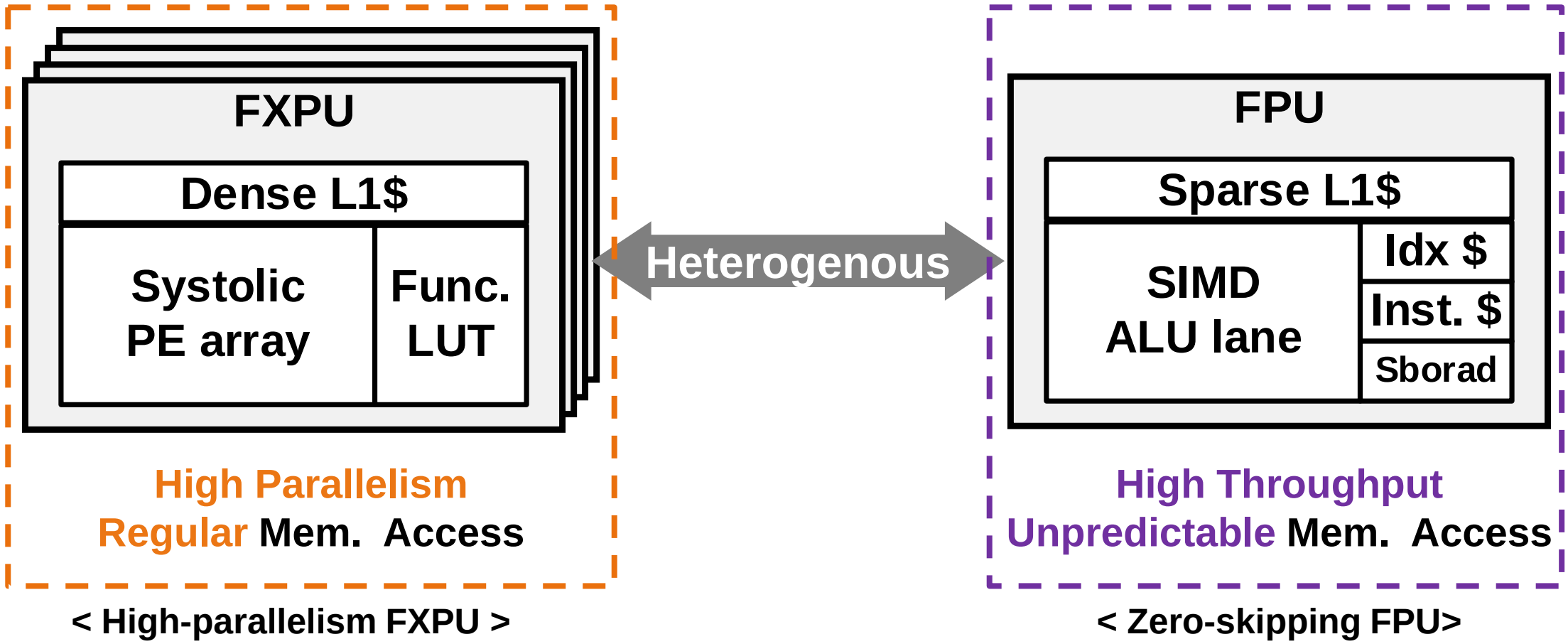
< Previous Mixed-Precision Arch. >



< Energy-efficiency on SR dataset >

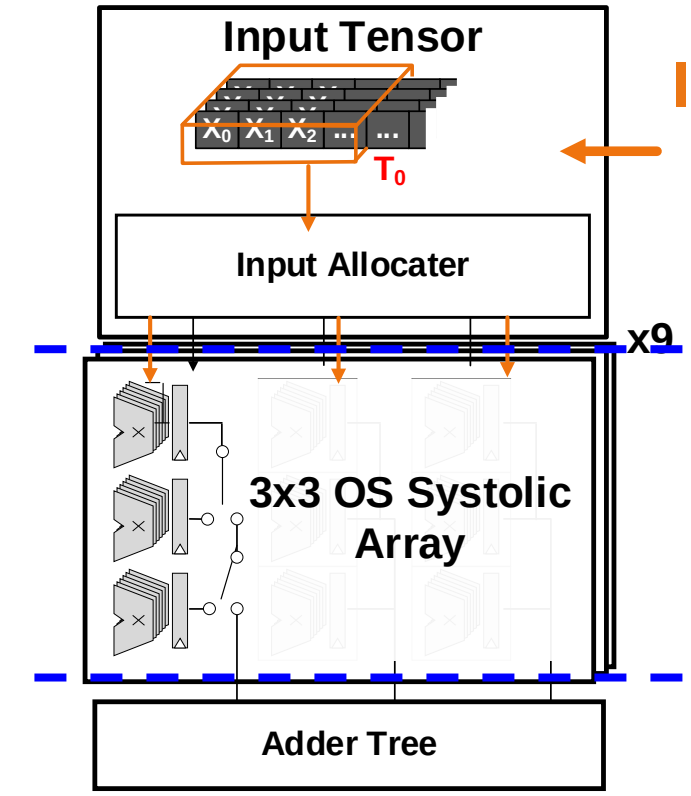
Heterogeneous Accelerating Architecture

- **Efficient Skipping- and Parallelism-exploit Accelerator**



Proposed High Parallelism FXPU

- **High Parallelism FXPU for Diverse Convolution Kernels**



< High Parallelism FXPU Arch.>

Input Allocator w/ Dense L1$:

Receive input data

Reform & feed to PE Array

3x3 2D PE Array:

MAC operation

Output Stationary

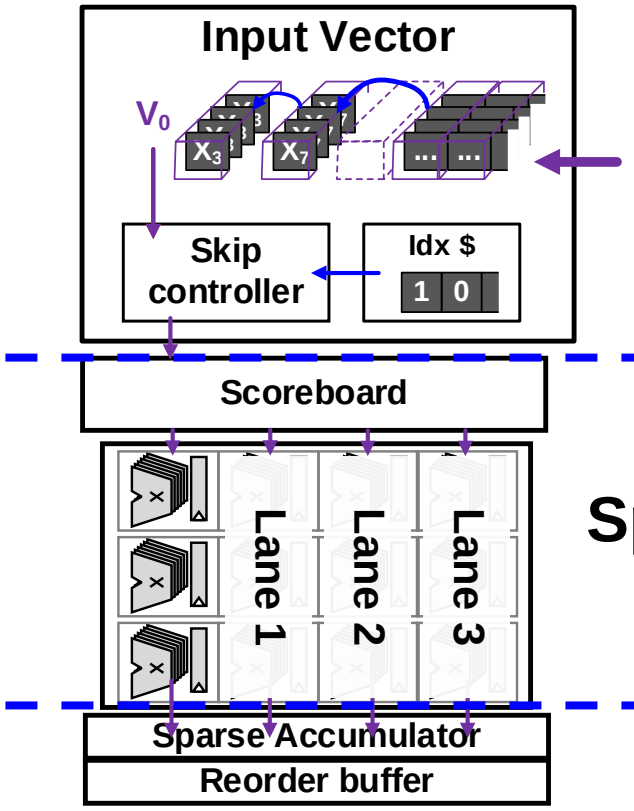
Systolic Data Reuse

Adder Tree w/ LUT

For Activation Func.

Proposed Zero-skipping Architecture

- **Zero-skipping FPU for Sparse Convolutions**



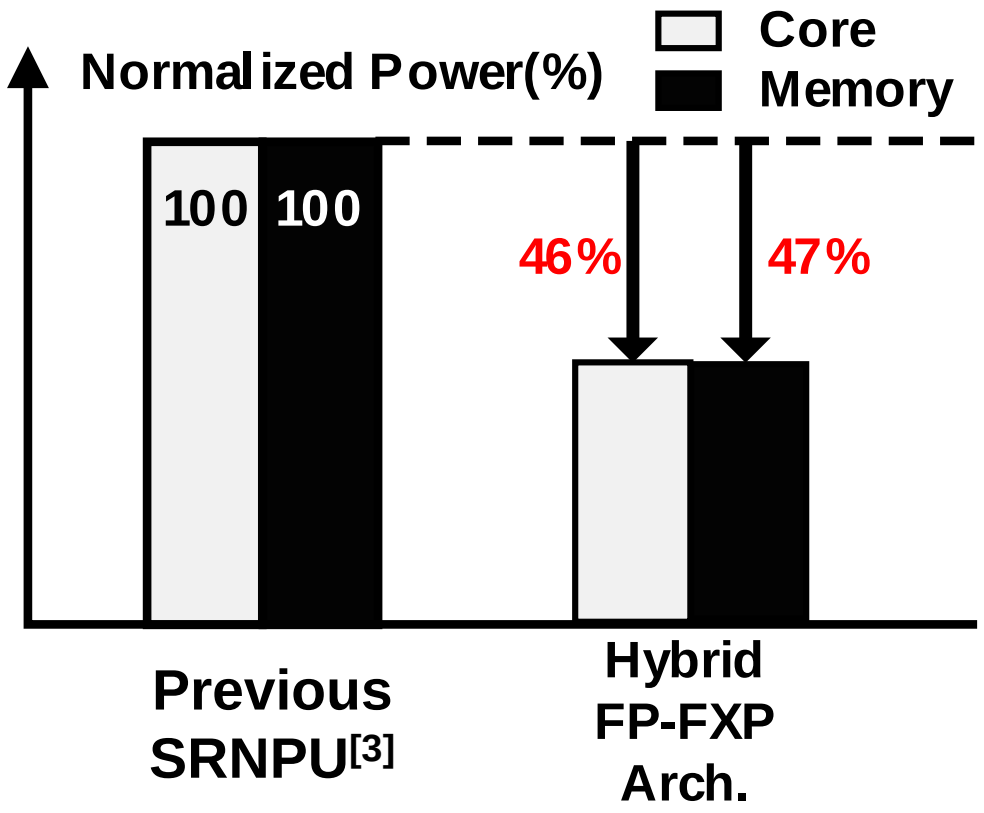
< Zero-skipping FPU Arch. >

Skip Ctrlr. w/ Sparse L1$:
Receive input data & skip index
Select weight w/ zero index

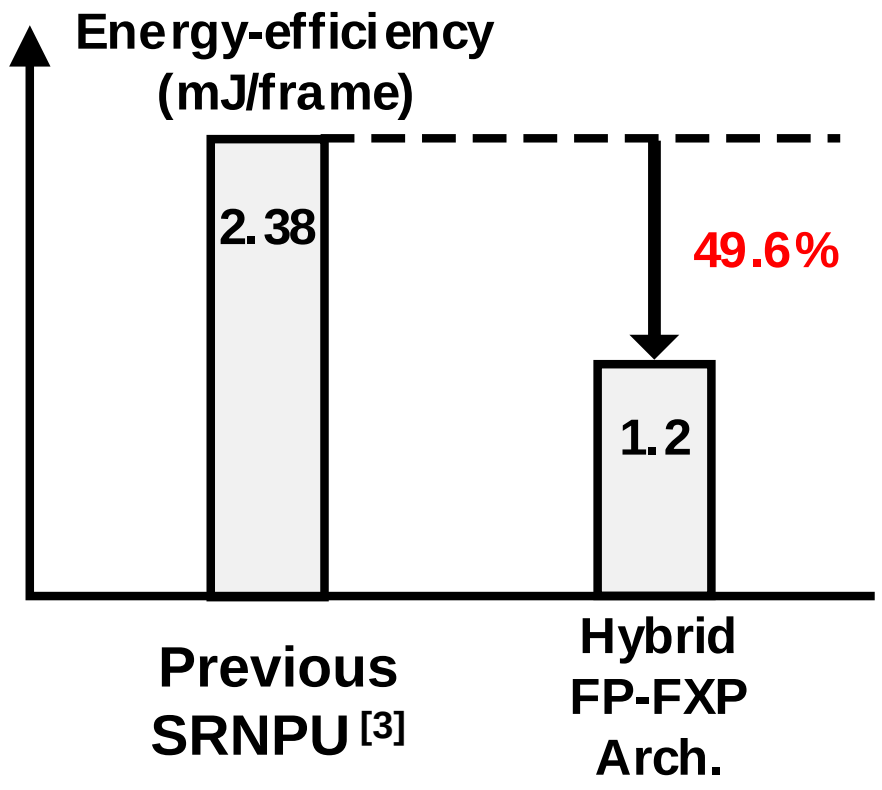
4 3x1 1D PE Lane:
Sparse Multiply & Act. Func. operation
Scoreboard based parallelism

Accumulator & Reorder L1$
Sparse Accumulate

Results of Heterogeneous Accelerating Arch.



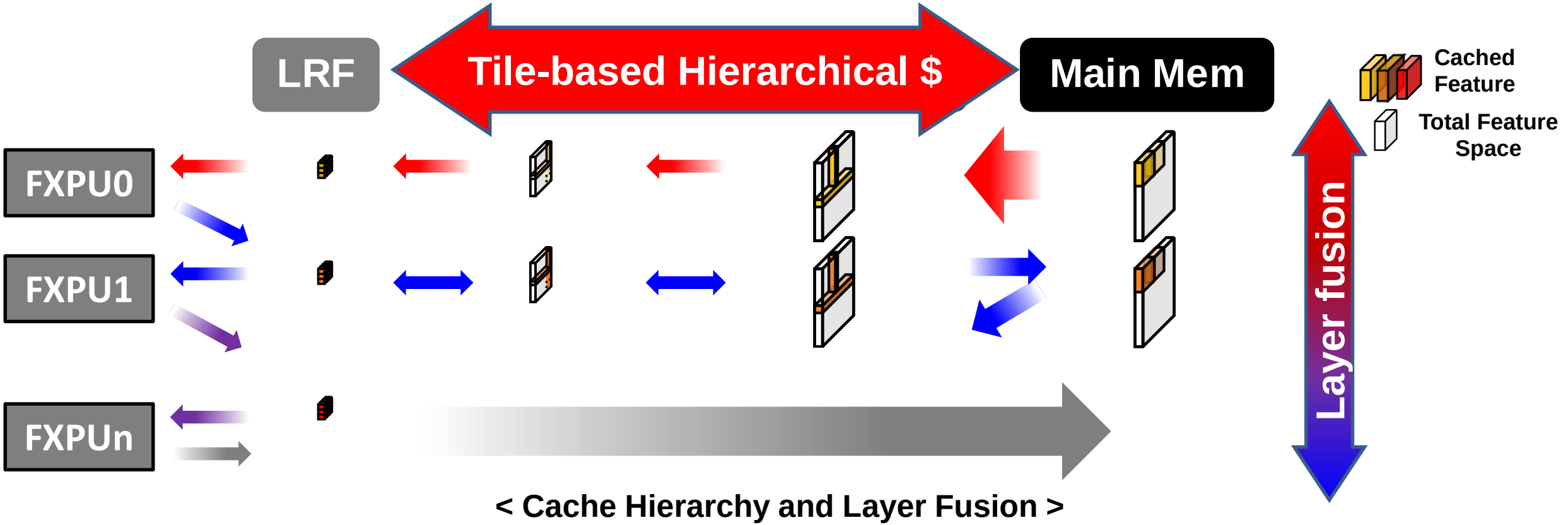
< Power Reduction by HAA >



< Energy-efficiency Improved by HAA >

Proposed Hierarchical Cache Subsystem

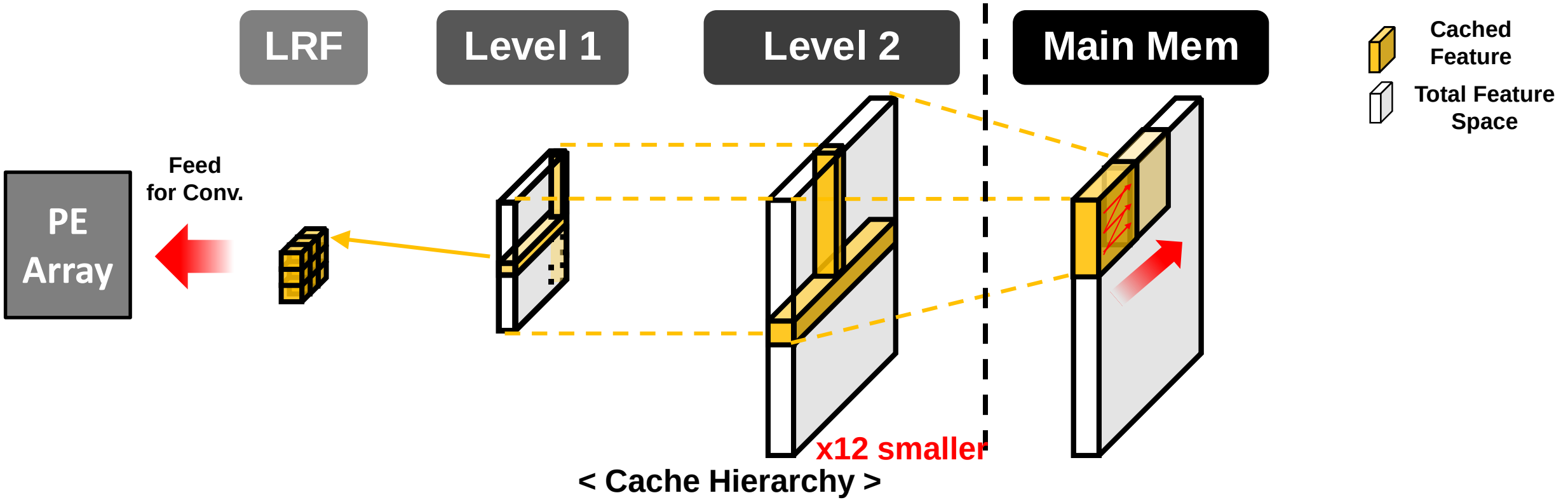
- **2-level Hierarchical Line Cache**
  - Tile-based hierarchical cache to **reduce cache size & power** ☺
  - Multi-core layer fusion to **reduce EMA** ☺



Proposed Hierarchical Cache Subsystem

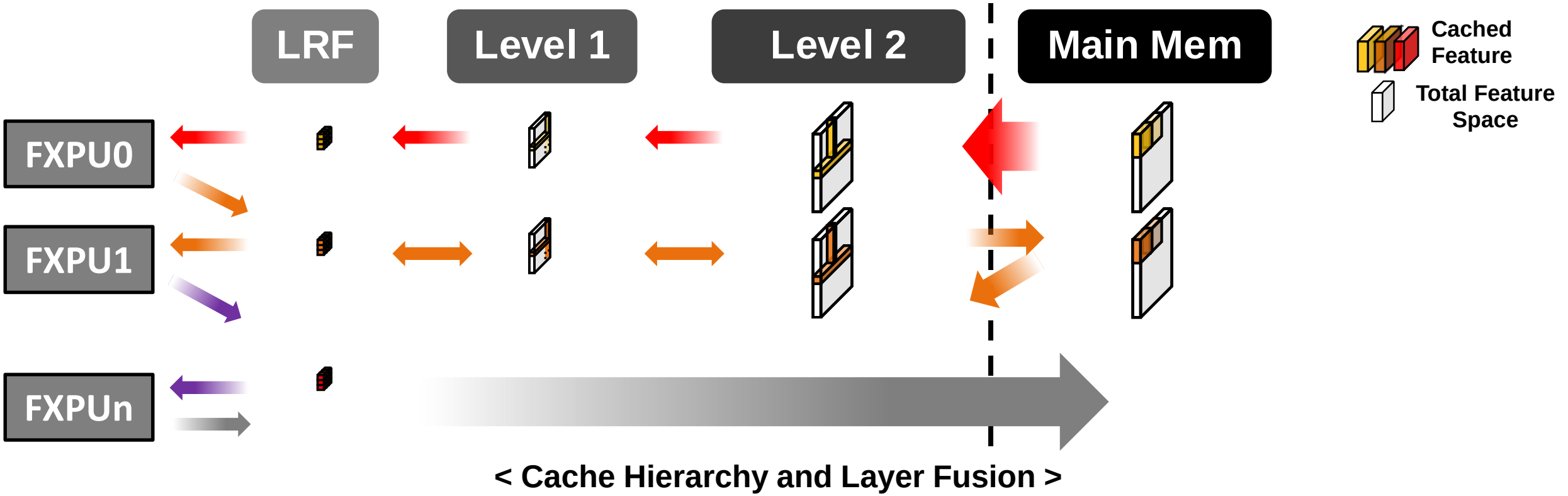
- **Tile-based Execution**
  - Fetch small tile of entire input each time

  ➔ Shorter reuse distance between lines ➔ x12 smaller on chip cache ☺



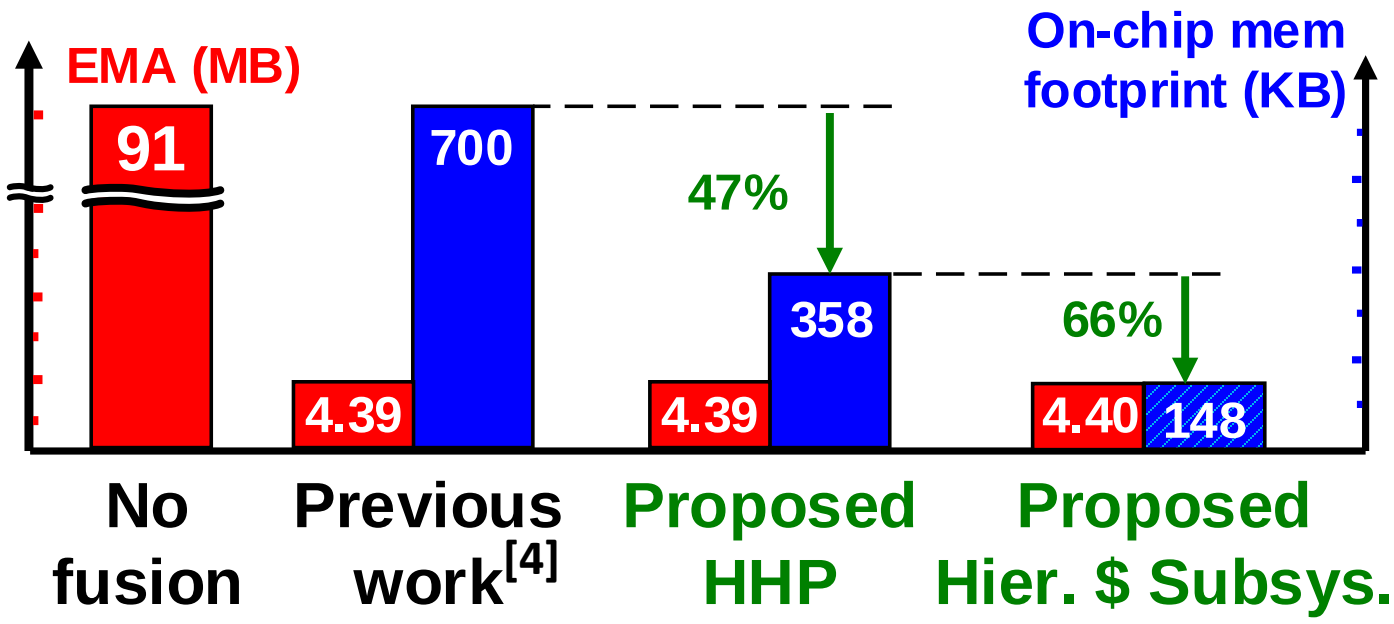
Proposed Hierarchical Cache Subsystem

- **2-level Hierarchical Line Cache**
  - Tile-based hierarchical cache to reduce cache size & power ☺
  - Multi-core layer fusion to reduce EMA ☺

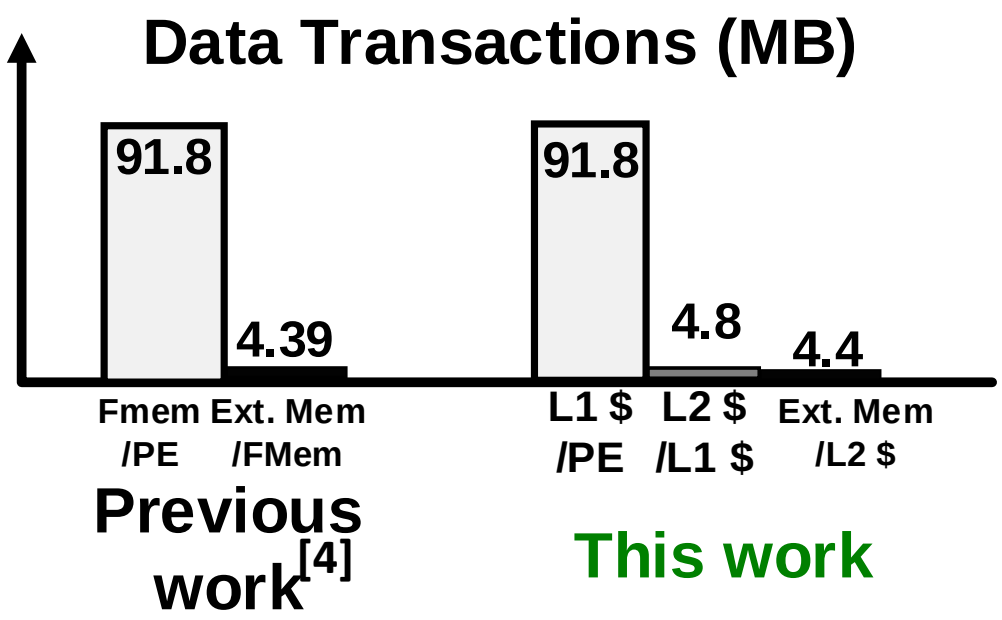


Result of Hierarchical Cache Subsystem

- Additionally reduce 66% global memory footprint
- Reduce 18x L2 cache(Large memory) access



< On-chip Memory Footprint for fusion>

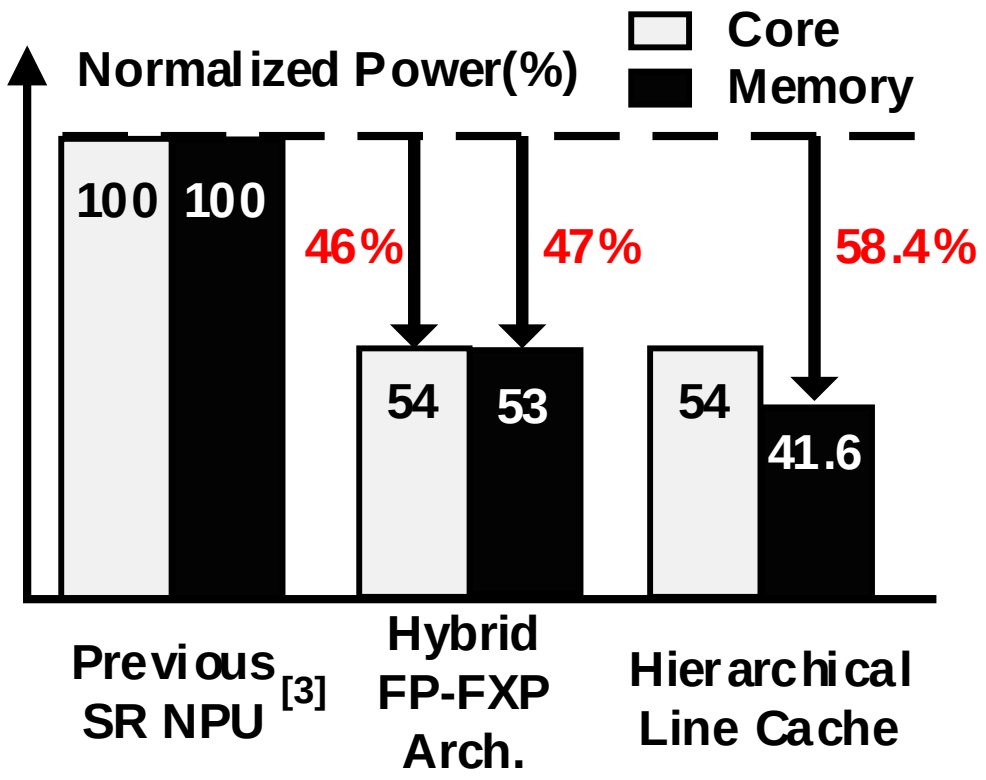


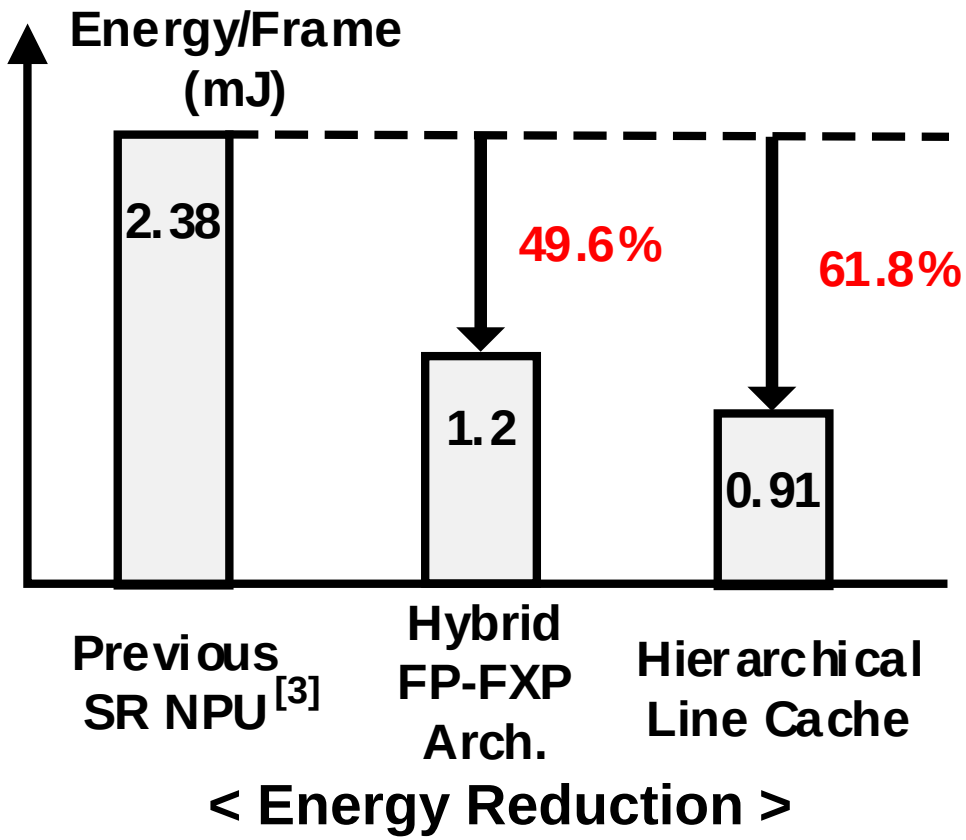
< On-/Off-chip Data Transactions>

[4] K. Goetschalckx et al. S.VLSI 2021;

Result of Hierarchical Cache Subsystem

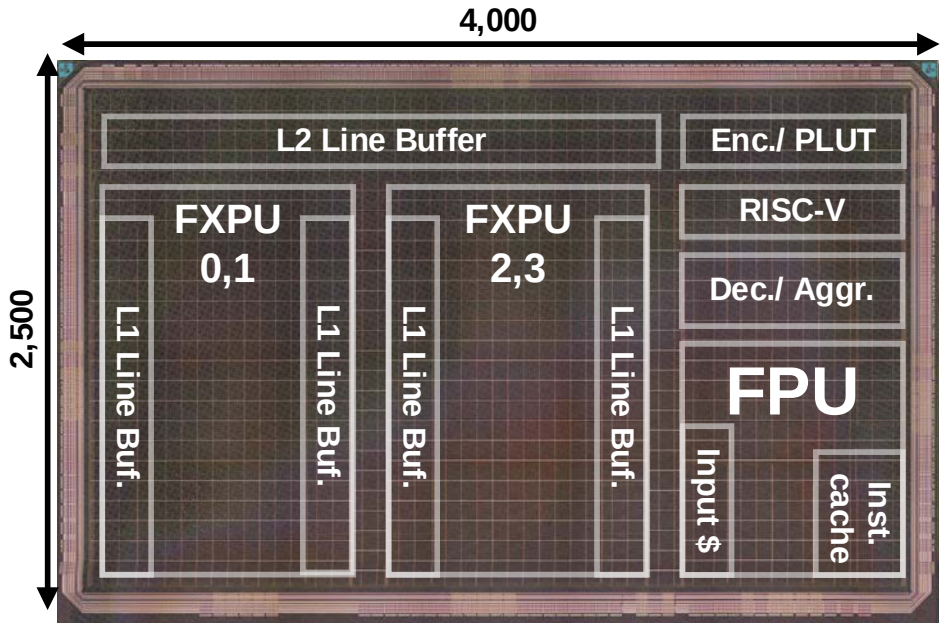
- With smaller cache subsystem **58.4 %** power ▼
- With fused layer **61.8%** energy ▼





[3] J. Lee et. al S. VLSI 2019

Chip Photography and Summary



< Chip Photograph >

Resolution	270p→1080p	540p→4k	GAN
Method	FSRCNN	ClassSR	OMGD
Framerate	107 fps	41 fps	86 fps

< ISP Performances >

Process	65nm
Supply Voltage (V)	1.0
Frequency (MHz)	200
SR Algorithm	FSRCNN / ClassSR
Activation Precision	FXP8 + 5~10% FP8
Frame Rate* (fps)	107
SR Energy* (mJ/frame)	0.92

< Chip Summary>

* x4 FSRCNN on Set5, Set14 Dataset

Conclusion

- **Energy-Efficient Non-sparse High-quality SRCNN**
  - **Heterogeneous Accelerating w/ Hybrid-precision**
    - 46% Processing Power ▼ & 47% Mem Access Power ▼ & 47% EMA ▼
  - **Data Lifetime-aware Hierarchical Line cache**
    - 53.7% Mem Access Power ▼ & 71.8% EMA ▼

A 0.92 mJ/frame Super-resolution SoC for Resource-limited Mobile Applications

Thank You!

- **Questions? Feel Free to Contact Me!**
  - E-mail: zhiyong_li@kaist.ac.kr
  - Lab homepage: https://ssl.kaist.ac.kr
  - Zoom Meeting: https://hc34.slack.com/?redir=%2Fmessages%2Fp14-kaist-fhd_mobile_soc