Footprint Modeling of Cache Associativity and Granularity

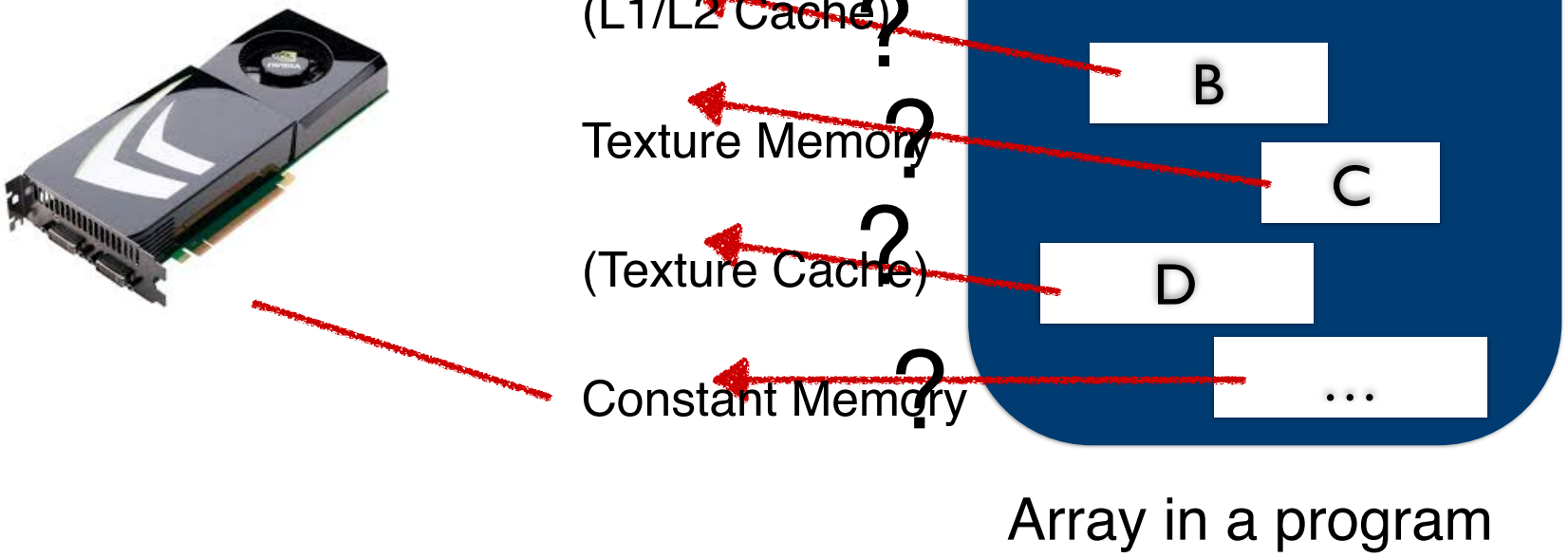
Hao Luo[1] (UR), Guoyang Chen[2] (NCSU), *Fangzhou Liu* (UR), Pengcheng Li[1] (UR), Chen Ding (UR), Xipeng Shen (NCSU)

University of Rochester, Rochester, NY
North Carolina State University, Raleigh, NC

1. Now at Google Inc. This work was done when the student was a graduate student
2. Now at Alibaba Group US Inc. This work was done when the student was a graduate student

Motivation

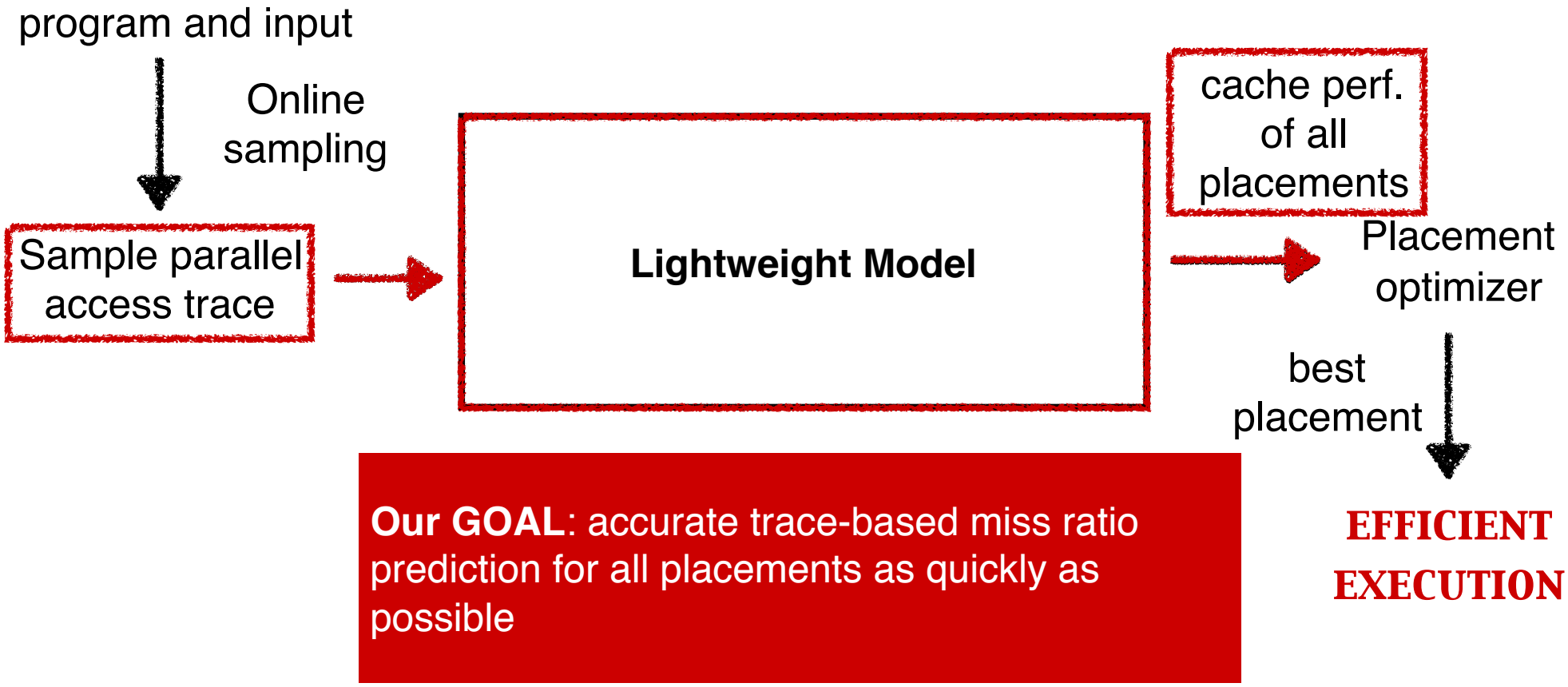
Data placement problem



3X performance difference

Motivation

PORPLE[1] -Online Optimization of Data Placement



1. Guoyang Chen, Bo Wu, Dong Li and Xipeng Shen. 2014. PORPLE: An Extensible Optimizer for Portable Data Placement on GPU. *In Proceedings of MICRO.*

Motivation

- Data Non-uniformity
  - A program uses multiple arrays, which have different access frequency and locality.
- Uneven Address Mapping
  - Data addresses may be mapped unevenly to cache sets.
  - Cache can be reconfigured to different capacity
    - Which changes address mapping
- Dual-grained Cache
  - Data was fetched and managed in different granularity.

```
1 for (i = 1; i < 1025; i++) {
2     for (j = 1; j < 1025; j++) {
3         B[i][j] = A[i][j]
4                 + A[i][j+1]
5                 + A[i][j-1]
6                 + A[i-1][j]
7                 + A[i+1][j];
8     }
9 }
```

Agenda

- Footprint Metric
- Three footprint models
- Evaluation

Background

Footprint Metric

- **Trace** is a series of memory accesses (interleaved)

Example: a b c a c d

- **Working Set Size** is the number of distinct elements within a window (fixed-length time interval)

- **Footprint** is the average number of distinct items accessed in windows of length “x”

$$fp(x) = \frac{1}{N - x + 1} \sum_{t=x}^{N} \omega(t, x)$$

Background

Footprint Metric

Time:	1	2	3	4	5	6
Trace:	a	b	c	a	c	d

AVG(3 3 2 3) = 2.75

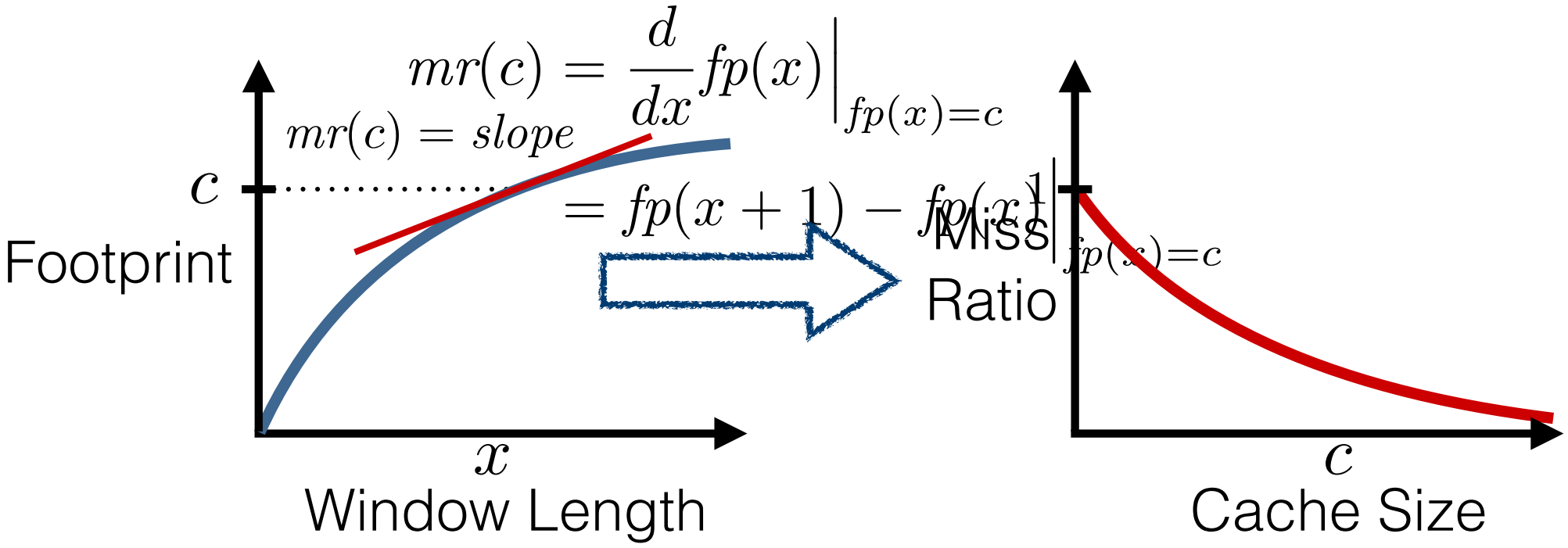
- **Footprint** is the average number of distinct items accessed in windows of length “x”

Window length	1	2	3	4	5	≥ 6
Footprint	1	2	2.75	3	3.5	4

Background

Denning-HOTL Conversion

- The **Cache Miss ratio** can be computed by the derivative of footprint function.



1.Xiaoya Xiang, Chen Ding, Hao luo, and Bin Bao. 2013. HOTL: a higher order theory of locality. *In Proeedings of ASPLOS*. pp 343-356

Agenda

- Footprint Metric

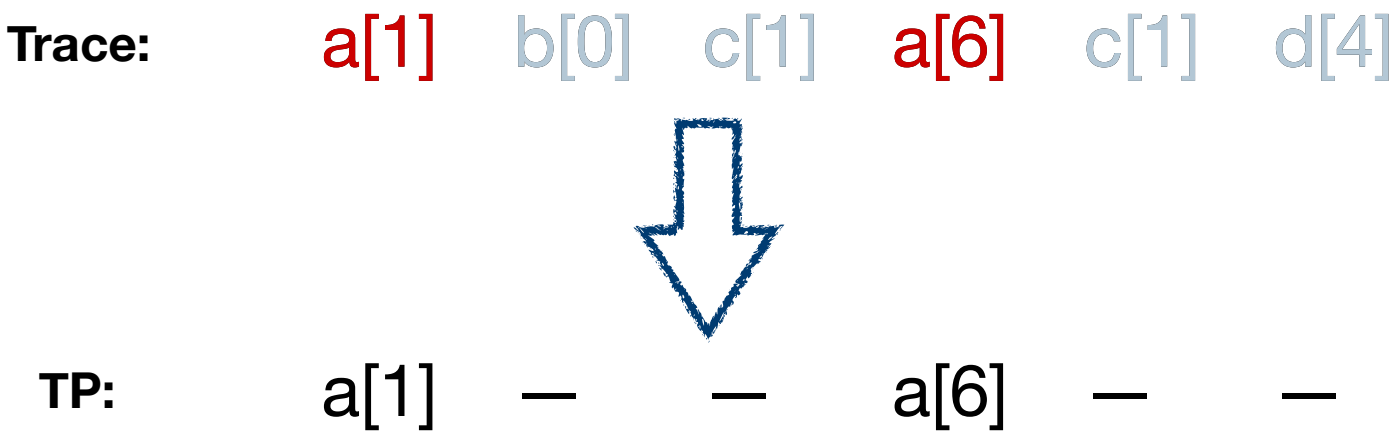
- Three new footprint models
  - **Nonuniform data locality -> Partial Footprint (This talk)**
  - Uneven address mapping -> Mapped Footprint (See paper)
  - Dual-grained Cache -> Dual-grained Footprint (See paper)

- Evaluation

Partial Footprint

Time-preserving trace (TP)

- **Time-preserving trace** is a copy of the original trace with 0 or more accesses replaced with placeholders.



Partial Footprint

Partial Footprint Model

- Let s_i be a TP trace, its **Partial Footprint** is defined as:

$$pfp(x) = \frac{1}{N - x + 1} \sum_{t=x}^{N} \omega(s_i, t, x)$$

TP: a[1] — — a[6] — —

AVG(1 1 1 1) = 1

- Let $\Omega = \{d_i\}$ be the set of arrays used by the program. We can define the partial footprint of the TP trace that preserves array d_i, denoted as $pfp(d_i, x)$.

Partial Footprint

Partial Footprint Model

TP: a[1] — — a[6] — —

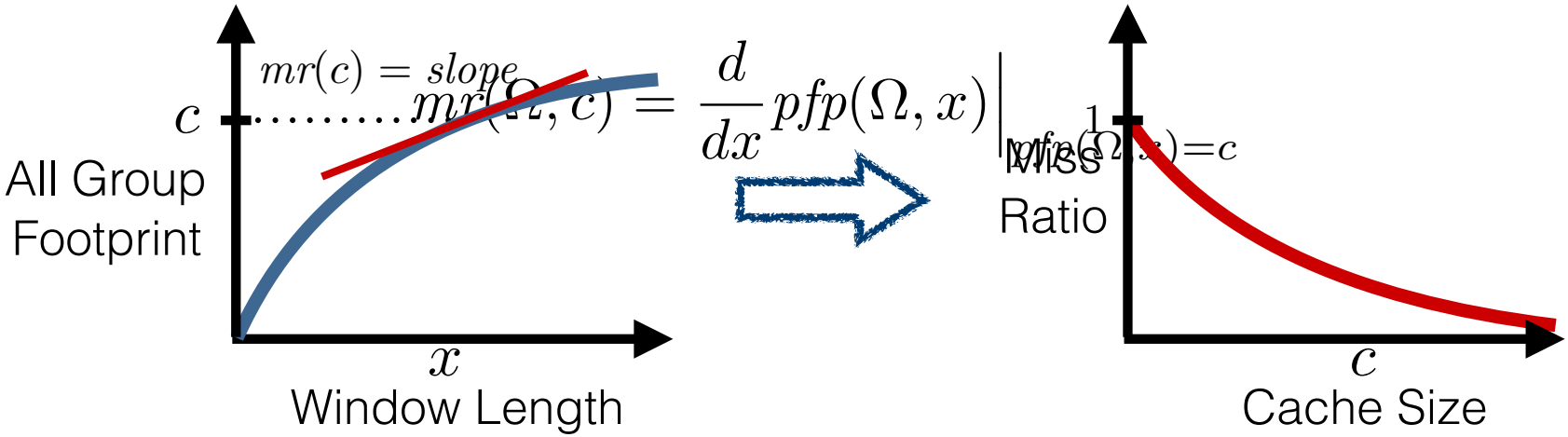
Window length	1	2	3	4	5	≥ 6
Partial Footprint	1/3	3/5	1	4/3	3/2	2

Cache Size	1	2	3	4	5	≥ 6
Miss Ratio	4/15	2/5	1/3	1/6	1/2	0

Partial Footprint

All-group Cache Performance

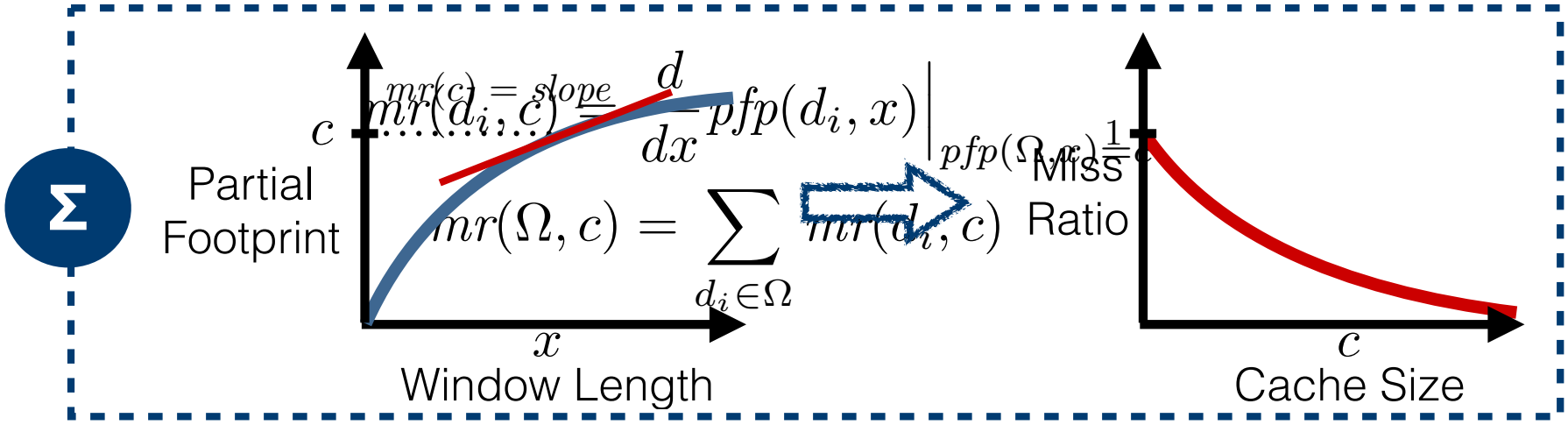
- The all-group cache miss ratio can be calculated in two ways
  1. Compute the all-group cache miss ratio using the Denning-HOTL conversion.



Partial Footprint

All-group Cache Performance

- The all-group cache miss ratio can be calculated in two ways
  1. Compute the all-group cache miss ratio using the Denning-HOTL conversion.
  2. Compute the cache miss ratio for each data item and then sum them up.



Partial Footprint

All-group Cache Performance

- The all-group cache miss ratio can be calculated in two ways
  1. Compute the all-group cache miss ratio using the Denning-HOTL conversion.

$$mr(\Omega, c) = \frac{d}{dx} pfp(\Omega, x)\Big|_{pfp(\Omega,x)=c}$$

  2. Compute the cache miss ratio for each data item and then sum them up.

$$mr(d_i, c) = \frac{d}{dx} pfp(d_i, x)\Big|_{pfp(\Omega,x)=c}$$

$$mr(\Omega, c) = \sum_{d_i \in \Omega} mr(d_i, c)$$

- The equality of this two solutions shows the ***Composition Invariance***.

Agenda

- Footprint Metric
- Three footprint models
- Evaluation

Evaluation

Setup

- 13 benchmarks from SHOC and CUDA Code Samples
- Traces were collected on Tesla M2075 GPU
- 48KB 6-way L1 Cache and 12 KB, 8-way Texture Cache
- Compared our models with Simulation (Baseline), and two modeling techniques, Set-RD[1] and PORPLE[2]

1. Rathijit Sen and David A. Wood, 2013, Reuse-based online models for caches. *In Proceedings of SIGMETRICS*. pp279-292
2. Guoyang Chen, Bo Wu, Dong Li and Xipeng Shen. 2014. PORPLE: An Extensible Optimizer for Portable Data Placement on GPU. *In Proceedings of MICRO.*

Evaluation

Experiment Result

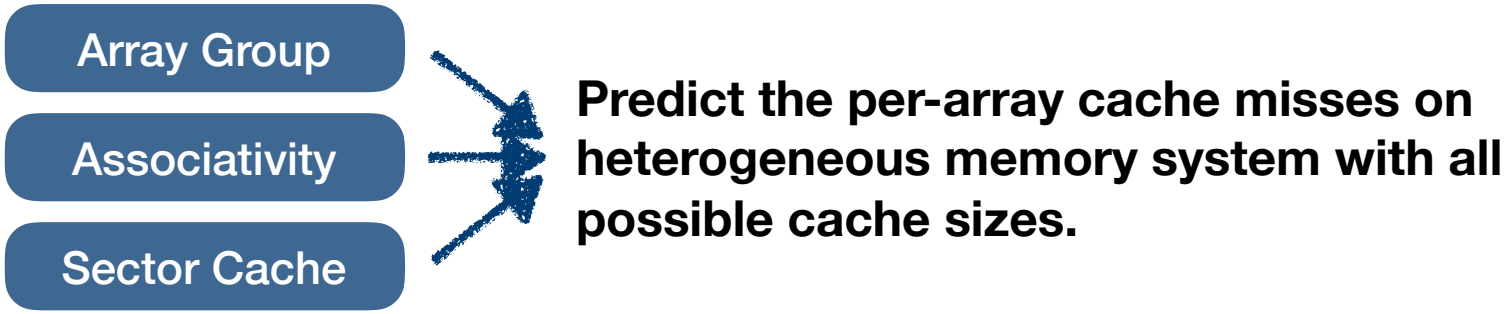
	Simulator	Footprint	PORPLE	Set-RD
Accuracy	100%	99.2%	96.6%	100%
Cost	14.0	1.6	4.7	524

PORPLE is **111** times faster than Set-RD, and Footprint analysis is **3** times faster than PORPLE.

The arithmetic average across the 13 benchmarks is **0.0%** for Set-RD, **3.4%** error for PORPLE and **0.8%** for Footprint, which is **4.3 times** smaller error than PORPLE.

Summary

- Three models solve Data Non-uniformity, Uneven Address Mapping and Dual Granularity.



- Experiment results show lower overhead and higher accuracy.
  - ❖ 327 times faster than Set-RD, 3 times faster than PORPLE
  - ❖ 99.2% accurate, 4.3 times lower error than PORPLE

Thanks for Listening
Any Questions?