ScaleML: Machine Learning based Heap Memory Object Scaling Prediction

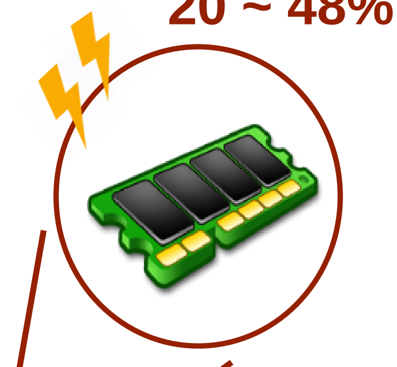
Joongeon Park[1], Safdar Jamil[1], Awais Khan[1], Sangkeun Lee[2], Youngjae Kim[1]

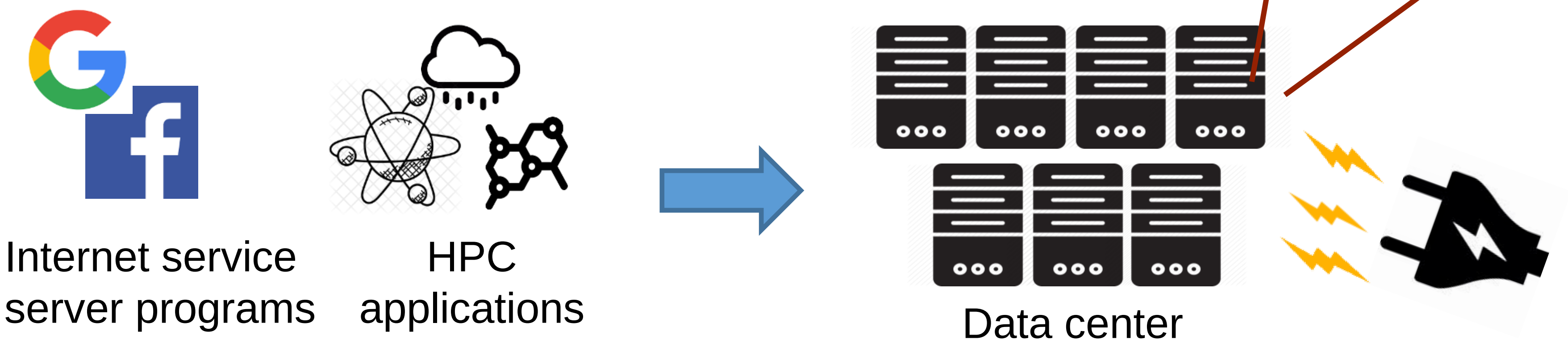
[1]Sogang University, Seoul, Korea
[2]Oak Ridge National Laboratory, TN, US

LASS Laboratory for AI System Software

Immense Energy Consumption

- Internet service servers & large-scale HPC applications running in data center consume tremendous energy
  - 173% increase in data throughput per year [1]
  - 1.8 Mega-Ton of CO_2 emission by Google data center [2]
- Considerable portion is consumed in memory!
  - 20 ~ 48% of total machine's energy consumption [3]

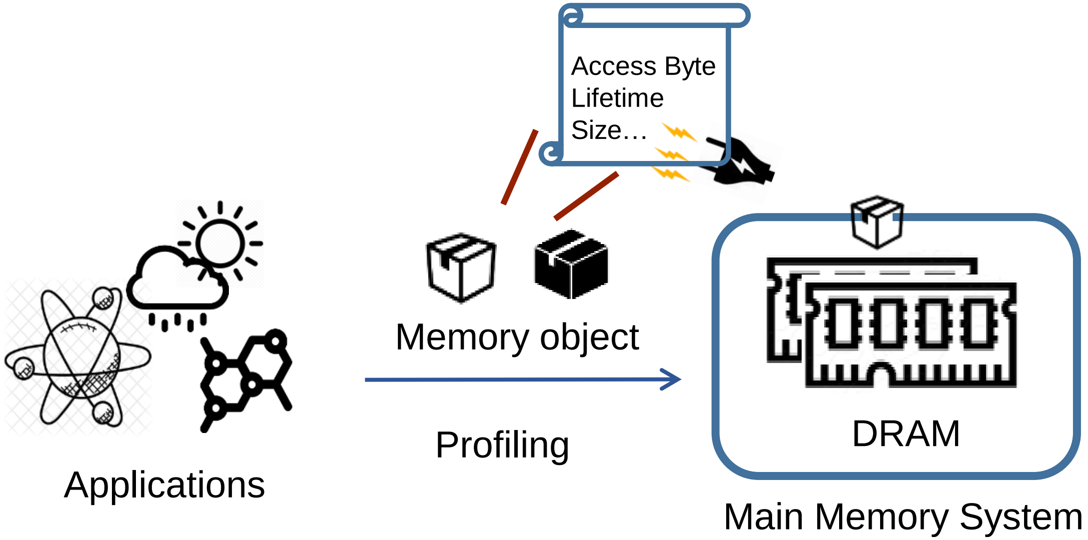




[1] Z. Jia, L. Wang, J. Zhan, L. Zhang, and C. Luo, "Characterizing data analysis workloads in data centers," in Proceedings of the IEEE International Symposium on Workload Characterization (IISWC), pp. 66–76, 2013.https://www.forbes.com/sites/forbestechcouncil/2017/12/15/why-energy-is-a-big-and-rapidly-growing-problem-for-data-centers/
[2] The Guardian. "How viral cat videos are warming the planet." theguardian.com. https://www.theguardian.com/environment/2015/sep/25/server-data-centre-emissions-air-travel-web-google-facebook-greenhouse-gas
[3] M. Dayarathna, Y. Wen and R. Fan, "Data Center Energy Consumption Modeling: A Survey," in *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 732-794, Firstquarter 2016.

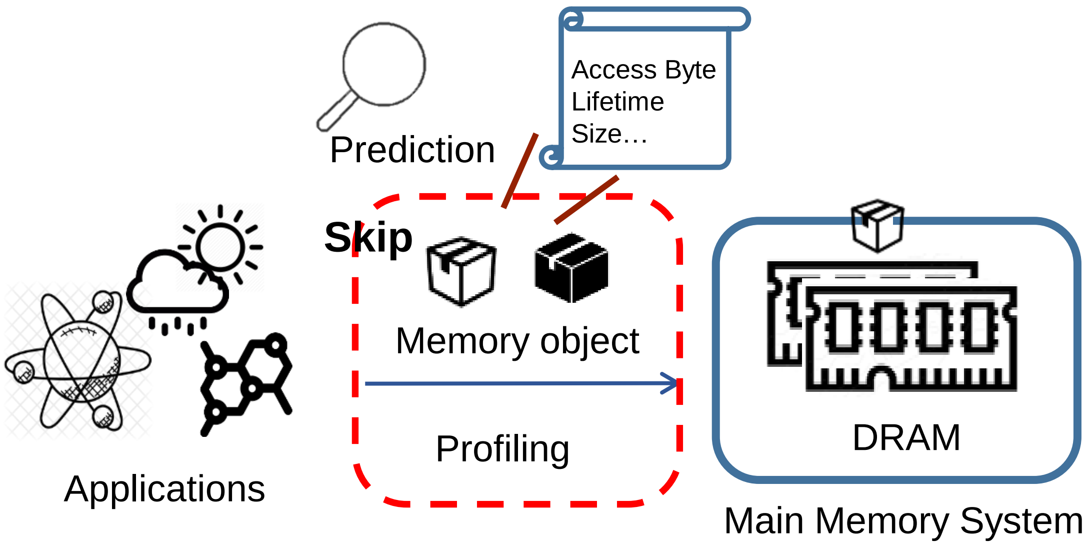
Immense Energy Consumption

- Software-based solutions to improve the memory-level energy efficiency have been proposed.
  - Previous studies have been conducted on energy-efficient object placement into DRAM by analyzing memory object access patterns.
  - However, profiling the access pattern of the memory object consumes a lot of energy.



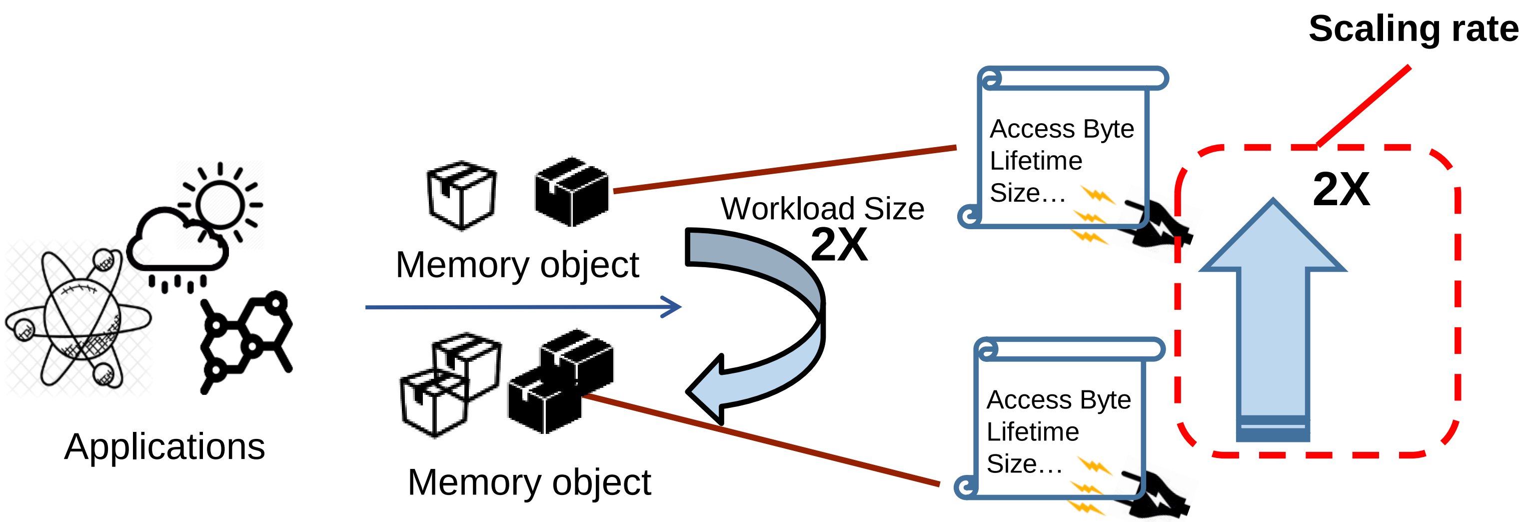
Existing Studies

- Studies have been conducted to predict profiling pattern of the memory object and skip the profiling process.
  - To predict profiling pattern of the memory object, memory access patterns of various workload sizes are used.
  - But, whenever application workload changes, the object access patterns also vary.



Existing Studies

- Linear Scaling Rate (LSR) is one of the solutions to address the energy-efficiency.
  - When the application workload size increases, the memory object access patterns also increase proportionally [4].
  - Existing energy-efficient object placement study [5] proposed LSR.



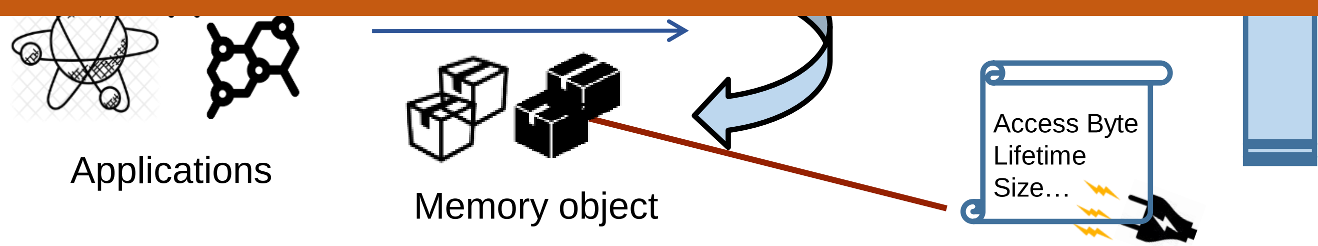
[4] Xu Ji, Chao Wang, Nosayba El-Sayed, Xiaosong Ma, Youngjae Kim, Sudharshan S. Vazhkudai, Wei Xue, and Daniel Sanchez. 2017. Understanding object-level memory access patterns across the spectrum. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '17).

[5] T. Kim, S. Jamil, J. Park and Y. Kim, "Optimizing Heap Memory Object Placement in the Hybrid Memory System With Energy Constraints," in IEEE Access, vol. 8, pp. 130323-130339, 2020.

Existing Studies

- Linear Scaling Rate (LSR) is one of the solutions to address the energy-efficiency.
  - When the application workload size increases, the memory object access patterns also increase proportionally [4].
  - Existing energy-efficient object placement study [5] proposed LSR

LSR has a limitation because it statically calculates the scaling rate according to the increase in the workload size.



[4] Xu Ji, Chao Wang, Nosayba El-Sayed, Xiaosong Ma, Youngjae Kim, Sudharshan S. Vazhkudai, Wei Xue, and Daniel Sanchez. 2017. Understanding object-level memory access patterns across the spectrum. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC ’17).
[5] T. Kim, S. Jamil, J. Park and Y. Kim, "Optimizing Heap Memory Object Placement in the Hybrid Memory System With Energy Constraints," in IEEE Access, vol. 8, pp. 130323-130339, 2020.

Motivation : Experiment Setup

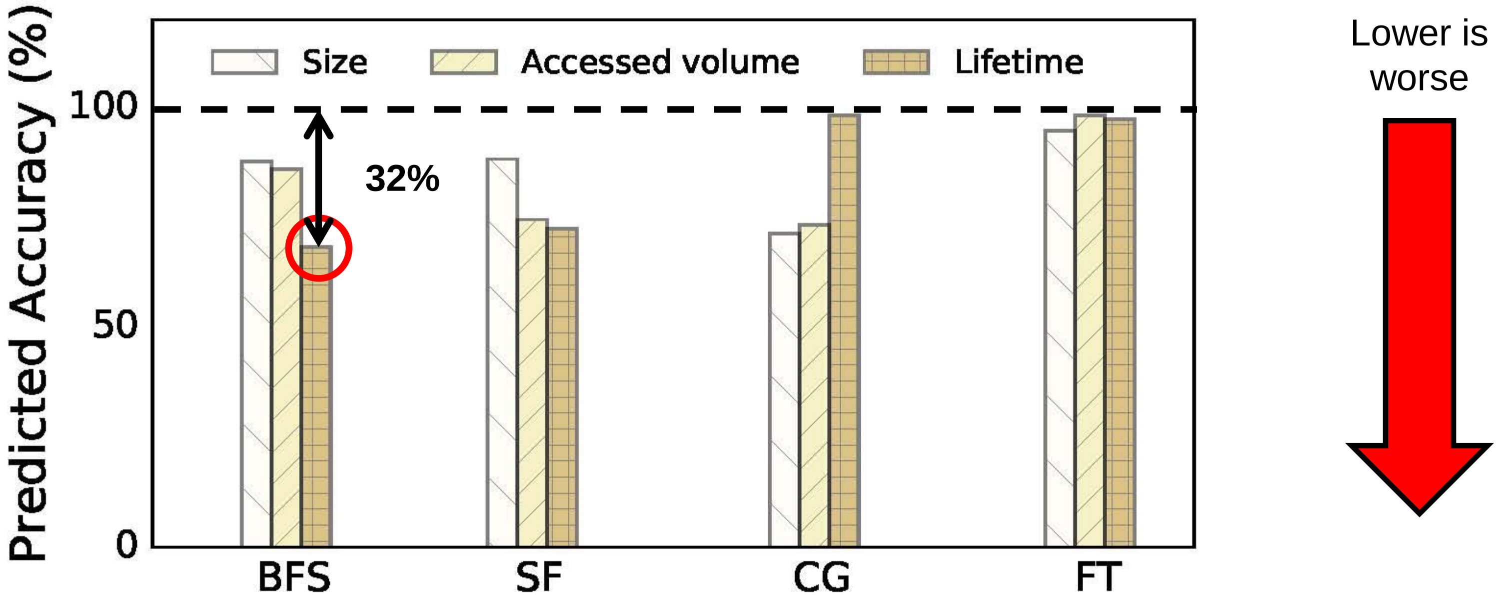
- ML Tool : ASCENDS [6]

- Benchmark
  - Problem Based Benchmark Suite (PBBS) : Breadth First Search (BFS), Spanning Forest (SF)
  - NAS Parallel Benchmark (NPB) : Conjugate Gradient (CG) and 3D fast Fourier Transform (FT)

[6] S. Lee, J. Peng, A. William, D. Shin, ASCENDS: Advanced data science toolkit for non-data scientists, Journal of Open Source Software, 5 (2020) 1656. https://doi.org/10.21105/joss.01656.

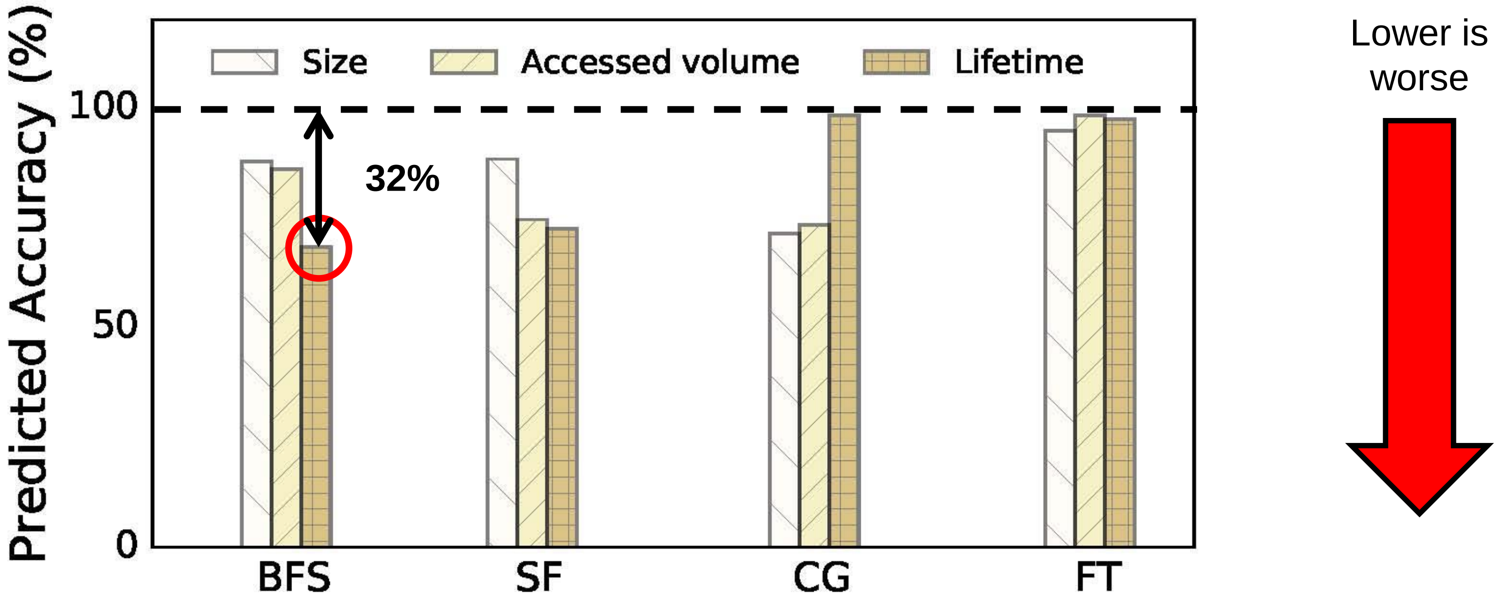
Existing Studies: Limitations

- Linear Scaling Rate (LSR) is one of the solutions to address the energy-efficiency.
  - When predicting the memory object access through LSR, the predicted value and the actual value showed a difference of about **32%**.



Existing Studies: Limitations

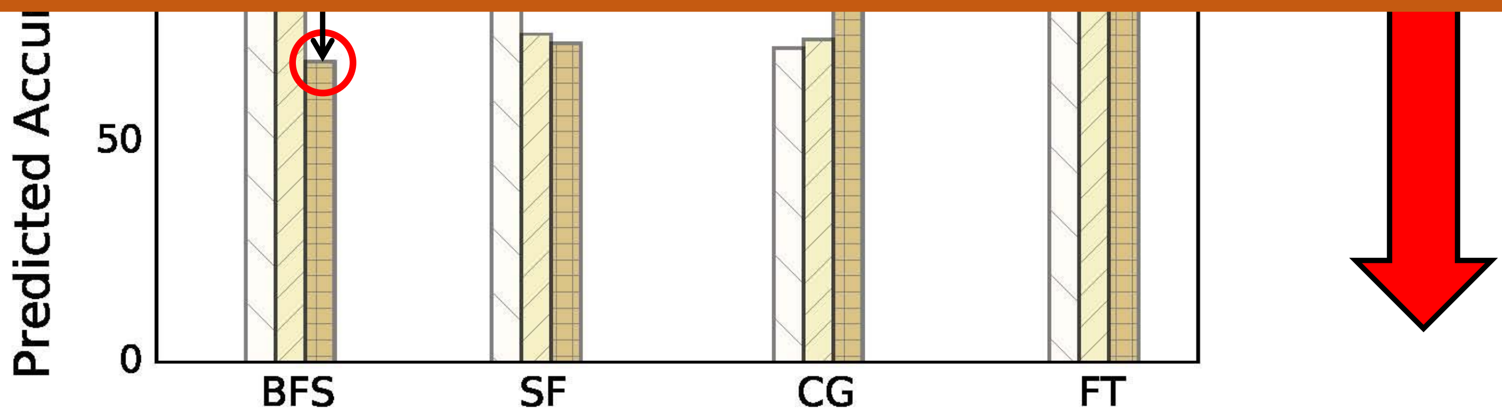
- Linear Scaling Rate (LSR) is one of the solutions to address the energy-efficiency.
  - When predicting the memory object access through LSR, the predicted value and the actual value showed a difference of about **32%**.
  - Moreover, the scaling rate is different for each memory object pattern in the a pplication, so it does not follow the LSR.



Existing Studies: Limitations

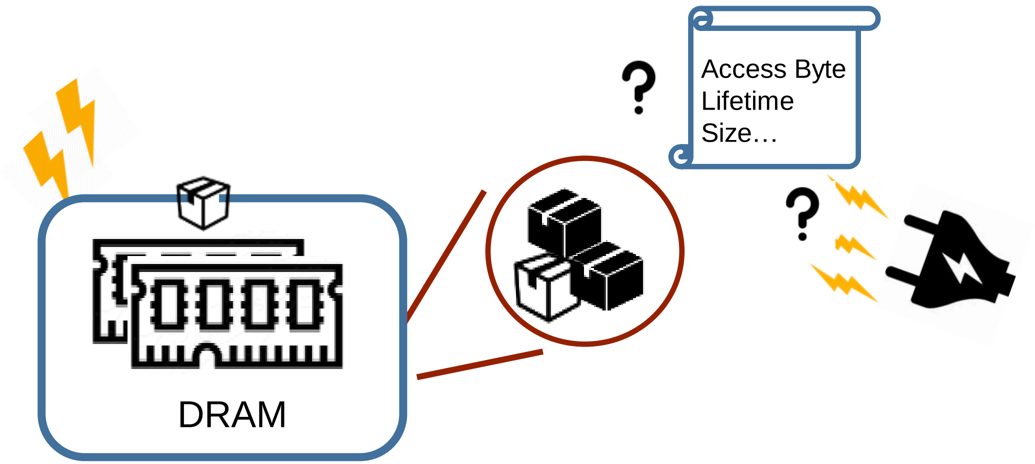
- Linear Scaling Rate (LSR) is one of the solutions to address the energy-efficiency.
  - When predicting the memory object access through LSR, the predicted value and the actual value showed a difference of about **32%**.

ML is used to make accurate predictions and to consider the Memory object of various applications.



Existing Studies: Limitations

- Which memory object pattern should be predicted?
  - Since different objects have different patterns, it should be analyzed the access patterns for each memory object.
  - Among memory object access patterns, a pattern related to energy consumption of memory should be used.



Our Solution: ScaleML

- ScaleML: **ML-based memory object access pattern's scaling rate prediction framework**

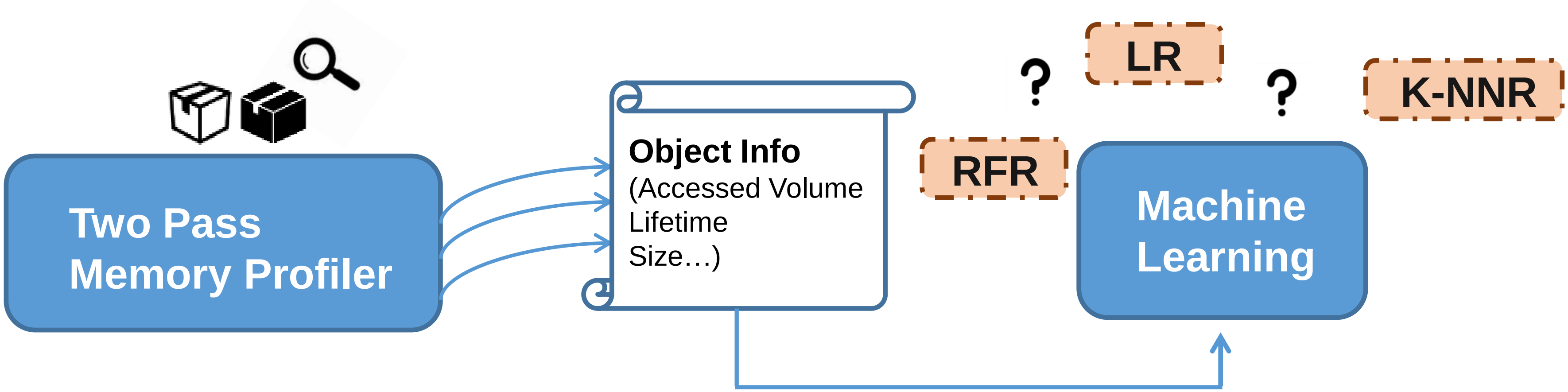
➢ How can we profile the Memory object access pattern?

➢ Which ML model to use?

➢ What input/output fits the Memory object access pattern?

Our Solution: ScaleML

- ScaleML
  - **How can we profile the Memory object access pattern?**
    - Use Two-Pass Memory Profiler
  - **Which ML method to use?**
    - Compare Linear Regression (LR), Random Forest Regression (RFR), and K-Nearest Neighbor (K-NNR) to find the most suitable ML method.
  - **What input/output fits the Memory object Access pattern?**
    - Consider the Accessed volume, Lifetime, Size among various memory object patterns.



Our Solution: ScaleML

- ScaleML
  - **How can we profile the Memory object access pattern?**

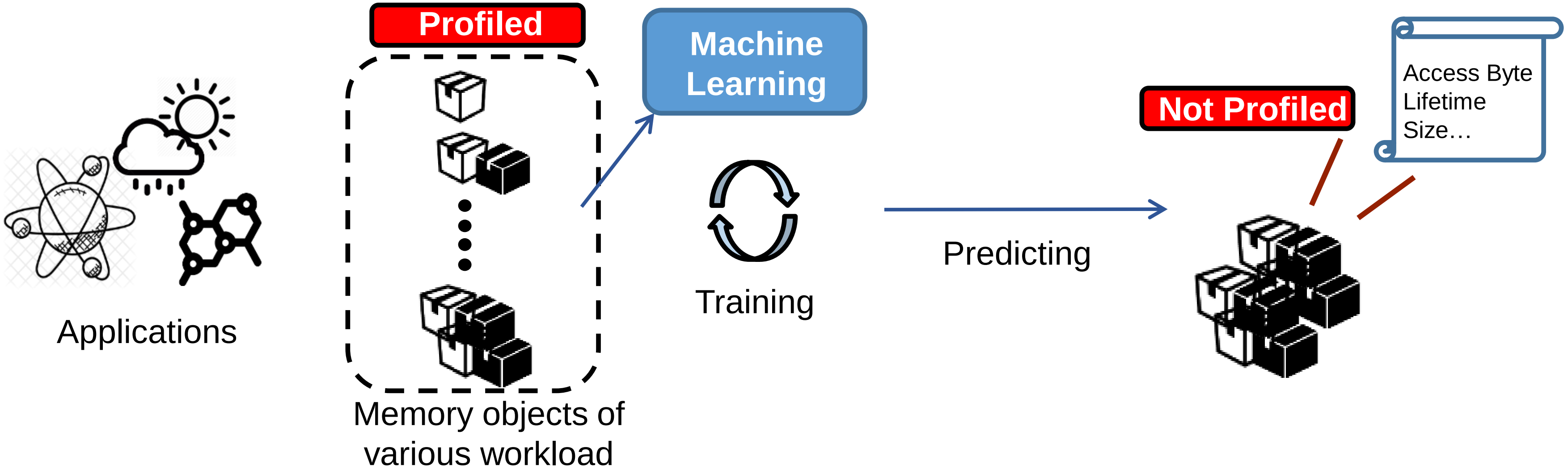
    - Use Two-Pass Memory Profiler

  - **Which ML method to use?**

    - Compare Linear Regression (LR), Random Forest Regression (RFR), and K-Nearest Neighbor (K-NNR) to find the most suitable ML method.

  - **What input/output fits the Memory object Access pattern?**

    - Consider the Accessed volume, Lifetime, Size among various memory object patterns.



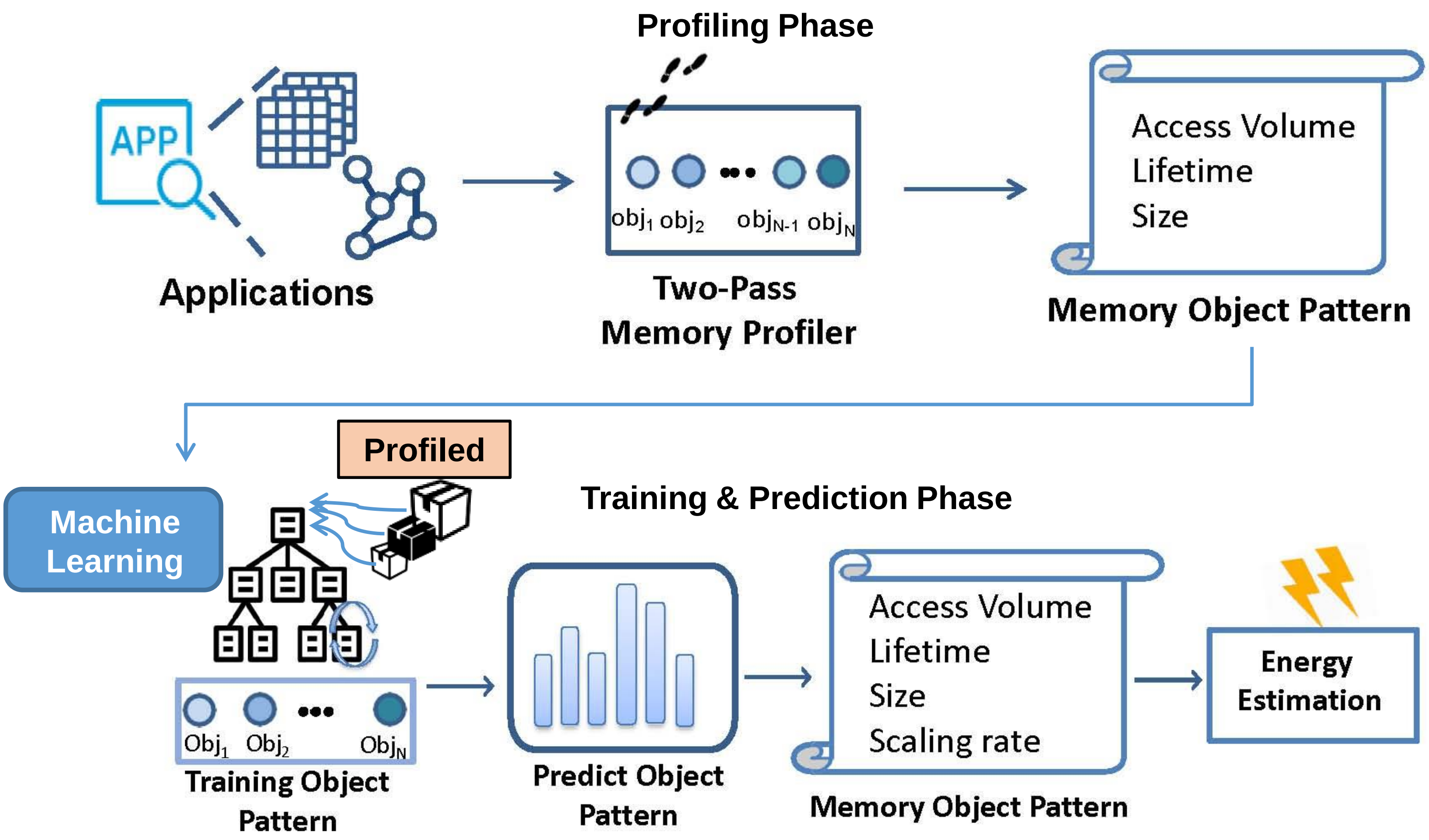
Memory Energy Consumption Model

- Why consider Accessed volume, Lifetime, Size?
  - DRAM sense amplifier acts as row buffer.
    - DRAM operates destructive read.
    - Sense amplifier maintains sensed data and restore it after operation.
  - i-th object energy estimation on DRAM
    - DRAM energy components : Activate and Precharge ($dE_{ACT+PRE}$), Read/Write (dE_{RW}), Refresh (dE_{REF})
- Then, how to estimate the energy consumption?

  - DRAM energy consumption : $DE_i = dE_{ACT+PRE} \cdot AV_i + dE_{RW} \cdot AV_i + dE_{REF} \cdot S_i \cdot T_i$

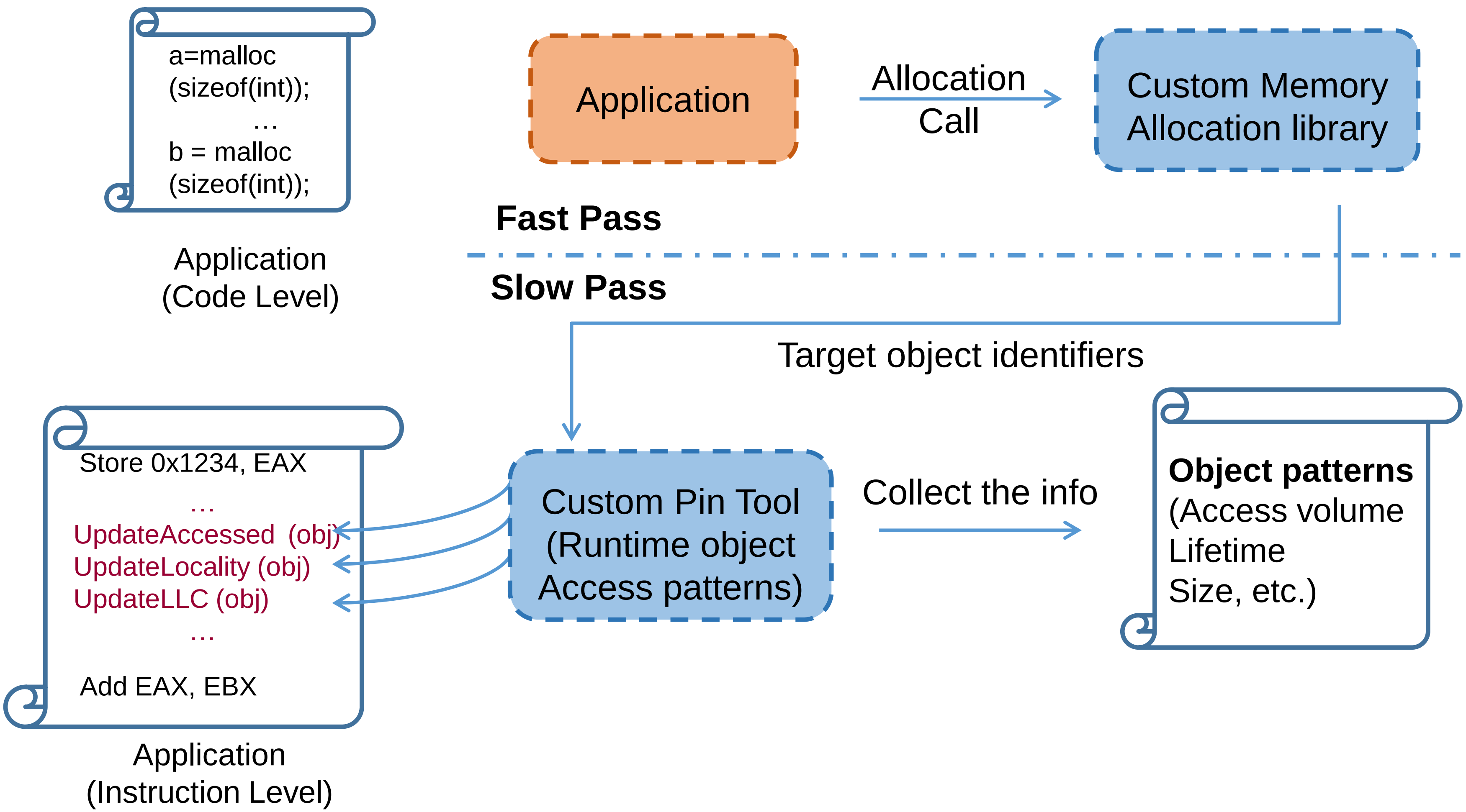
  (AV_i : Accessed volume, S_i : Size, Ti : Lifetime of i-th object)

ScaleML: Overview



SCALEML: Memory Object Profiling

- Two Pass Memory Profiler



SCALEML: Machine Learning Models

- Which of the various ML models should be used?
  - Linear Regression (LR)

    - The accuracy of the prediction is high if Memory object patterns have linear pattern.

  - K-Nearest Neighbor Regression (K-NNR)

    - The accuracy of the prediction is high if Memory object patterns have relationship(linear, exponential, non-linear, etc…).

  - Random Forest Regression (RFR)

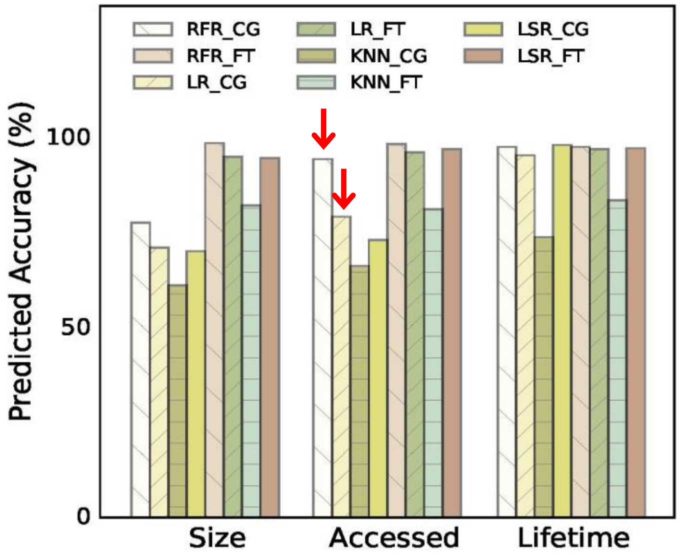
    - RFR can independently learn the change in each access pattern of memory object as the workload changes.
    - Each tree gets random samples that are different from the whole data when it is split, so it has a randomness to avoid over-fitting.

  - Common property of each considered ml models

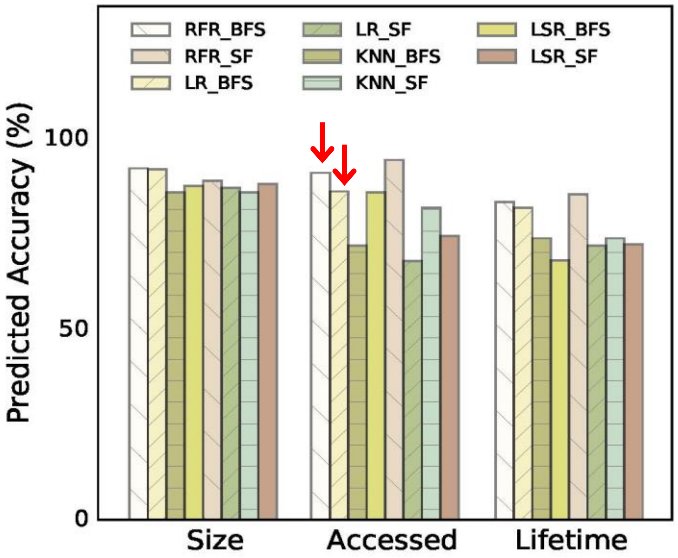
    - Light-weight to execution

Comparison of ML Models

- Comparative analysis of prediction accuracy of various ML models
  - Compared to LR, RFR is up to **16%** higher in the NPB benchmark and up to **6.8%** higher in PBBS benchmark.

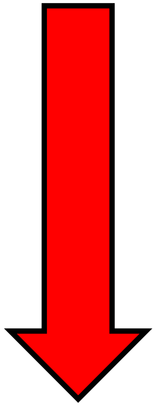


Accuracy of ML Model in NPB benchmark



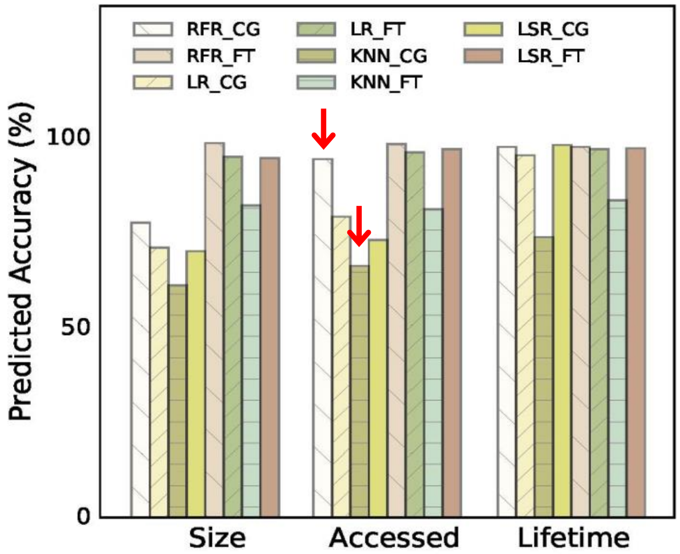
Accuracy of ML Model in PBBS benchmark

Lower is worse

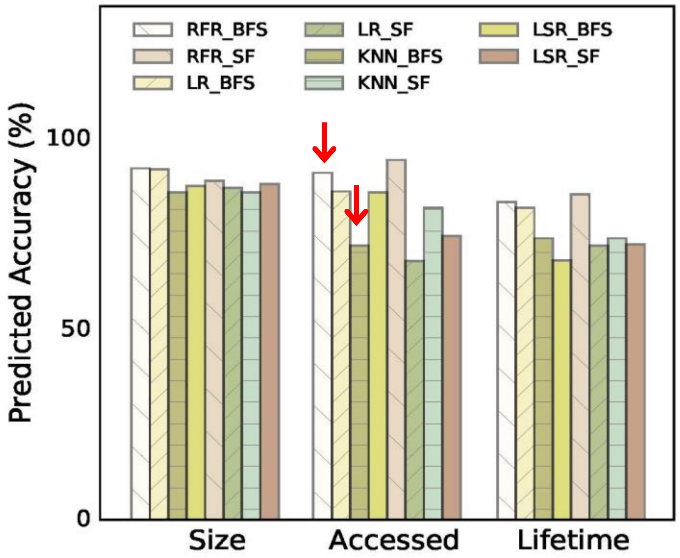


Comparison of ML Models

- Comparative analysis of prediction accuracy of various ML models
  - Compared to K-NNR, RFR is up to **23.6%** higher in the NPB benchmark and up to **19.8%** higher in PBBS.



Accuracy of ML Model in NPB benchmark

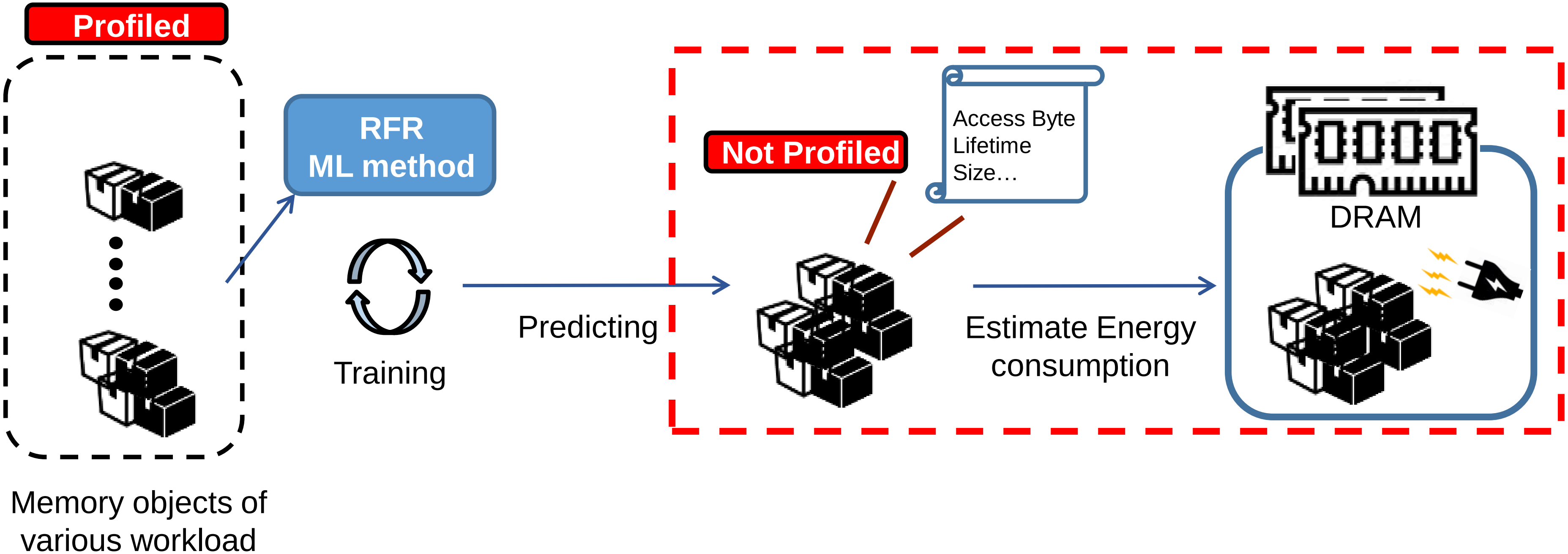


Accuracy of ML Model in PBBS benchmark

Lower is worse

SCALEML: Energy Prediction Phase

- Predict Memory object access patterns & energy consumption
  - Use trained model through RFR
  - Use energy consumption model of DRAM



Evaluation: Experiment Setup

- System Configuration
  - CPU : Intel Core i7 8700 CPU, 6 core, 3.2GHz
  - Main Memory : 16GB DDR4 1340MHz
  - Interface : PCIe 3.0 x8
- Benchmark & Dataset
  - We used two applications from each benchmark NPB, and PBBS
    - NPB Benchmark: CG, FT…
    - PBBS Benchmark: BFS, SF…
  - For each application, we profiled the 4 different workloads to train the ML models by varying the size of workload.
- Training Ratio
  - Training : 80%, Test : 20%

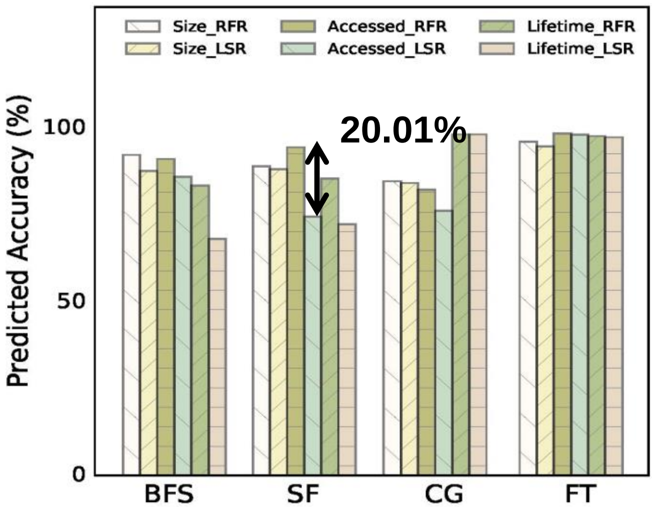
Evaluation: Experiment Setup

- Benchmark workload sizes

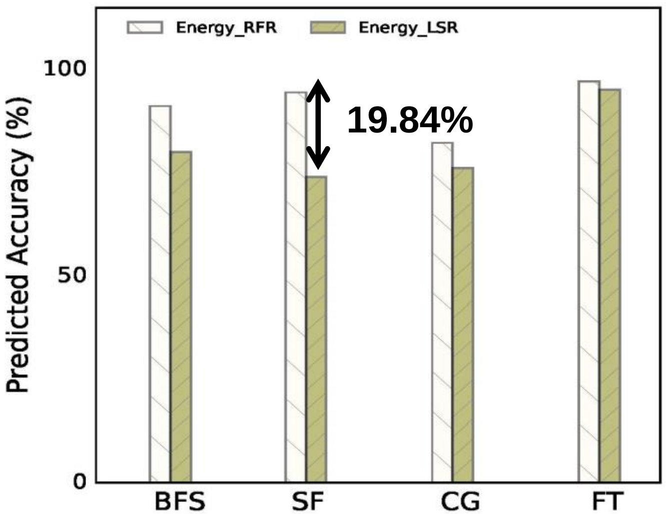
BFS(Vertex)	SF(Vertex)	CG(Num of Row)	FT(Grid Size)
$25 * 10^4$	$25 * 10^4$	$14 * 10^2$	$64 * 64 * 64$
$50 * 10^4$	$50 * 10^4$	$70 * 10^2$	$128 * 128 * 32$
$10 * 10^5$	$10 * 10^5$	$14 * 10^3$	$256 * 256 * 128$
$40 * 10^5$	$40 * 10^5$	$75 * 10^3$	$512 * 256 * 256$

Evaluation : Energy Consumption Comparison

- Energy Consumption Comparison
  - In SF application, the prediction accuracy of RFR model is up to **19.84%** higher than that of the LSR method

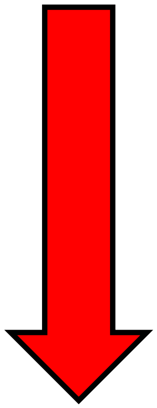


Comparison of prediction accuracy with ML model



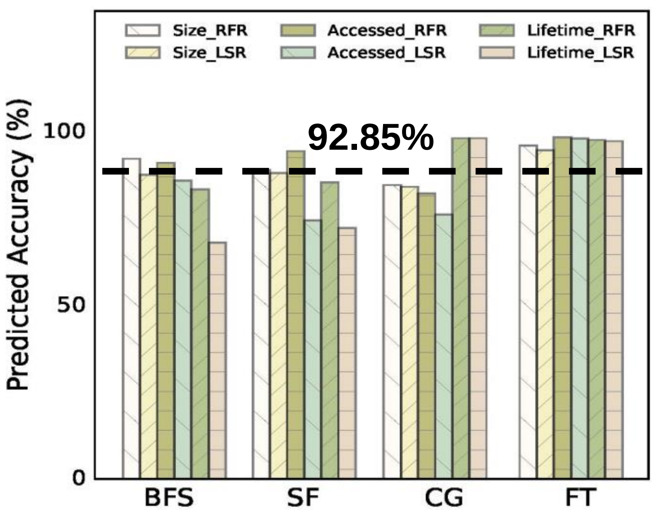
Comparison of energy consumption with ML model

Lower is worse

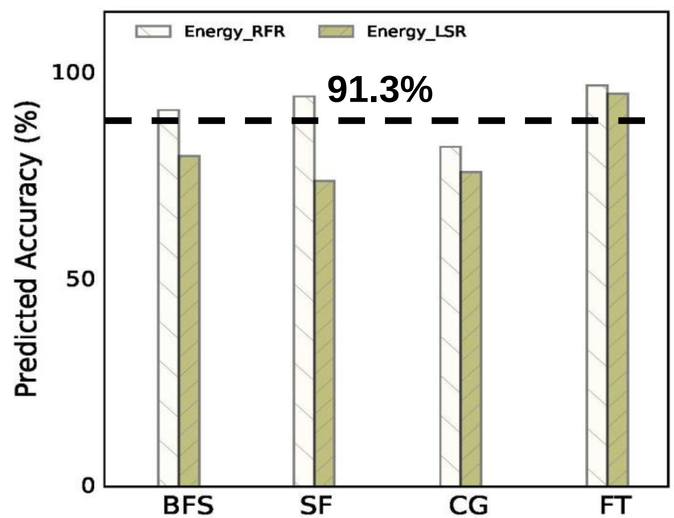


Evaluation : Energy Consumption Comparison

- Energy Consumption Comparison
  - The accuracy of the memory object access pattern predicted using the RFR model is **92.85%** on average, and the accuracy of estimated energy consumption is **91.3%**.

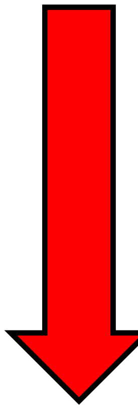


Comparison of prediction accuracy with ML model



Comparison of energy consumption with ML model

Lower is worse



Summary

- **ScaleML is a ML-based memory object access pattern's scaling rate prediction framework in conjunction with energy efficiency estimation.**

  - Bridges the existing prediction accuracy gap by 91.3%
  - Profiling object pattern information that directly affects energy consumption by using the Two Pass Memory profiler
  - Among various ML methods, RFR suitable for memory object pattern prediction is used.

Question?

Parkjoongeon@gmail.com
Laboratory for AI System Software
Sogang University, Seoul,
Republic of Korea