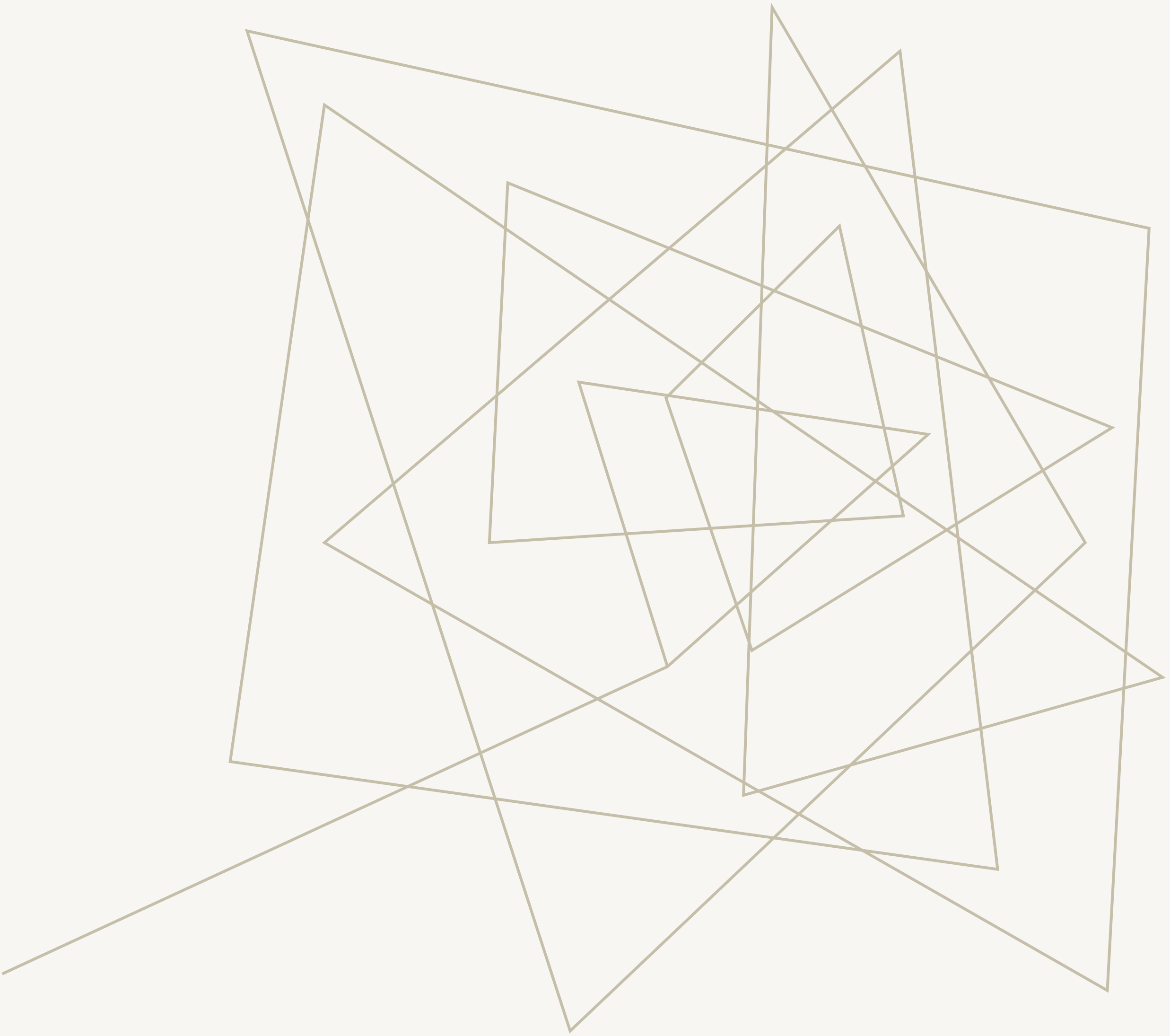
# BLOCK REPLACEMENT POLICY ON HYBRID CACHE

By Akash Pal



# What is a Hybrid Cache?

# HYBRID CACHE

- A hybrid cache refers to a caching architecture that combines different types of cache memories in a single system.
- The goal is to leverage the strengths of each type of cache to optimize overall performance.

# HYBRID CACHE ARCHITECTURE

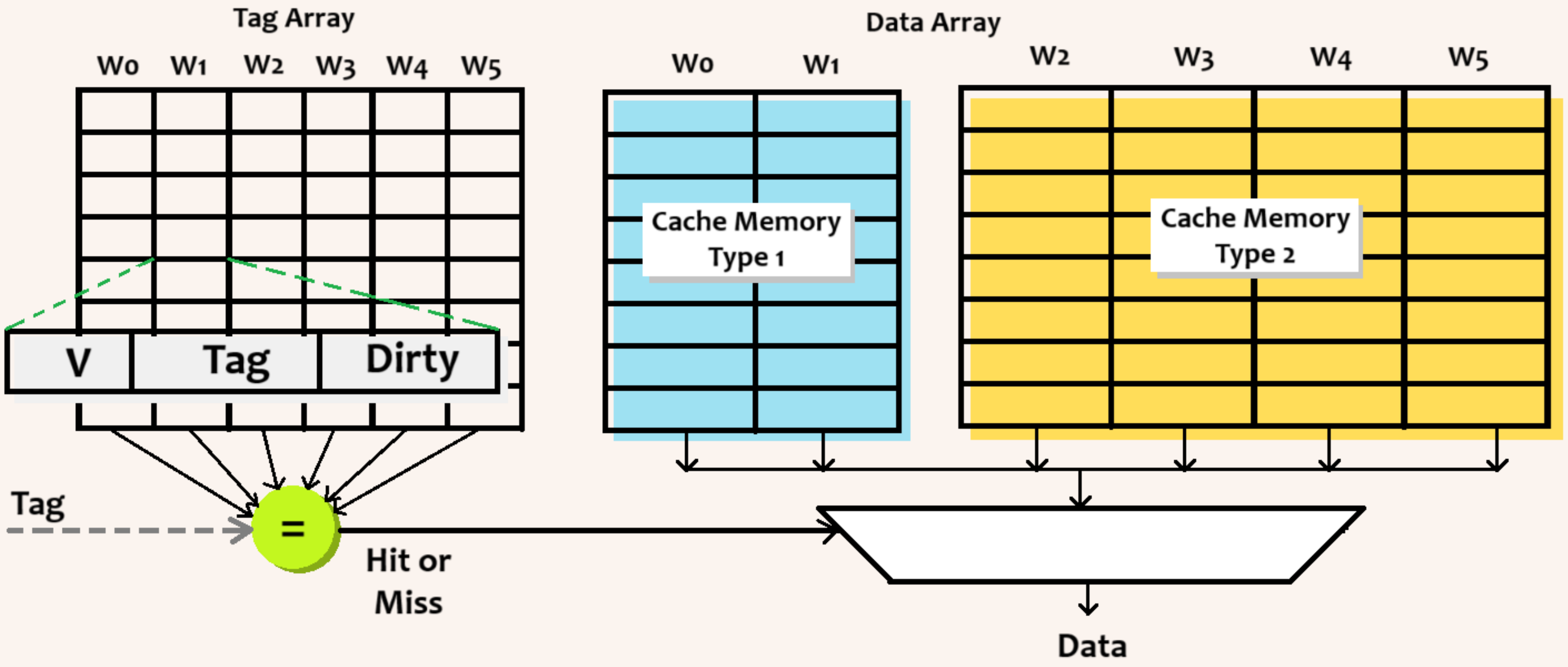


Fig: 1

# CACHE MEMORY TYPES

## STT-RAM

Spin-Transfer Torque RAM works by utilizing a spin-polarized current to manipulate the orientation of magnetic spins.

## MRAM

Magnetoresistive RAM works by storing data using the resistance changes in magnetic tunnel junction.

## PRAM

Phase-Change RAM works by using the reversible phase change of a chalcogenide glass material.

## SRAM

Static RAM works by using flip-flops to store binary data in cross-couples inverters.

NON-VOLATILITY

Retains Data even when power is turned off.

HIGH DENSITY

Allowing higher memory density results the proper utilization of space.

LOW LEAKAGE POWER

Due to no need of refreshing energy, it’s energy leakage is very low. Static power (on standby) is also very less.

HIGH-SPEED OPERATION

It has potential for high-speed read and write operations compared to certain non-volatile memory.

# STT-RAM: ADVANTAGES

# STT-RAM: DISADVANTAGES

## WRITE ENERGY CONSUMPTION

It consumes high energy to write on a memory cell. It is 172 pJ. Following graph shows the comparison. (Fig: 2)

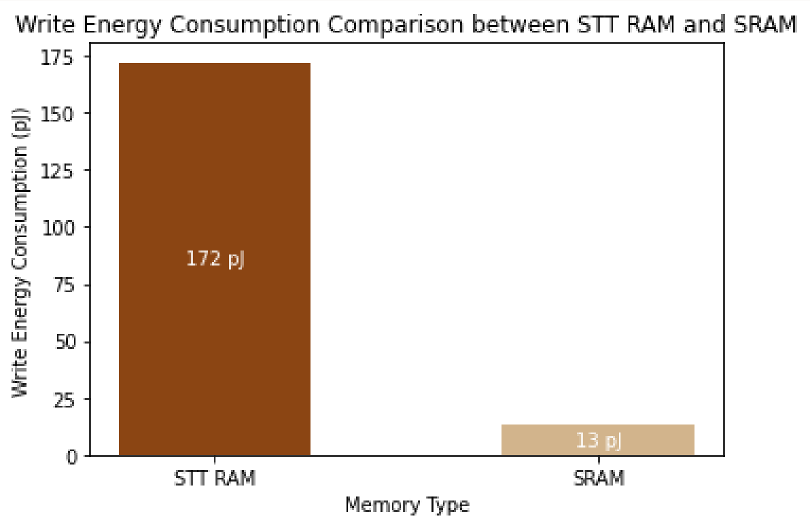


Fig: 2

## LOW WRITE ENDURANCE

Number of writes on STT-RAM is very less through out its life time. It is less than $10^{12}$.

Also write latency is very high. (Fig: 3)

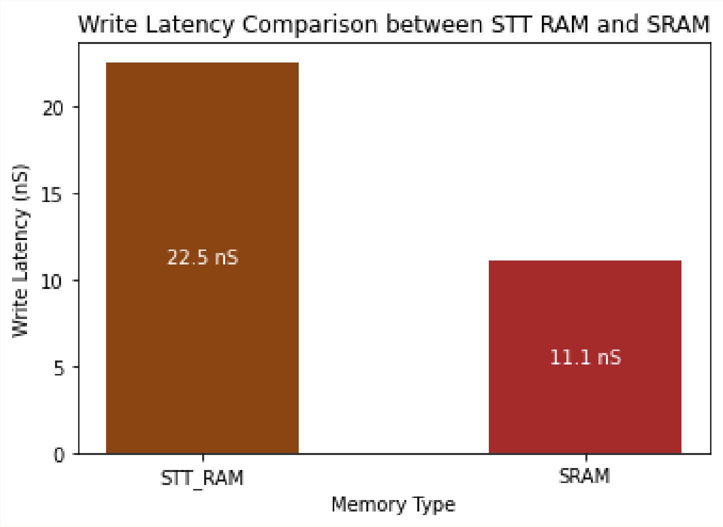


Fig: 3

# SRAM: ADVANTAGES

## HIGH WRITE ENDURANCE

Number of writes has negligible impact on the life time of SRAM.

We can write >>$10^{16}$ times on the cell of SRAM.

## LOW WRITE ENERGY CONSUMPTION

It consumes very low energy for write operation. It is 11.1 pJ to be specific. (Fig: 2)

## HIGH-SPEED OF READ-WRITE ACCESS

Due to flip-flop structure, it allows quick and direct access to stored data.

# SRAM: DISADVANTAGES

## LOW DENSITY

SRAM is not as dense as other memory technologies. Space utilization is very less.

## VOLATILITY

Stored Data will be lost by turning off the power.

## HIGH LEAKAGE POWER

Due to flip-flop structure implemented with capacitor, leakage power is very high. (Fig: 4)

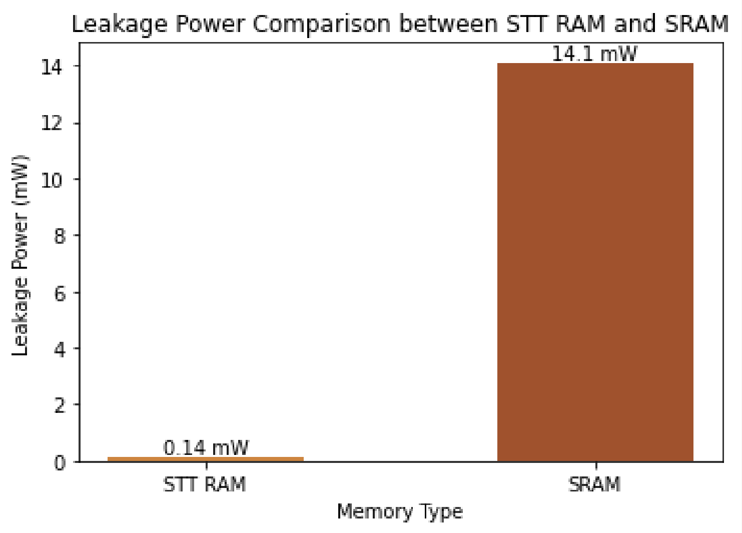


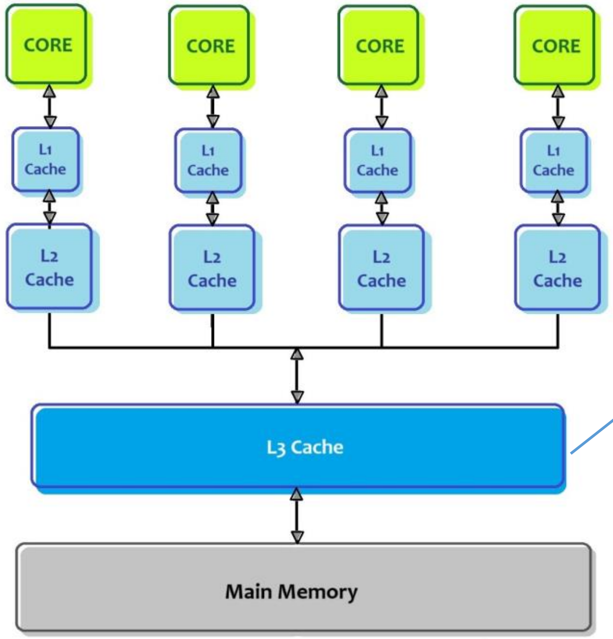
Fig: 4

# PROBLEM STATEMENT

Properly utilize hybrid cache architecture so that we can optimize latency, energy consumption, and cache lifetime in modern computing systems by implementing adaptive caching technique.

# BACKGROUND

How cache works?



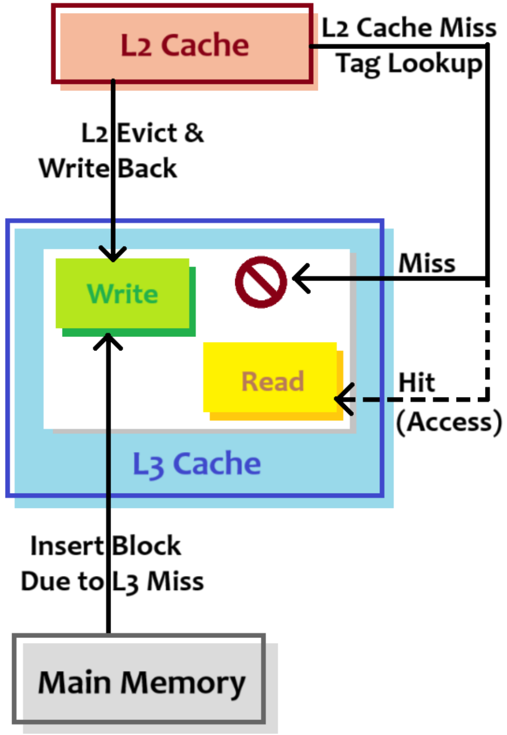


Fig 6: Cache Hierarchy

Fig 7: L3 Cache Read Write Operation

# PROPOSAL

- As SRAM has high write endurance and low write energy consumption, SRAM can be used as write region.
- Write intensive blocks are those on which heavy number of write operation will be performed.
- Write intensive blocks will be redirected towards write region.

HOW TO DECIDE WHETHER WRITE INTENSIVE OR NOT?

~ Paper[1] proposed that the block loaded into cache due to write miss on that cache is write intensive.

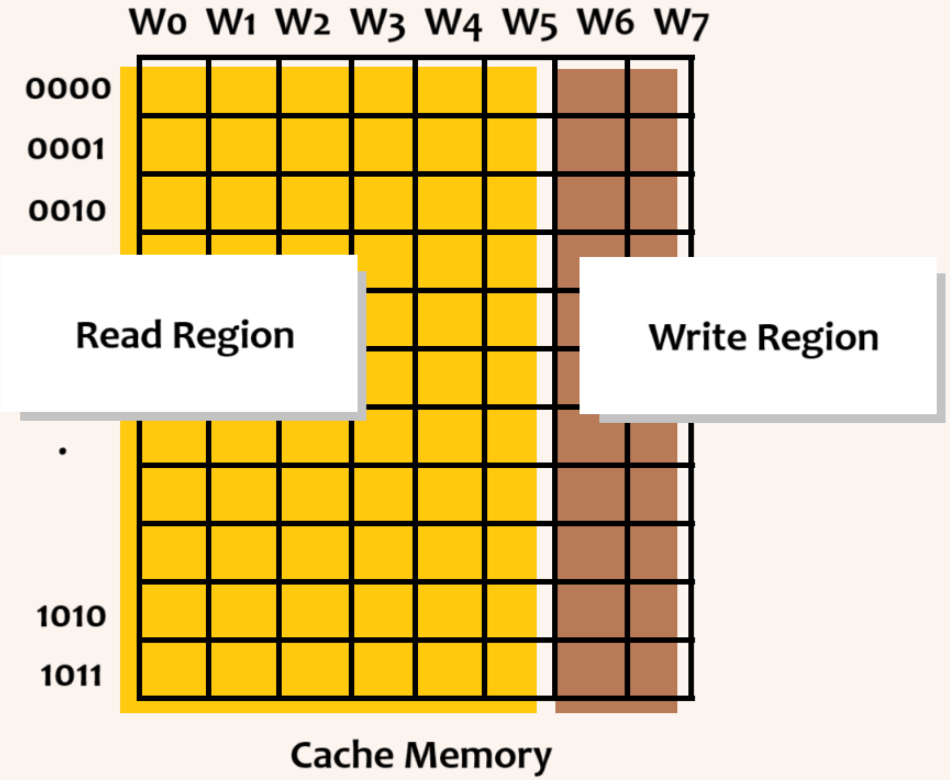


FIG 5: PARTITIONING

[1] : X. Wu, J. Li, L. Zhang, E. Speight, and Y. Xie, “Power and performance of read-write aware hybrid caches with non-volatile memories,” in 2009 Design, Automation Test in Europe Conference Exhibition, 2009, pp. 737–742.

# PROPOSAL

- Paper also proposed a saturated counter and a swap buffer.
- Saturated counter is implemented on each cell to count number of hits on wrong region.

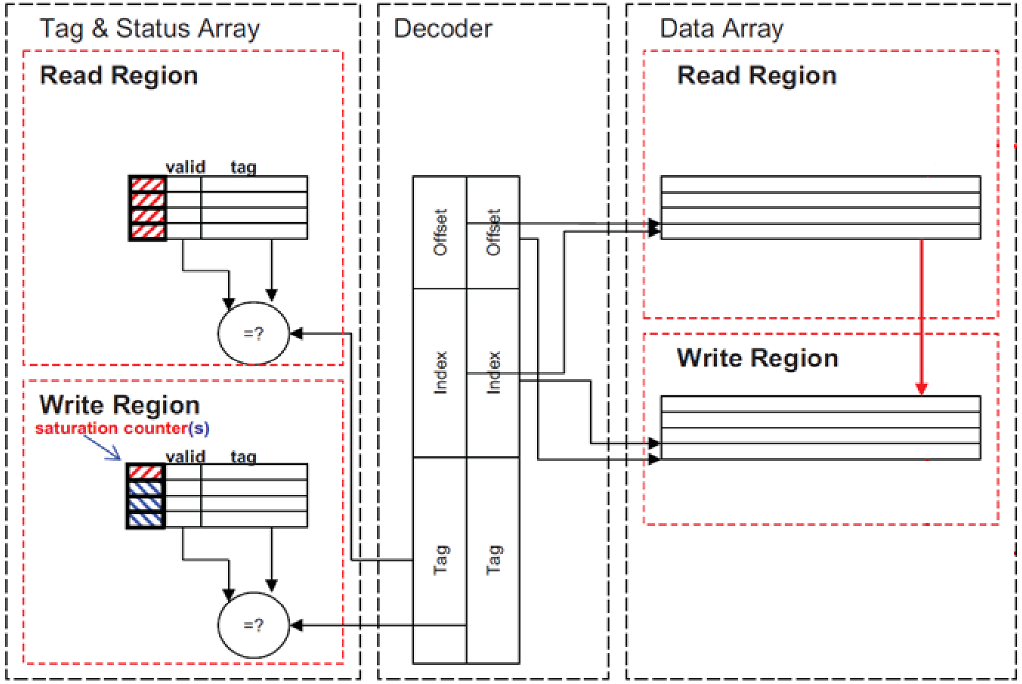
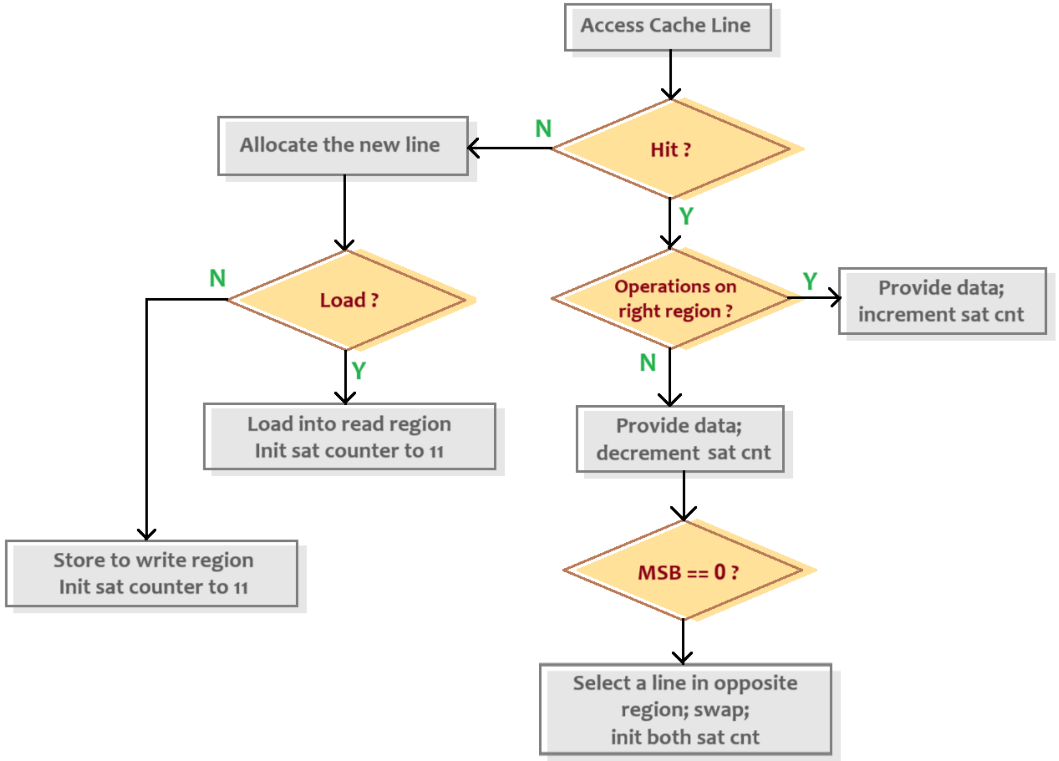
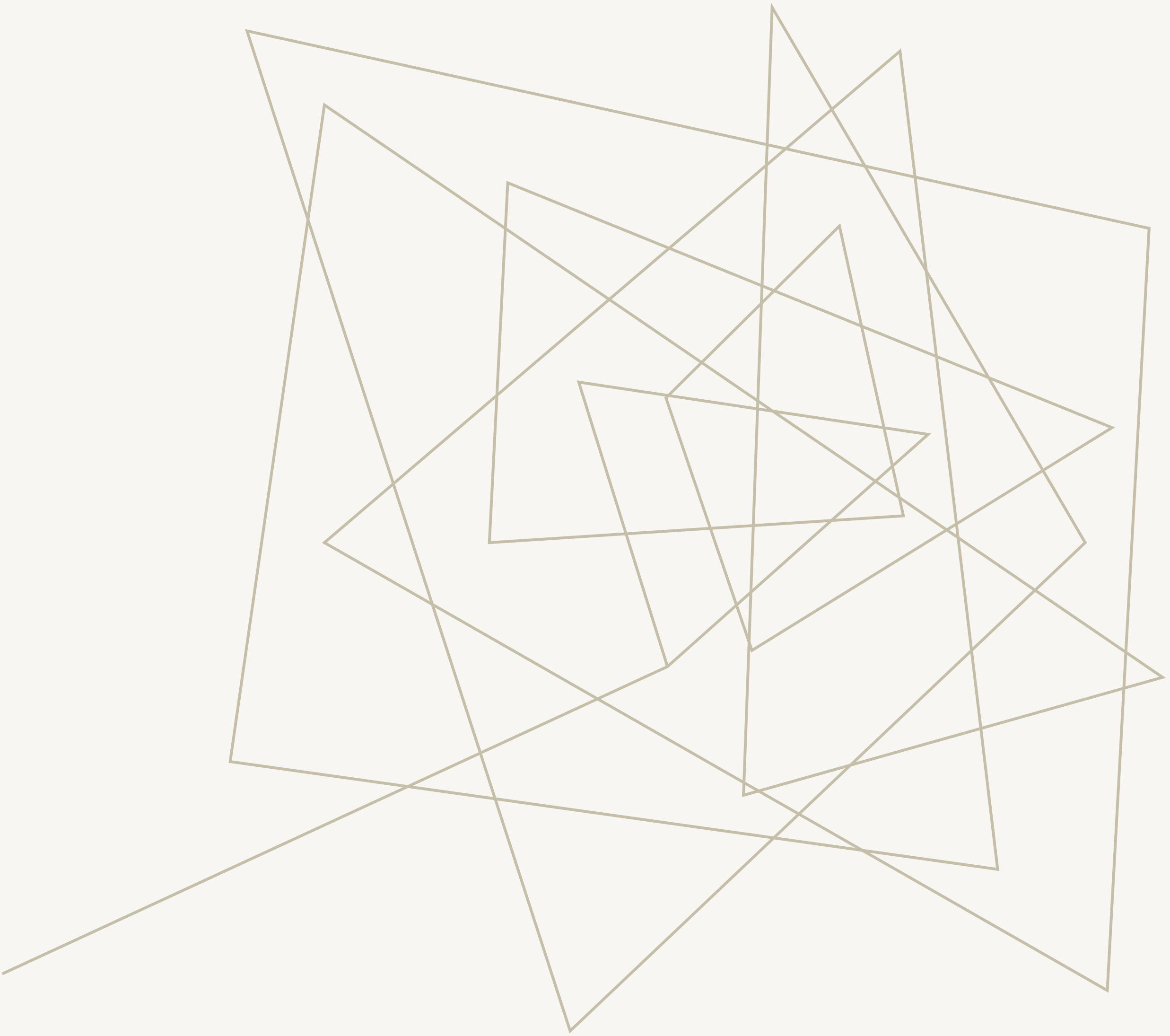


Fig 8: Block Diagram of Proposed Method

(Source: X. Wu, J. Li, L. Zhang, E. Speight, and Y. Xie, “Power and performance of read-write aware hybrid caches with non-volatile memories,” in 2009 Design, Automation Test in Europe Conference Exhibition, 2009, pp. 737–742. )





# LITERATURE REVIEW

# PROPOSAL

**Trigger Instruction (TI)** of a cache block is load/store instruction that makes the block loaded into cache.

**HOW PREDICTOR WORKS?**

- Another Paper[1] proposed a write intensive predictor to predict it.
- It also comes with a counter implemented with a threshold based on which the write intensive line is predicted.

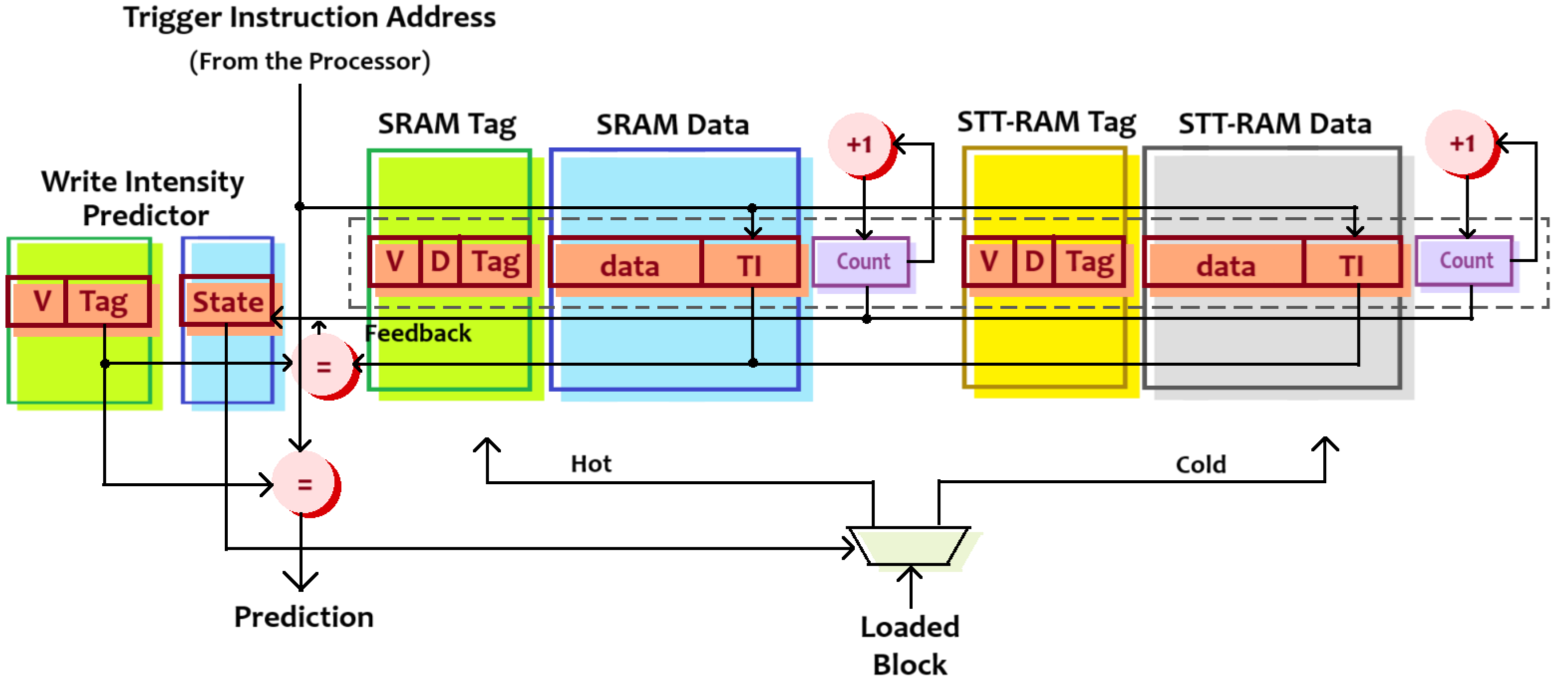


Fig 9: An overview of a hybrid cache architecture with the write intensity predictor.

[1] : J. Ahn, S. Yoo, and K. Choi, "Write intensity prediction for energy-efficient non-volatile caches," in International Symposium on Low Power Electronics and Design (ISLPED), 2013, pp. 223–228.

# PROPOSAL

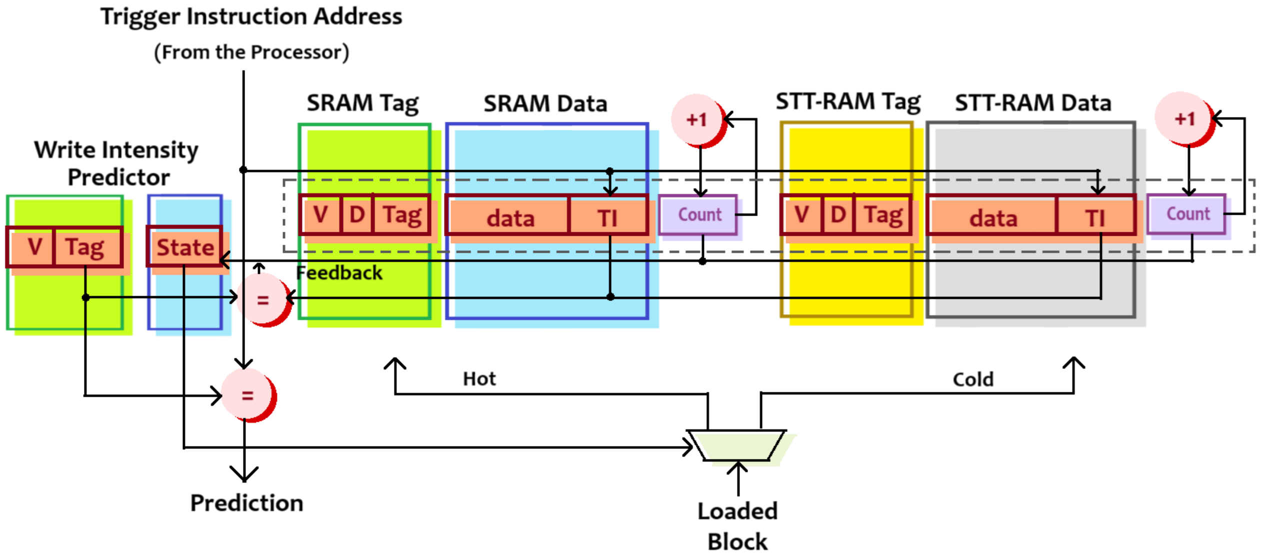


Fig 9: An overview of a hybrid cache architecture with the write intensity predictor.

## ON WRITE HIT

Data writes on data array and counter will increment. If threshold reaches, tag field of predictor is compared with TI field and state is updated using feedback mechanism.

## ON CACHE MISS

Along with previous work[1] (implementation), predictor is used also while loading the block.

- If TI of the block is store instruction, load into SRAM region.
- If TI of block is load instruction, predictor is searched on its list with TI address.

[1] : X. Wu, J. Li, L. Zhang, E. Speight, and Y. Xie, “Power and performance of read-write aware hybrid caches with non-volatile memories,” in 2009 Design, Automation Test in Europe Conference Exhibition, 2009, pp. 737–742.

# IMPLEMENTATION

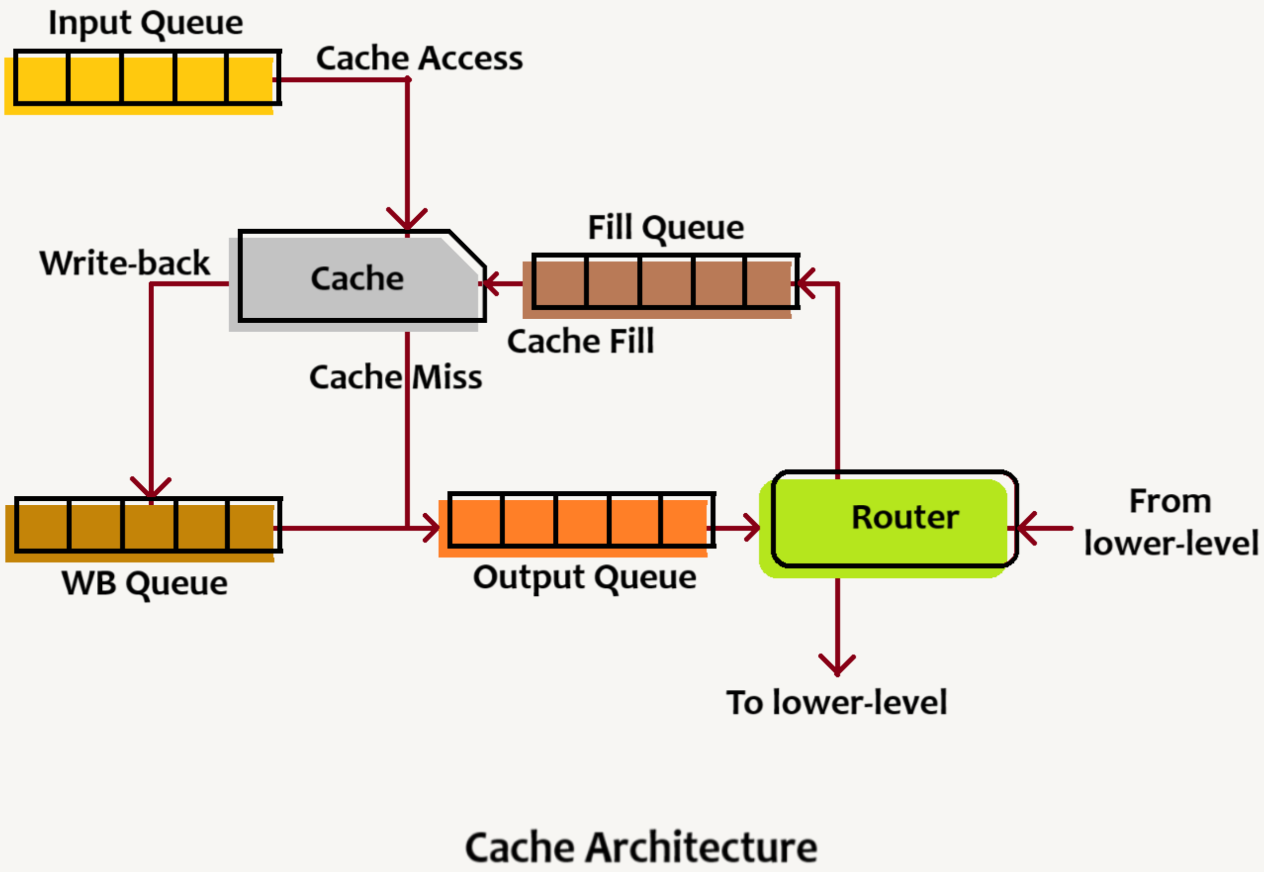
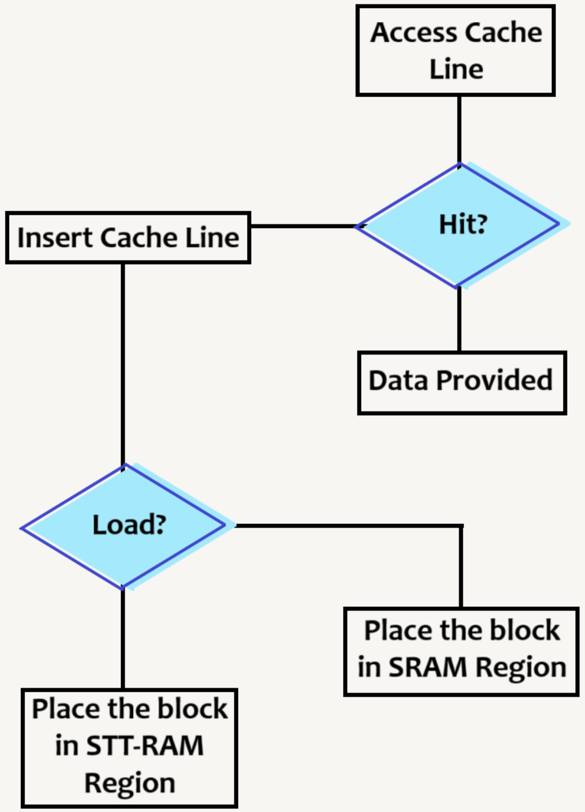


Fig. 10: Cache Memory Structure



Access Cache Line
Hit?
Insert Cache Line
Data Provided
Load?
Place the block in SRAM Region
Place the block in STT-RAM Region

Fig. 11: Flow Chart of Implementation

# IMPLEMENTATION

| COMPONENTS | CONFIGURATIONS |
|---|---|
| Processor | 3 GHz., 4 CPU cores type x86. with RR policy |
| L1 Cache | 64 KB, 8-way set-associative, 64 bytes line size, 3-cycle latency |
| L2 Cache | 256 KB, 8-way associative, 64 bytes line size, 8-cycle latency |
| L3 Cache | Hybrid, 1 MB, 16-way set-associative (12-way STT RAM, 4-way SRAM), 64 bytes line size, 30-cycle latency |
| Main Memory | 8-bus width, DRAM, 2048 row buffer size with FRFCFS policy |

Table 1: Configuration of Macsim Simulator.

# IMPLEMENTATION

Latency and Energy Consumption values for different types of memory technology.

| FEATURE | STT-RAM | SRAM |
|---|---|---|
| Read Energy (pJ) | 17.5 | 15.8 |
| Write Energy (pJ) | 172 | 13 |
| Read Latency (ns) | 11.4 | 11.1 |
| Write Latency (ns) | 22.5 | 11.1 |
| Power Leakage (mW) | 40 | 400 |

Table 1: Characteristic of STT-RAM and SRAM.

# RESULT

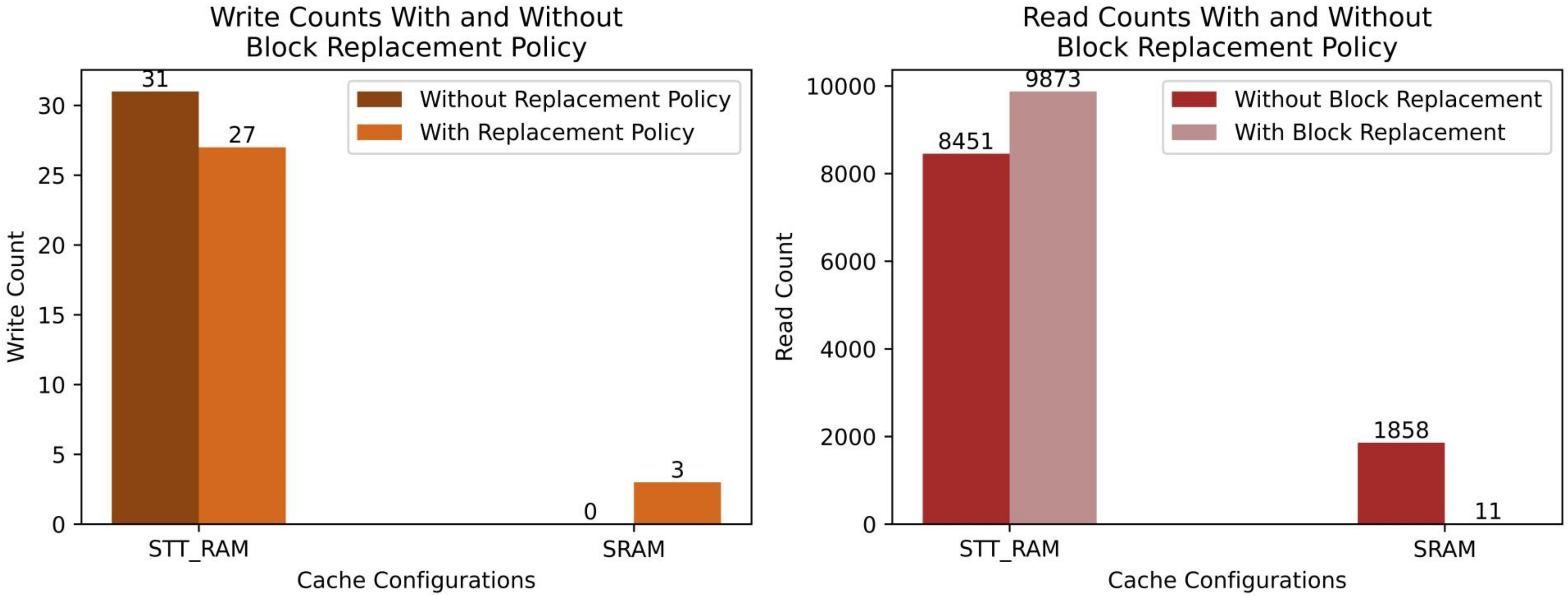


Fig. 12: Read & Write Count Improvement

# RESULT

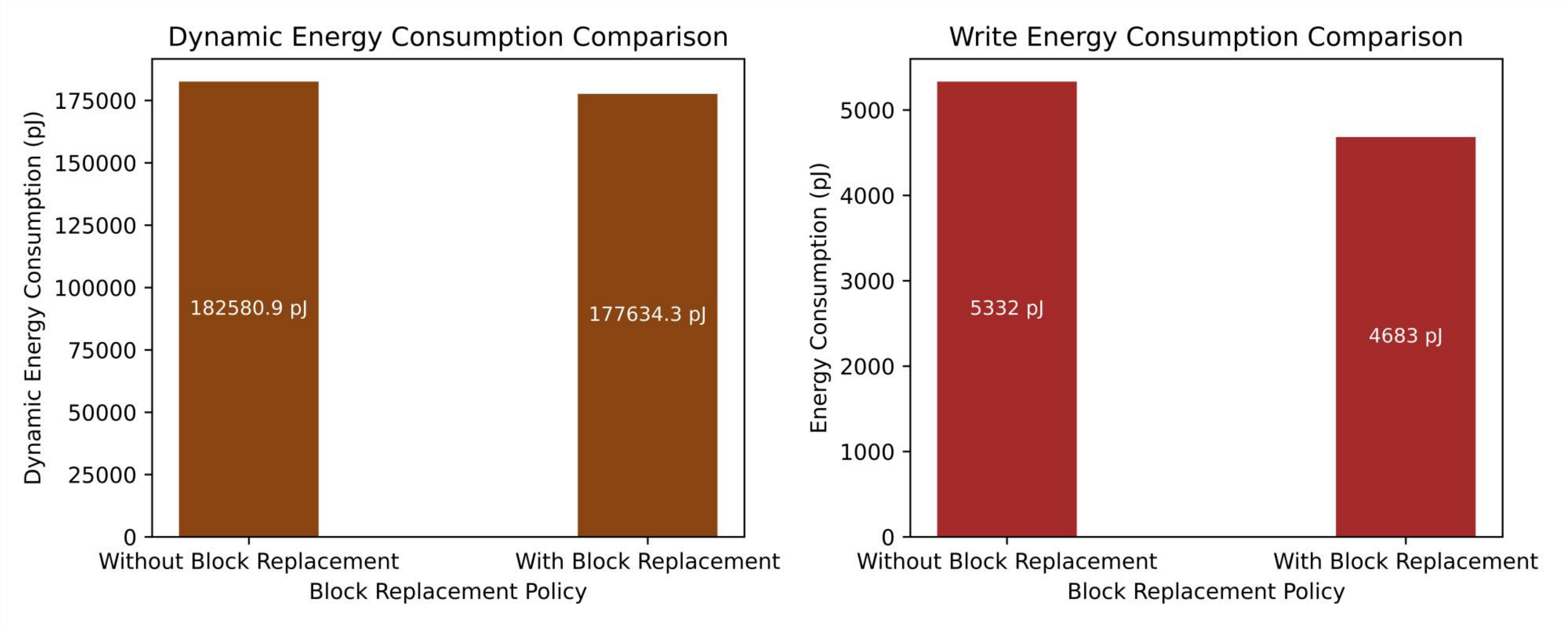


Fig. 13: Dynamic Energy and Write Energy Consumpsion Comparison

# RESULT

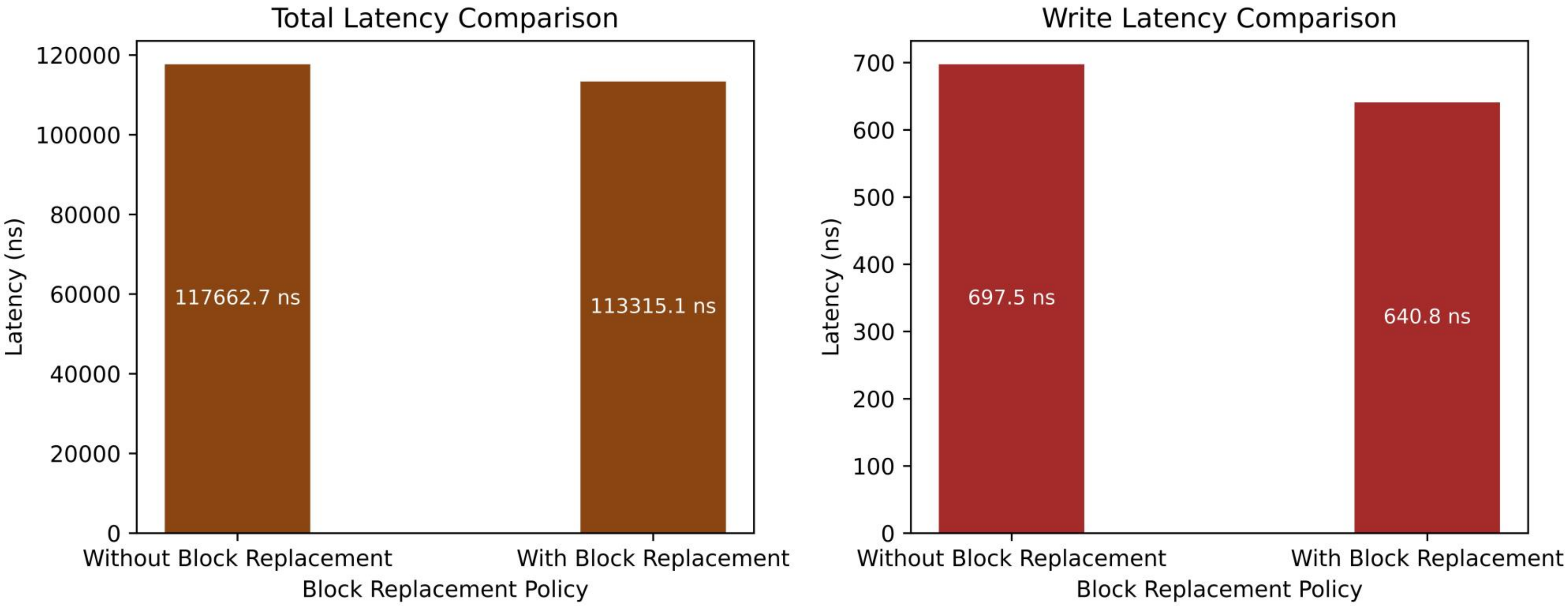


Fig. 14: Read and Write Latency Comparison

# IMPROVEMENT

10%

## WRITES ON RIGHT REGION

10% more write intensive block placed in write region in this implementation

3.69%

## LATENCY

3.69% reduction in total latency, and 8.14% reduction in write latency of L3 cache memory

2.71%

## ENERGY CONSUMPTION

2.71% reduction in total energy consumption, and 12.17% reduction in write energy consumption by L3 cache memory

# CONCLUSION

## Conclusion

This implementation shows very less improvement over hybrid cache.

## Future Work

We can try to implement block replacement policy on a heterogeneous system where shared cache faces huge data handling.

https://github.com/Zeekersky/MacSIM_Progress

# THANK YOU