ACM Web Conference 2025

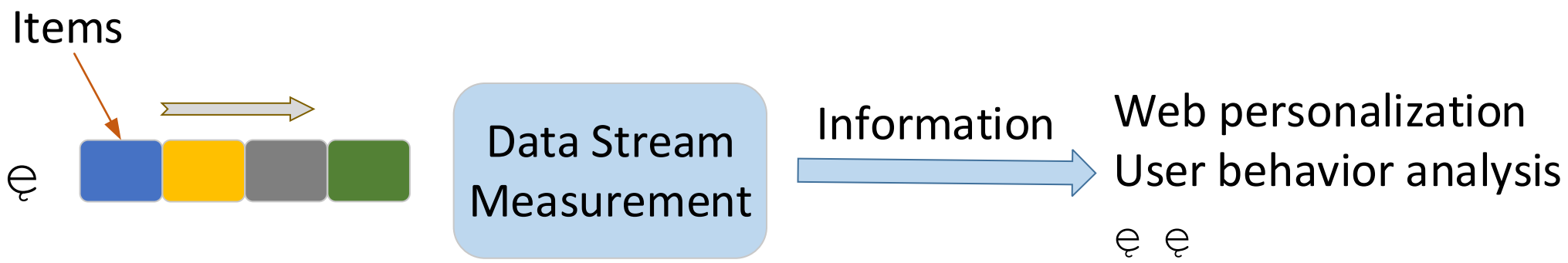
Pontus: A Memory-Efficient and High-Accuracy Approach for Persistence-Based Item Lookup in High-Velocity Data Streams

Weihe Li[1], Zukai Li[1], Beyza Bütün[2], Alec F. Diallo[1], Marco Fiore[2], Paul Patras[1]

[1]University of Edinburgh, Edinburgh, United Kingdom

[2]IMDEA Networks Institute, Madrid, Spain

Data Stream Processing



- User behavior analysis ⟶ important to improve user experience
- Persistently accessed websites (***persistent items***) ⟶ user preferences

Persistence

Appeared, **persistence = 1**

Appeared, **persistence = 2**

Appeared
persistence = 1

Item	Persistence
e_1	2
e_2	1
e_3	1
e_5	1

Challenges

- **Fast processing speed**

e.g. 10 Gb/s data stream: each item every 67 ns

- **Limited fast memory**

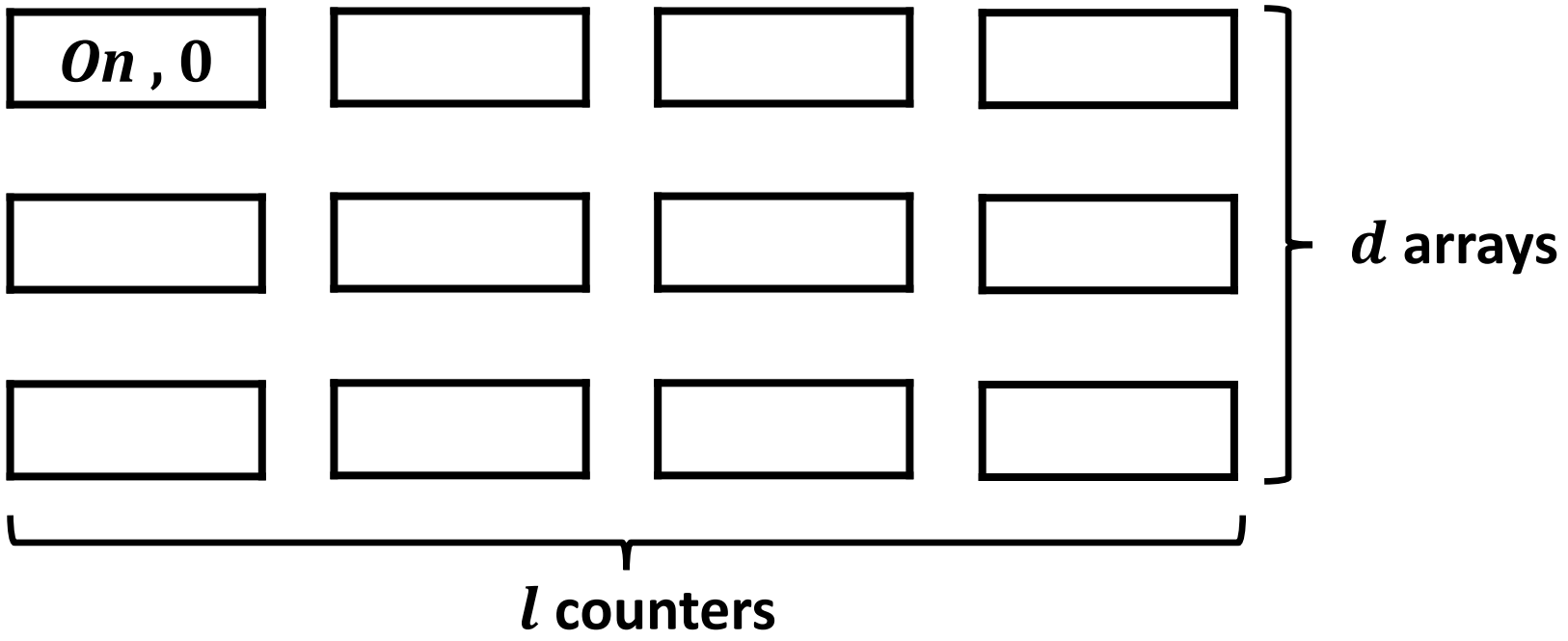
L1 Cache: around 64KB[1]

Infeasible to store information for all items

[1] Li, W. and Patras, P. Tight-sketch: A high-performance sketch for heavy item-oriented data stream mining with limited memory size. ACM CIKM 2023.

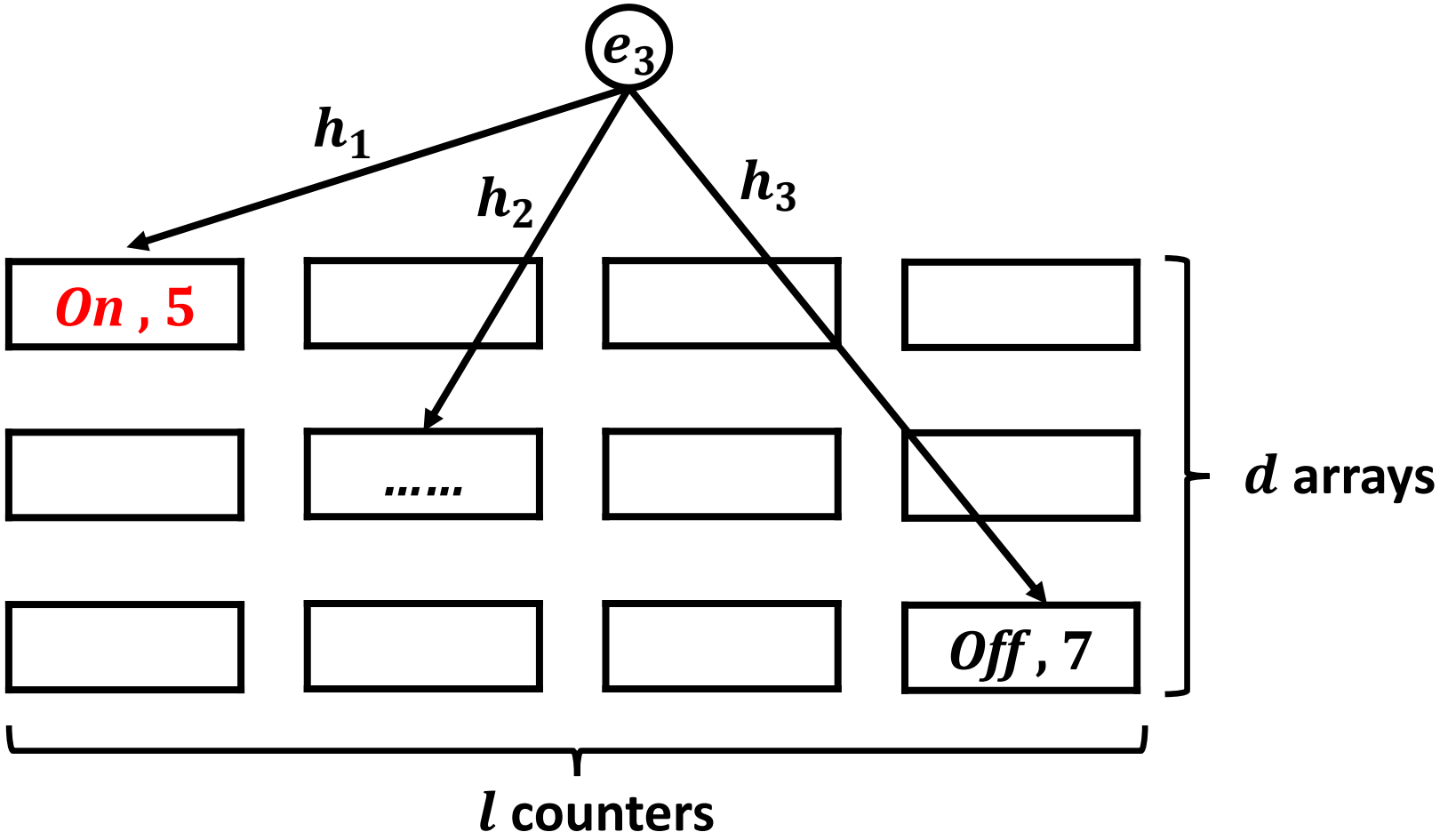
Sketches

- **Sketches**: **compact data structure by hashing**
  - Idea: hash data into limited space

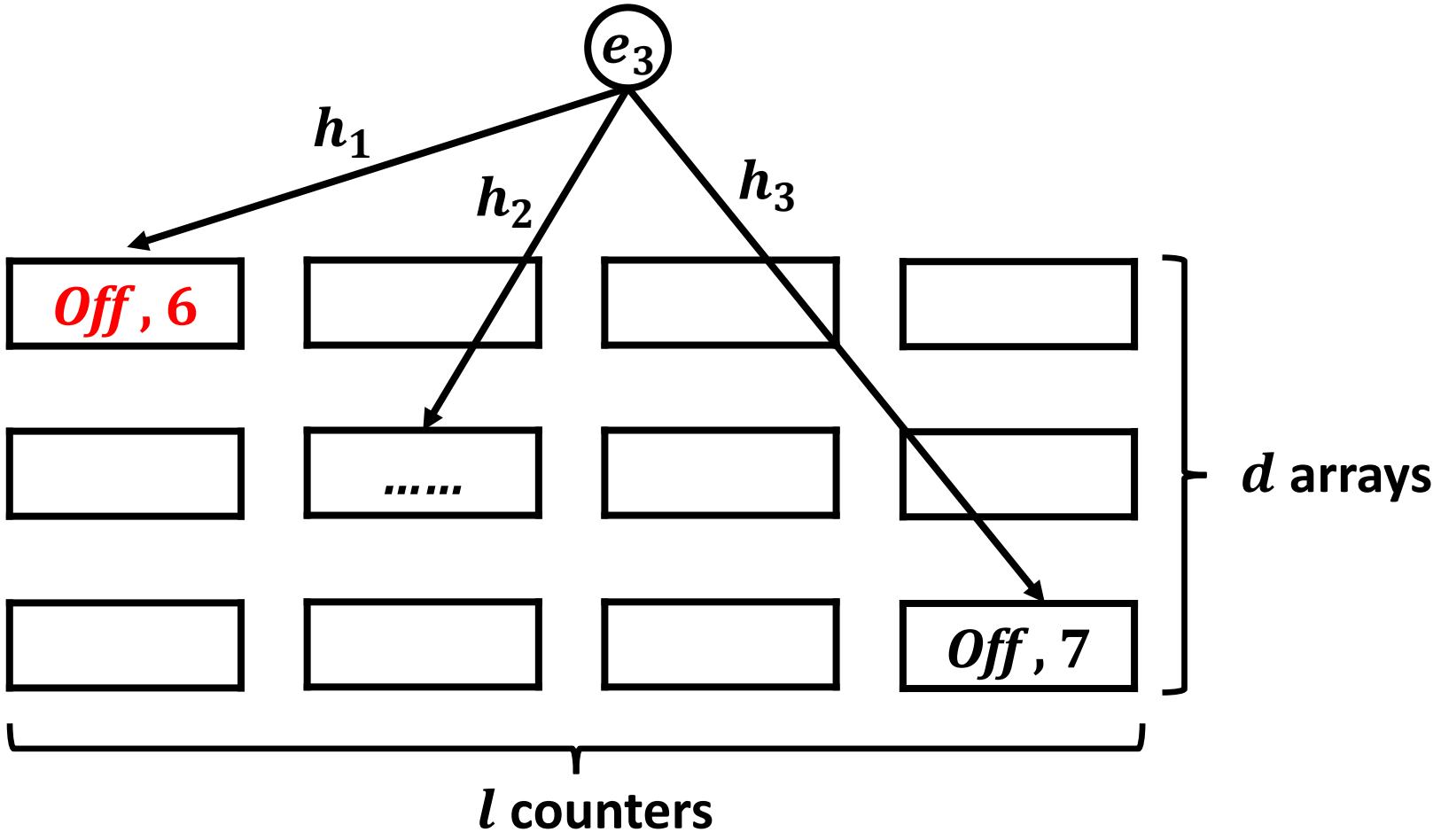


[1] Zhang, Yinda, et al. "On-off sketch: A fast and accurate sketch on persistence." *Proceedings of the VLDB Endowment* 14.2 (2020): 128-140.

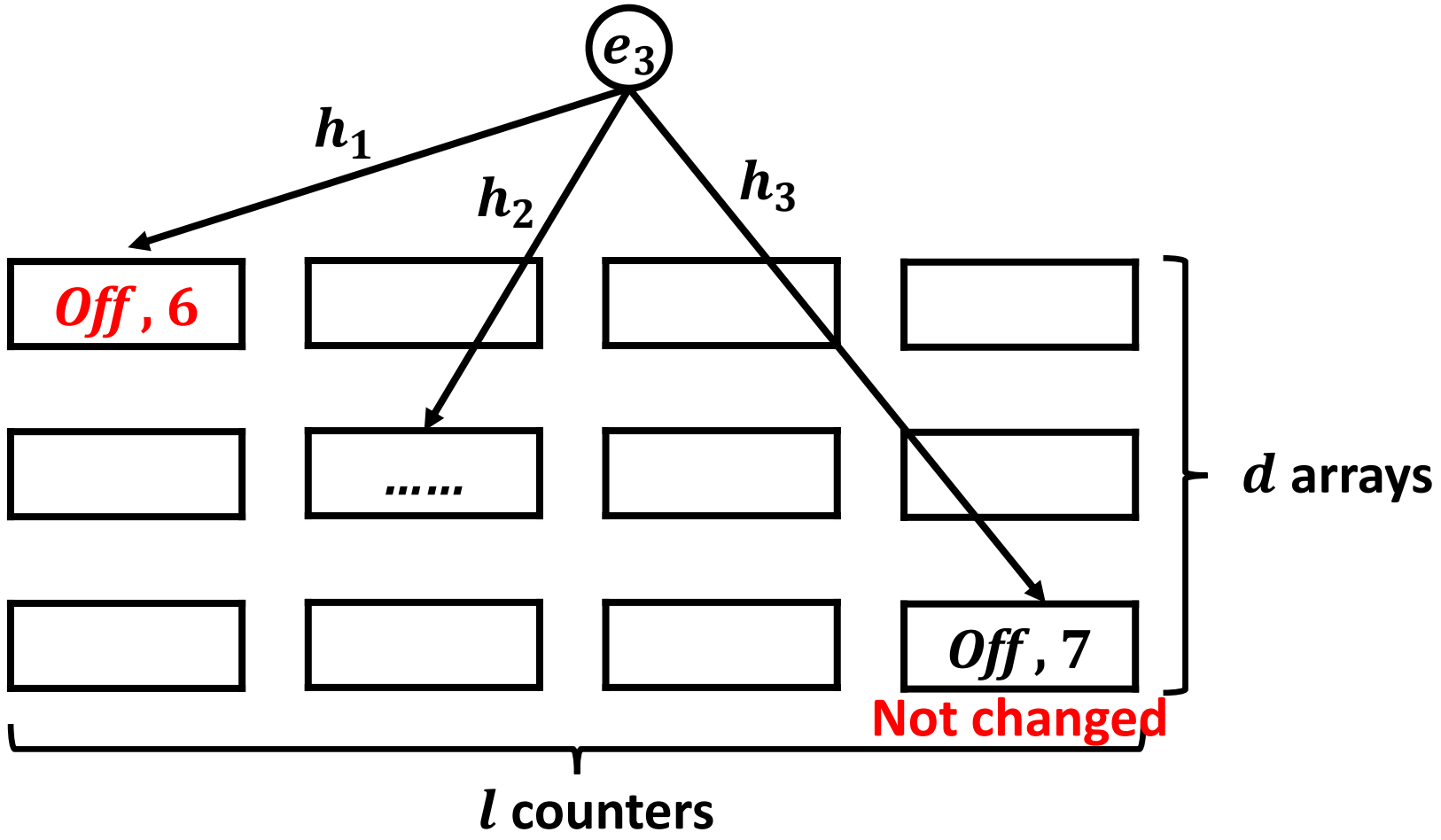
Sketches



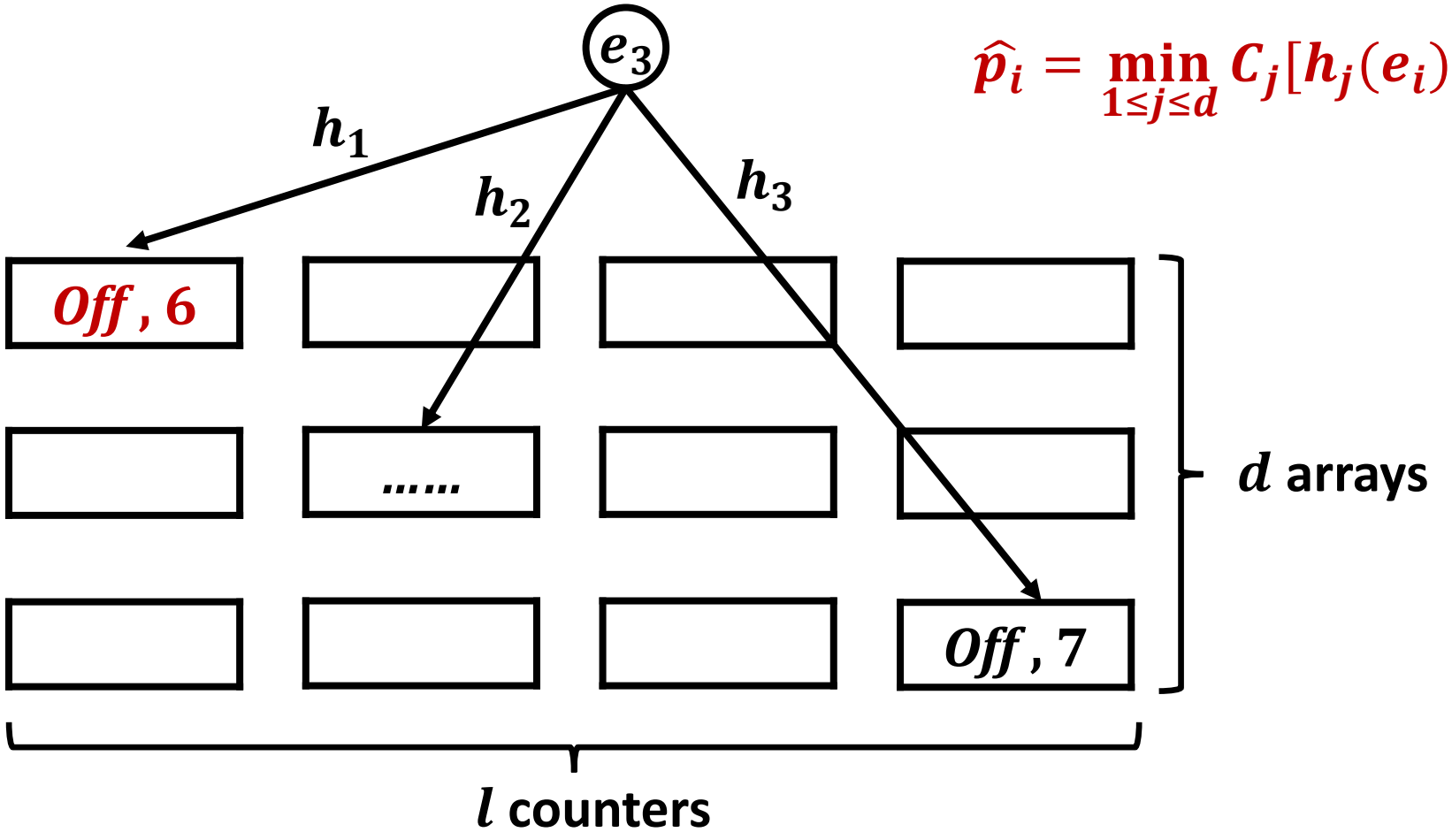
Sketches



Sketches



Sketches

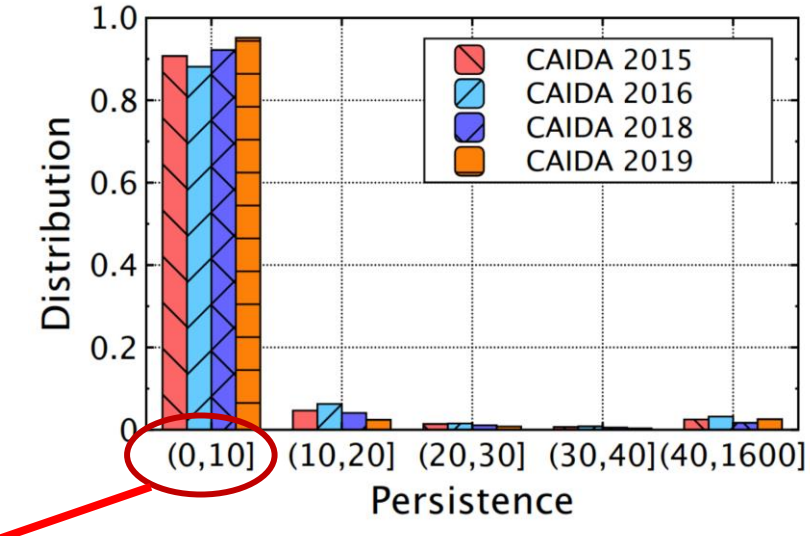


$$\hat{p}_i = \min_{1 \le j \le d} C_j[h_j(e_i)]$$

Limitations of Existing Sketch Methods

Low detection accuracy under limited memory budgets

Persistent flows being evicted from the bucket by non-persistent ones due to the **highly skewed** traffic distribution.

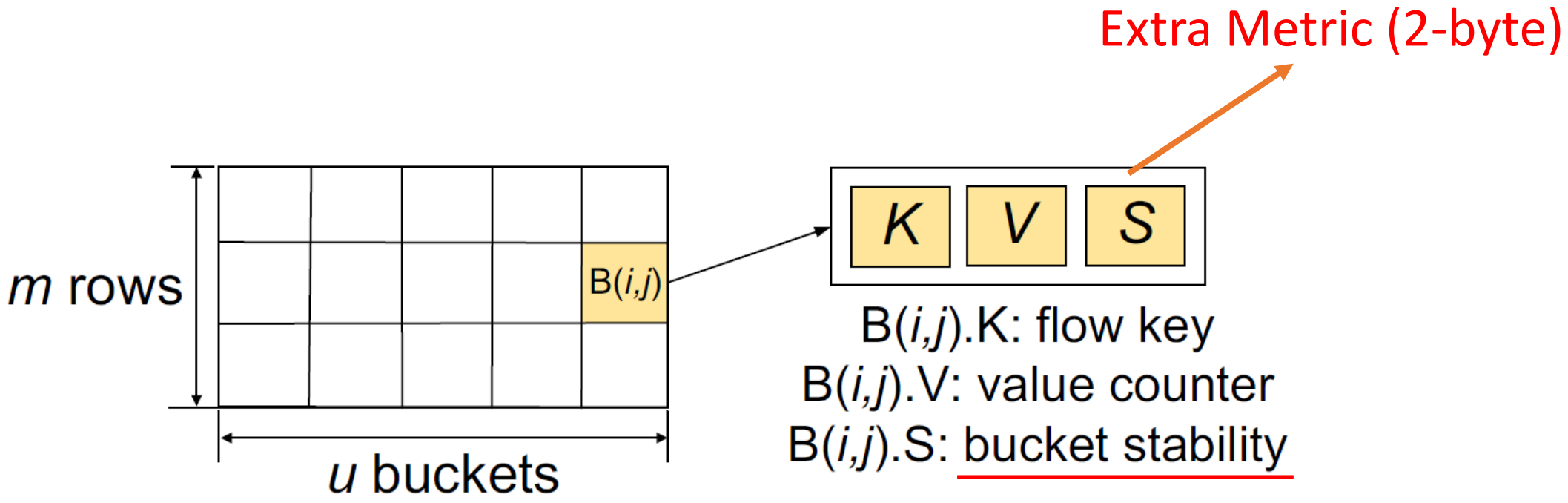


More than 85%

Limitations of Existing Sketch Methods

Low Memory Efficiency

Track multiple features per item to better protect persistent items.

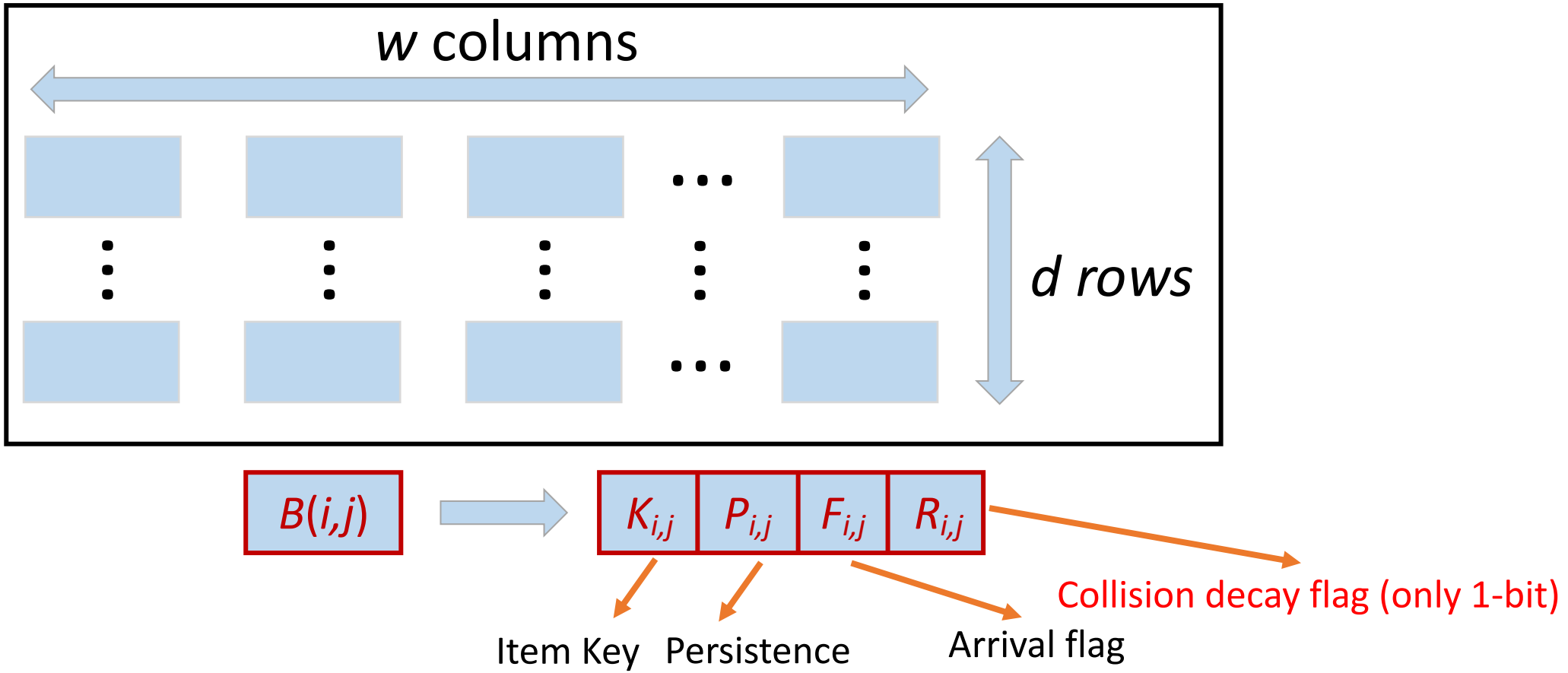


[2] Li, W. and Patras, P. Stable-sketch: A versatile sketch for accurate, fast, web-scale data stream processing. ACM WWW 2024.

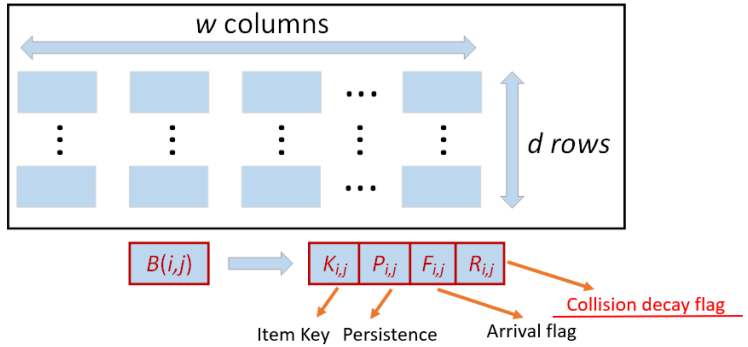
Our Contributions

- **Pontus**
  - A novel method for persistent item lookup
  - High accuracy, high memory-efficiency and fast processing speed
  - Deployable on the practical hardware, Tofino programmable switch

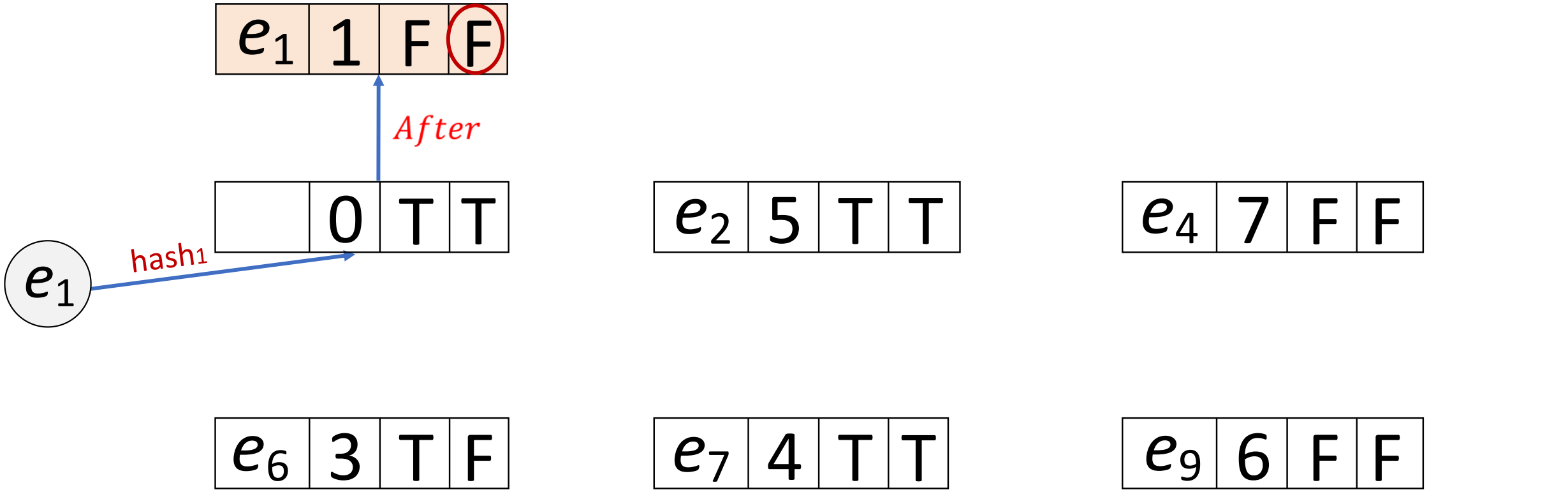
Data Structure



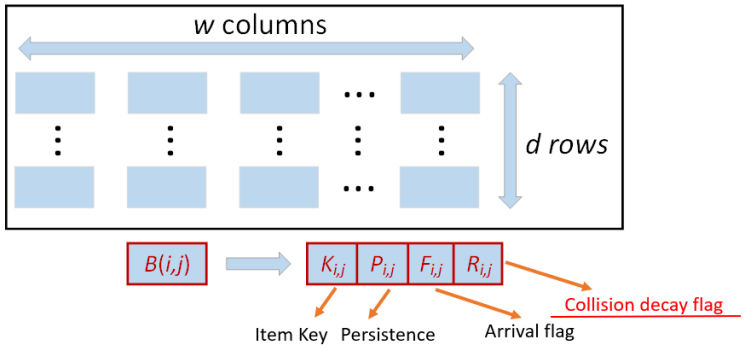
Update



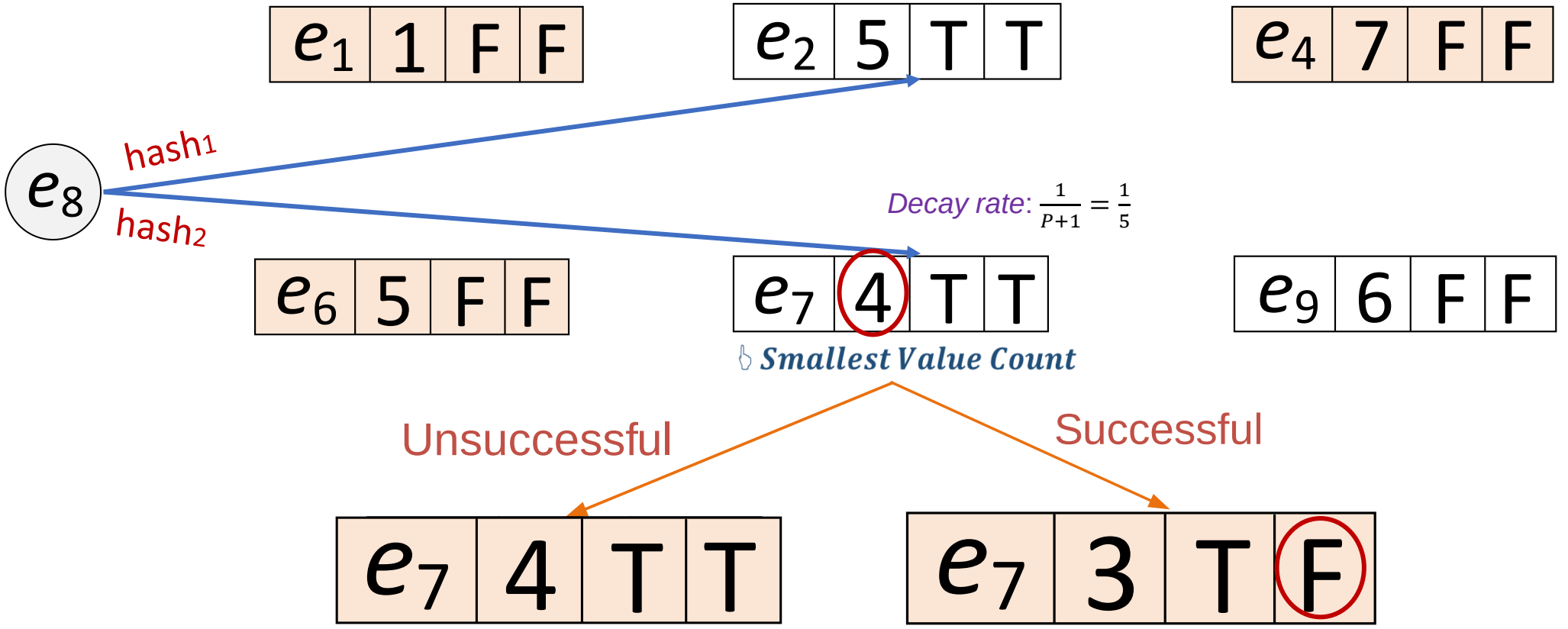
Case 1:



Update

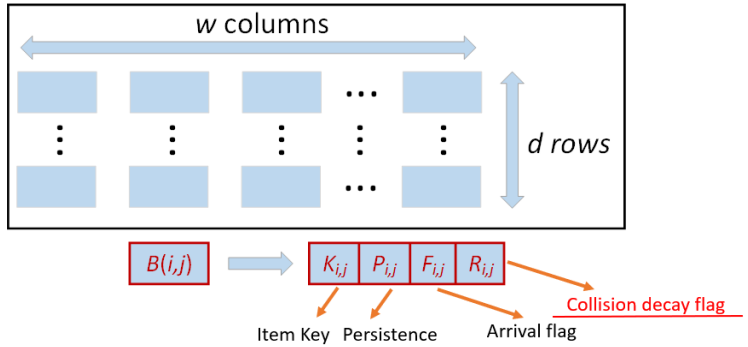


Case 2:

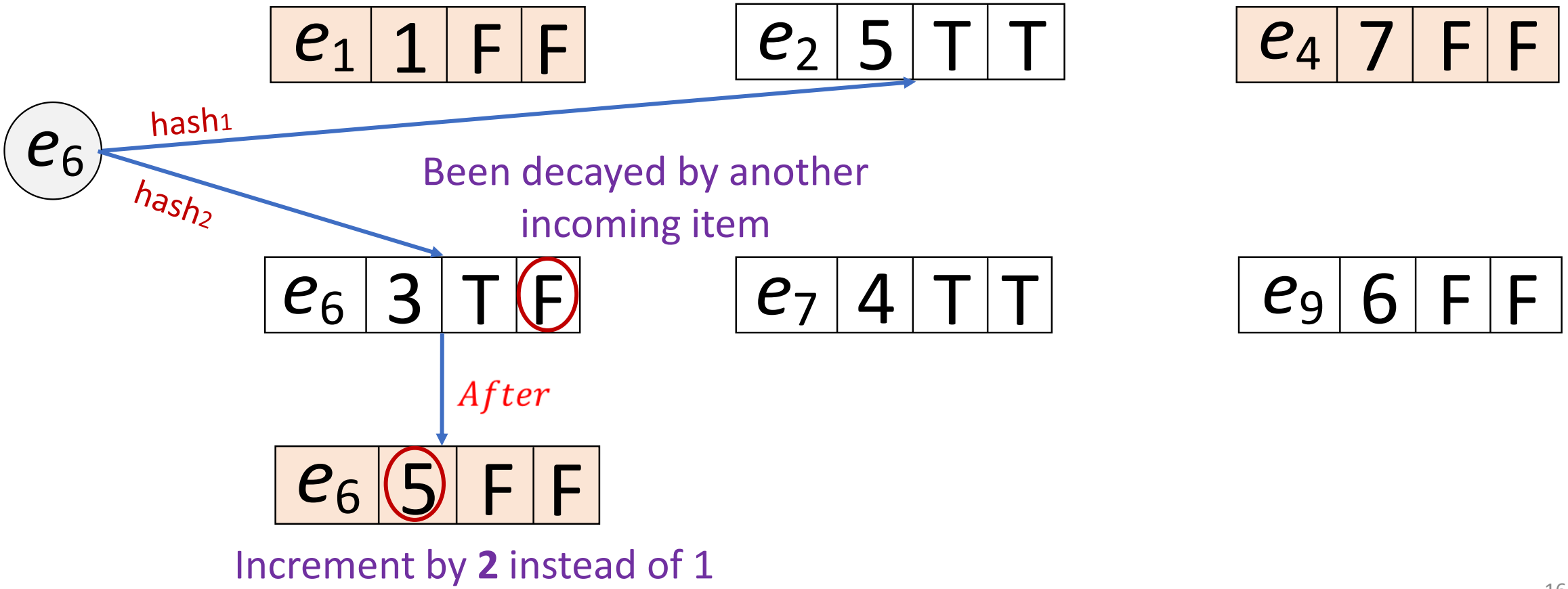


The incoming item can replace the tracked item only if its counter has decayed to **zero**.

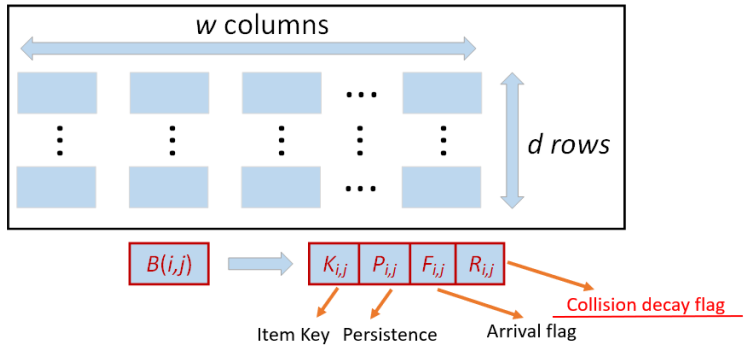
Update



Case 3:



Query



Only **a single scan** of all buckets is required to determine which bucket contains a value higher than the predefined threshold.

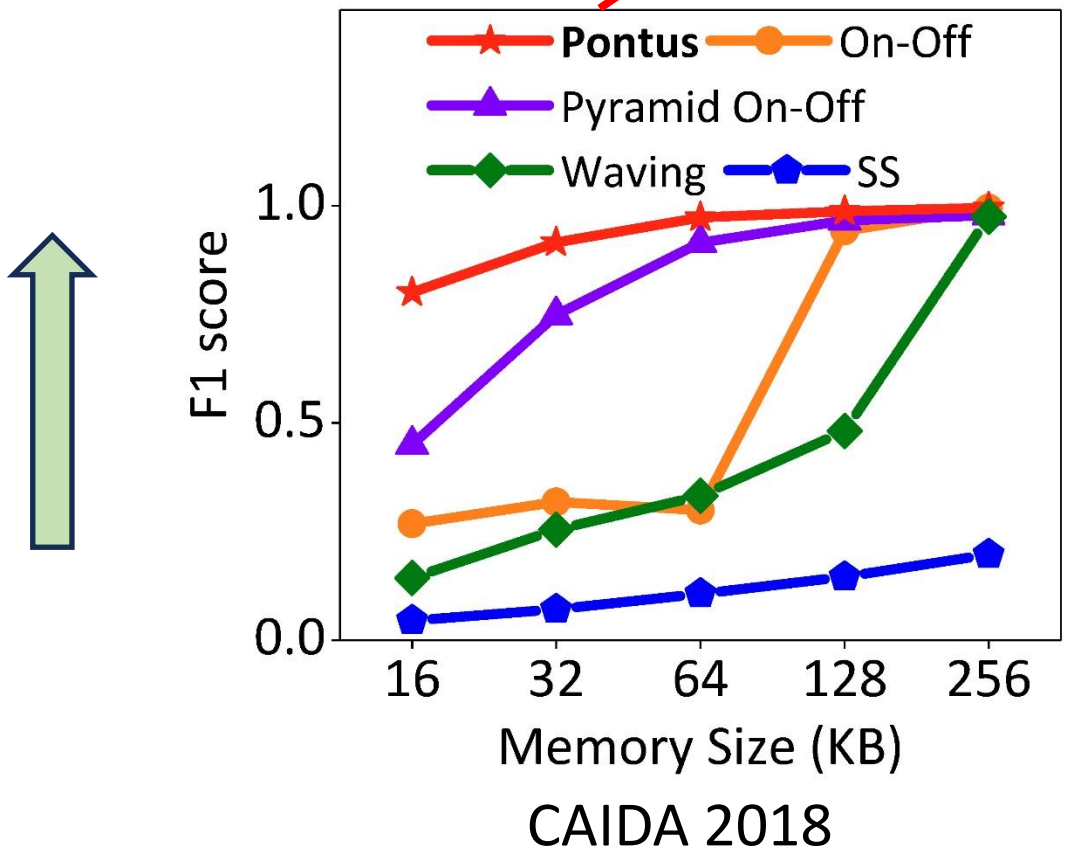
Evaluation

- **Software** -- CPU Platform (Intel(R) Core(TM) i5-1135G7 @ 2.40GHz processor, C++)
- **Hardware** -- Tofino Switch (P4)
- **Traces** -- CAIDA 2018 and 2019[3]

[3] CAIDA. Anonymized Internet Traces. https://catalog.caida.org/dataset.

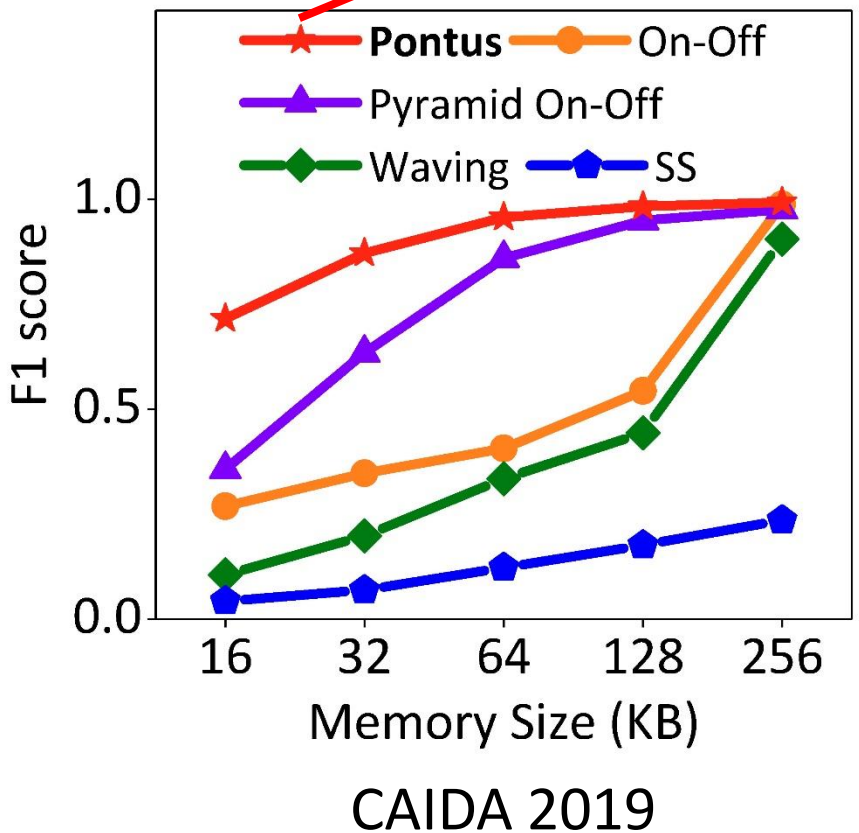
Evaluation – **Accuracy (CPU)**

Highest F1 score!
15.1% higher than Pyramid On-Off



CAIDA 2018

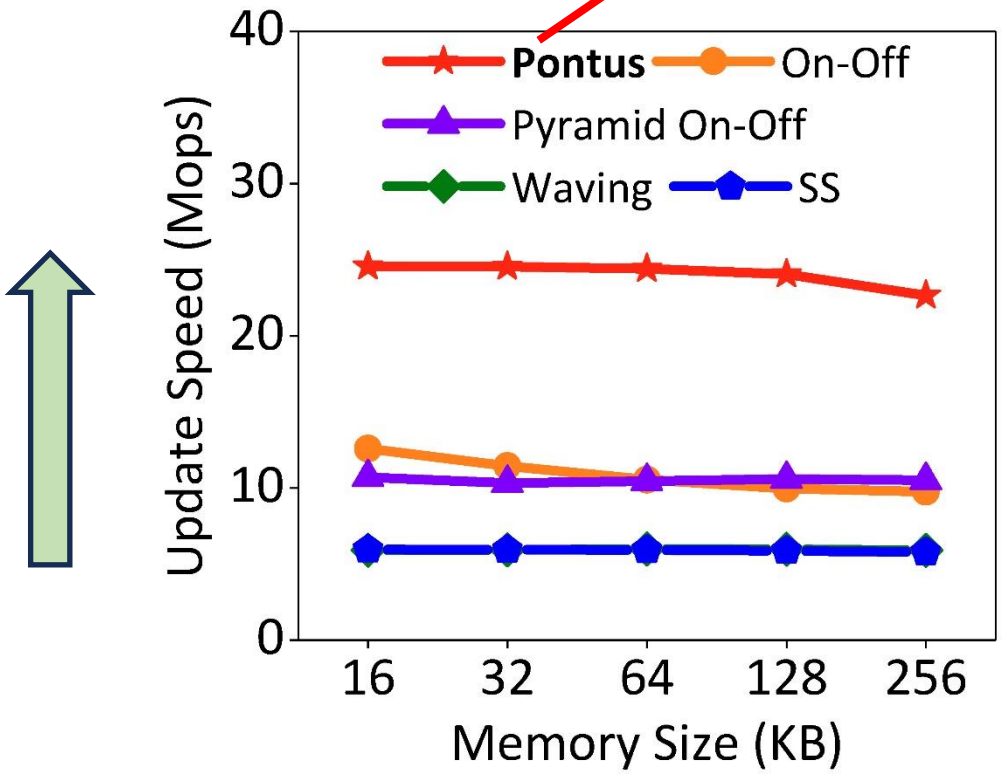
Highest F1 score!
19.7% higher than Pyramid On-Off



CAIDA 2019

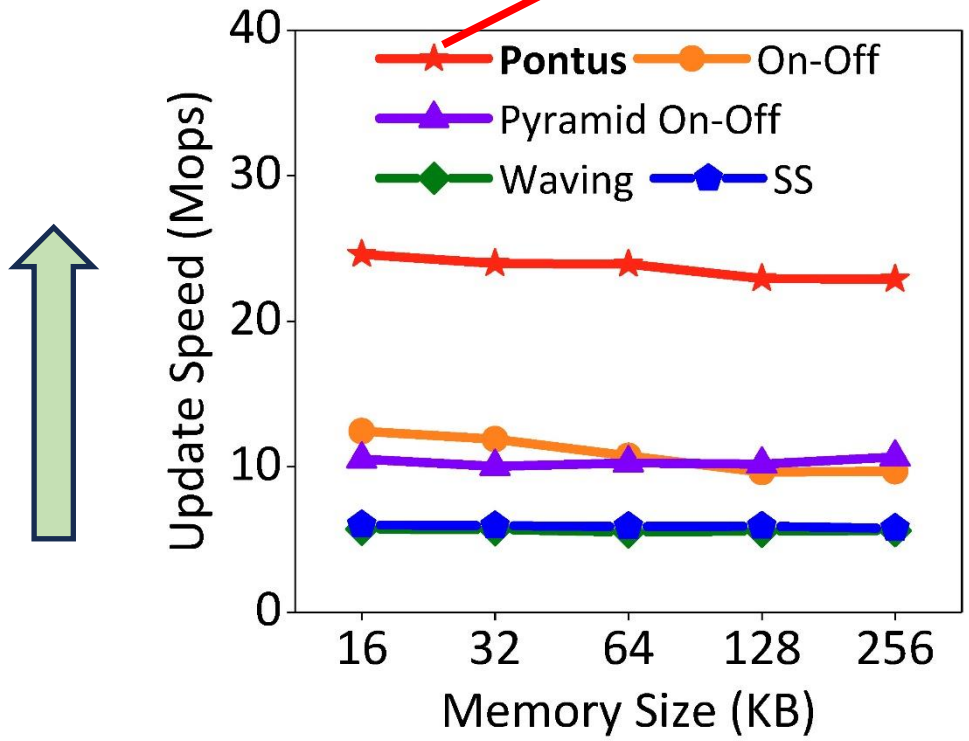
Evaluation – **Speed (CPU)**

Higher Speed!
121.4% higher than On-Off



CAIDA 2018

Higher Speed!
117.3% higher than On-Off



CAIDA 2019

Evaluation – **Resource Usage (Tofino)**

Resource	Usage	Resource	Usage
Hash Bit	5.7%	Match Crossbars	4.6%
Gateways	16.7%	Logical Table ID	21.9%
VLIW Instruction	7.3%	SRAM	4.3%
Total Average		8.2%	

Limited Overhead

- **More results**
  - Formal analysis
  - Evaluation on more tasks, like persistence estimation
  - More results on the CPU platform and Tofino switch

- **Code**: https://github.com/Mobile-Intelligence-Lab/Pontus

This research was supported by the SNS JU and the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101139270 (ORIGAMI). Beyza Bütün is a Comunidad de Madrid predoctoral fellow (PIPF-2022/COM-24867). Weihe Li was partially supported by Cisco through the Cisco University Research Program Fund (Grant no. 2019-197006).