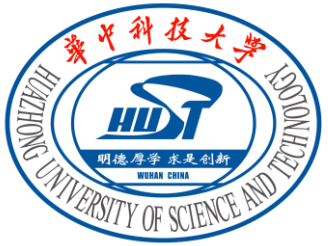
Efficiently Detecting Concurrency Bugs in Persistent Memory Programs

Zhangyu Chen, Yu Hua, Yongle Zhang*, Luochangqi Ding

Huazhong University of Science and Technology

**Purdue University*

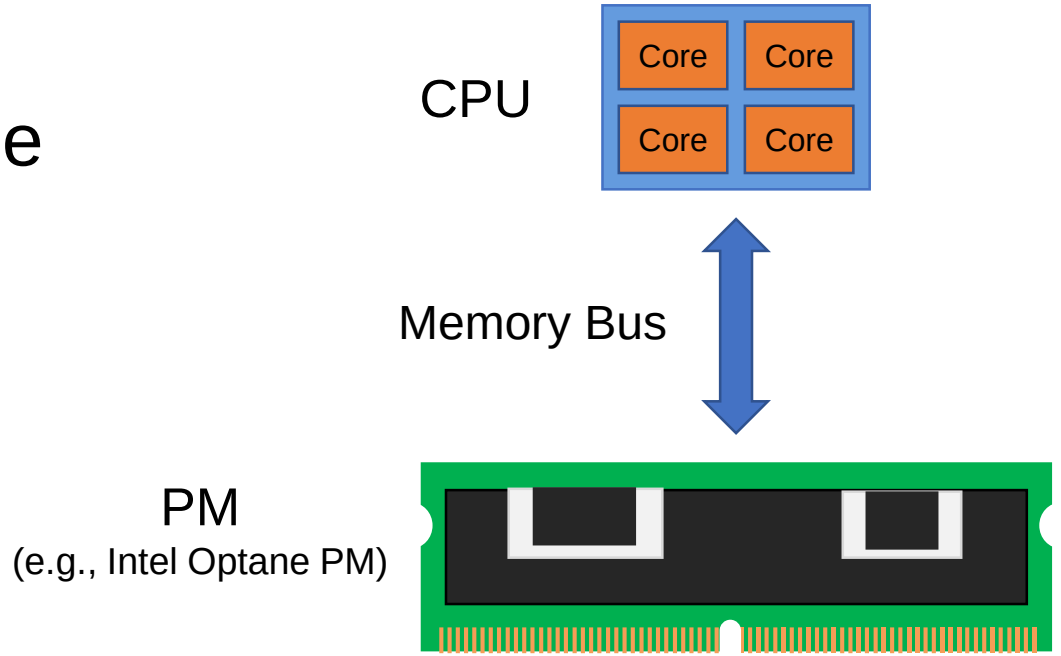
ASPLOS 2022



PURDUE UNIVERSITY

Persistent Memory (PM)

- PM characteristics
  - DRAM-comparable performance
  - TB-scale capacity
  - Non-volatility
  - Byte-addressability

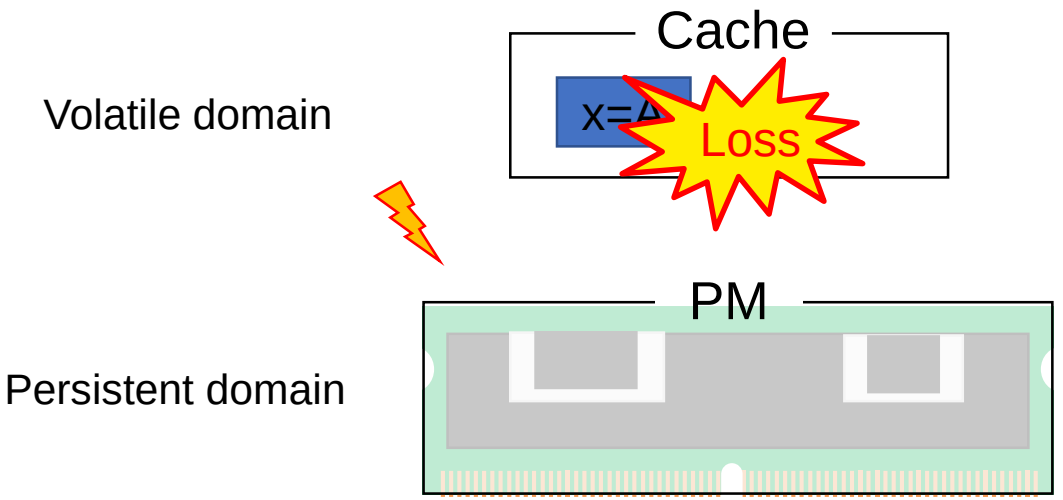


- New opportunities for memory systems
  - Lower cost/GB than DRAM
  - Instant recovery from PM

PM Programming

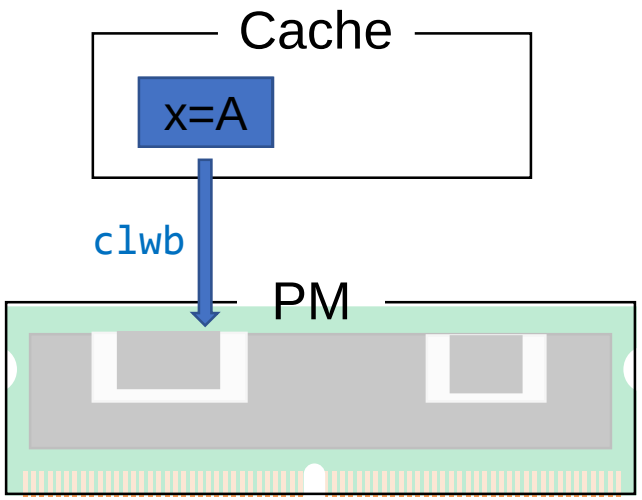
- PM programming is non-trivial
  - Volatile CPU caches

```
x = A;
```



- Architectural support for PM
  - Flush for durability (e.g., `clwb` from Intel)

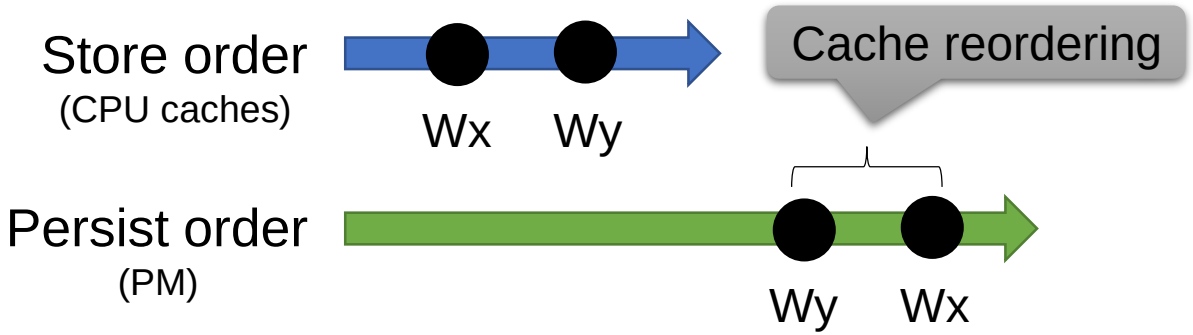
```
x = A;
clwb &x;
```



PM Programming

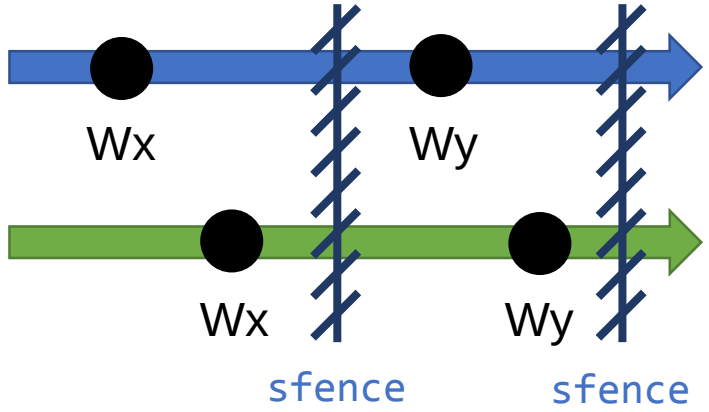
- PM programming is non-trivial
  - Volatile CPU caches
  - Persistency reordering

```
x = A;
clwb &x;
y = x;
clwb &y;
```



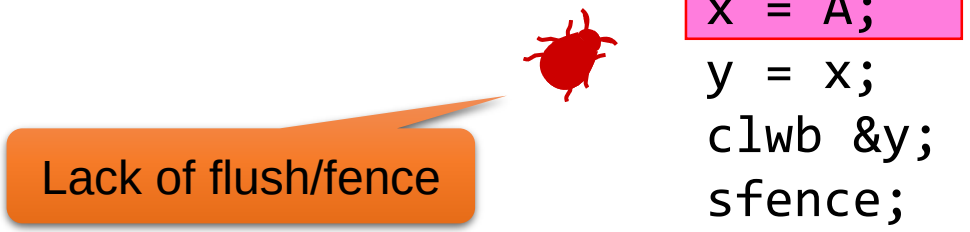
- Architectural support for PM
  - Flush for durability (e.g., `clwb` from Intel)
  - Fence for ordering (e.g., `sfence` from intel)

```
x = A;
clwb &x;
sfence;
y = x;
clwb &y;
sfence;
```

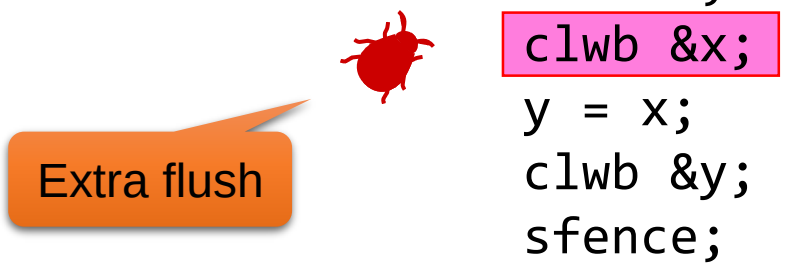


PM Crash-Consistency Bugs

- Correctness bugs
  - Causing correctness violation
  - Patterns:
    - Missing flush/fence
    - …



- Performance bugs
  - Causing performance degradation
  - Patterns:
    - Extra flush/fence
    - …



Existing Automatic PM Debugging Tools

- Correctness bugs
  - Missing flush/fence
    - AGAMOTTO [OSDI '20]
    - PMDebugger [ASPLOS '21]
  - Other patterns
    - Cross-failure race: XFDetector [ASPLOS '20]
    - Application-specific bugs: Witcher [SOSP '21]
    - ...
- Performance bugs
  - Extra flush/fence
    - AGAMOTTO, PMDebugger, …

```
x = A;
y = x;
clwb &y;
sfence;
```

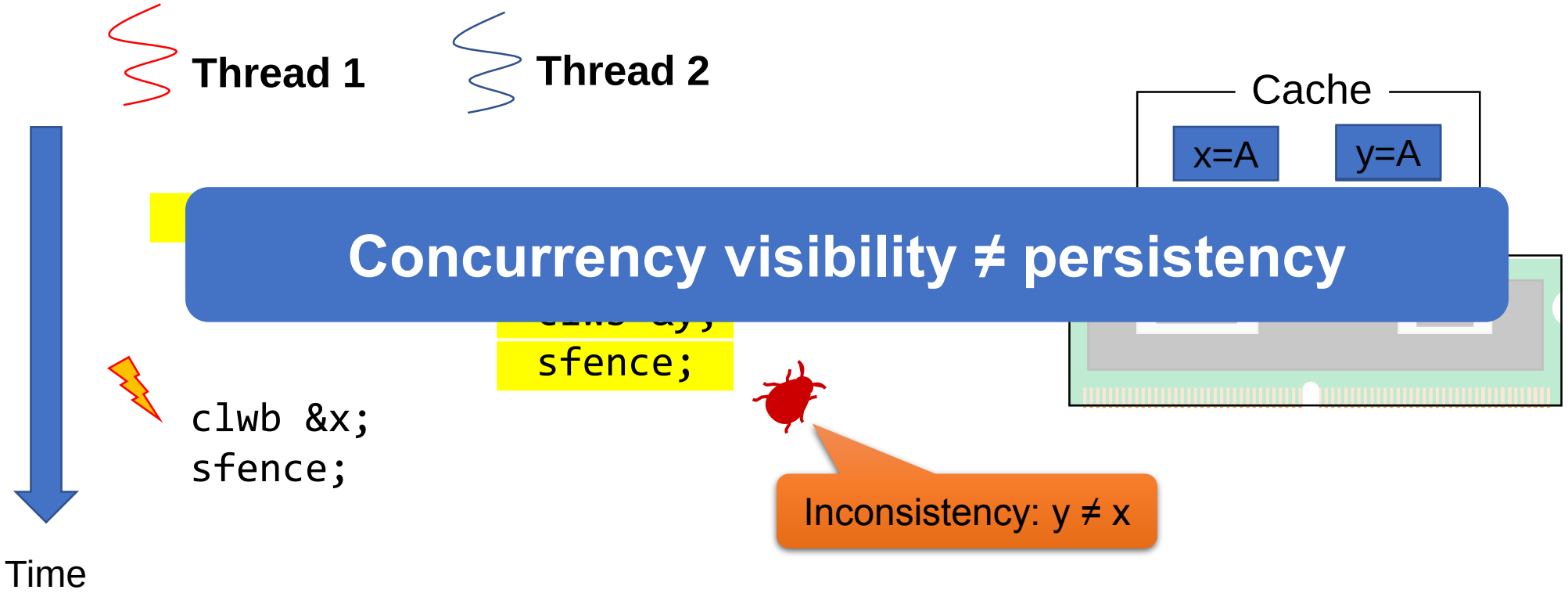
Are PM writes followed by flush/fence?

```
x = A;
clwb &x;
sfence;
clwb &x;
y = x;
clwb &y;
sfence;
```

Are flushes/fences necessary?

Crash-Inconsistency in Concurrent Executions

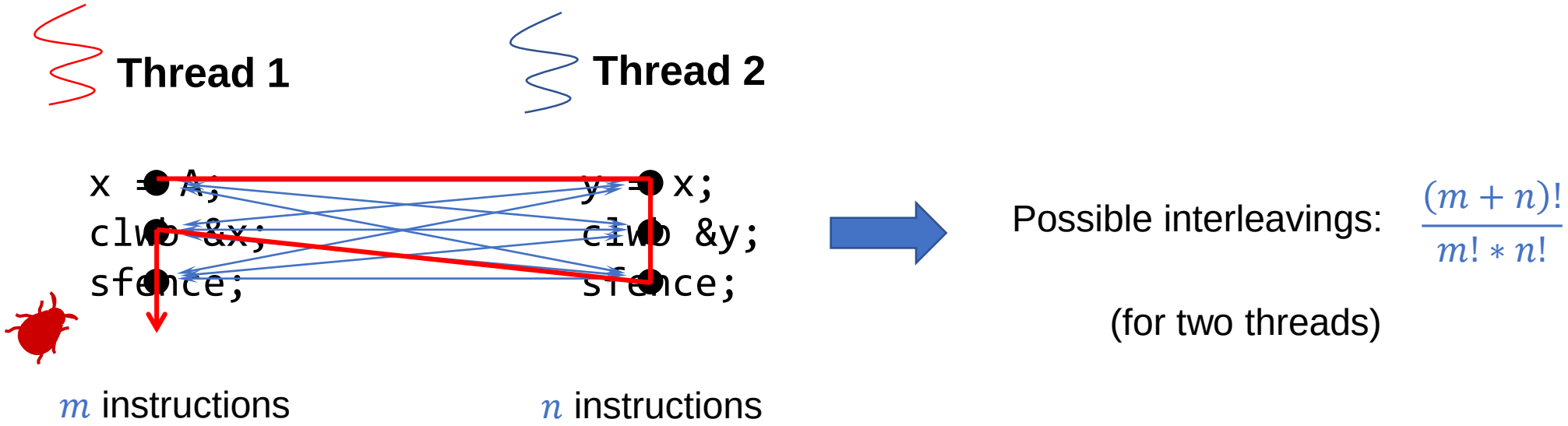
- Concurrency is important to PM system performance
- PM-specific concurrency bugs exist



No PM debugging tools detect buggy thread interleavings!

Challenges for PM Concurrency Bug Detection

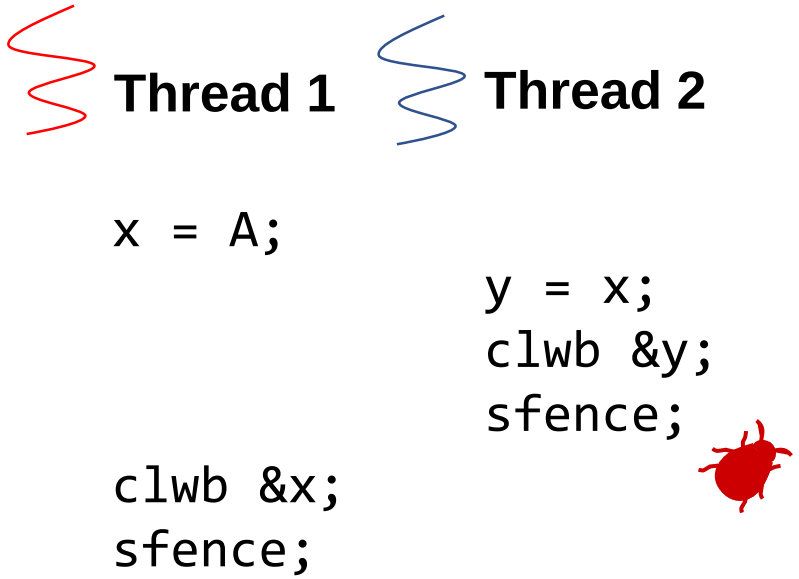
- Exponential interleaving search space
  - Exponential growth rate with respect to instructions



Challenges for PM Concurrency Bug Detection

- Exponential interleaving search space

- False positive
  - Definition: a detected bug is not a true bug
  - Reasons: inaccurate checkers, application-specific recovery…




```
Recover() {
  // Assume y = 0 initially
  if (y != 0 && y != x) {
    // Handle inconsistent x and y
  }
}
```

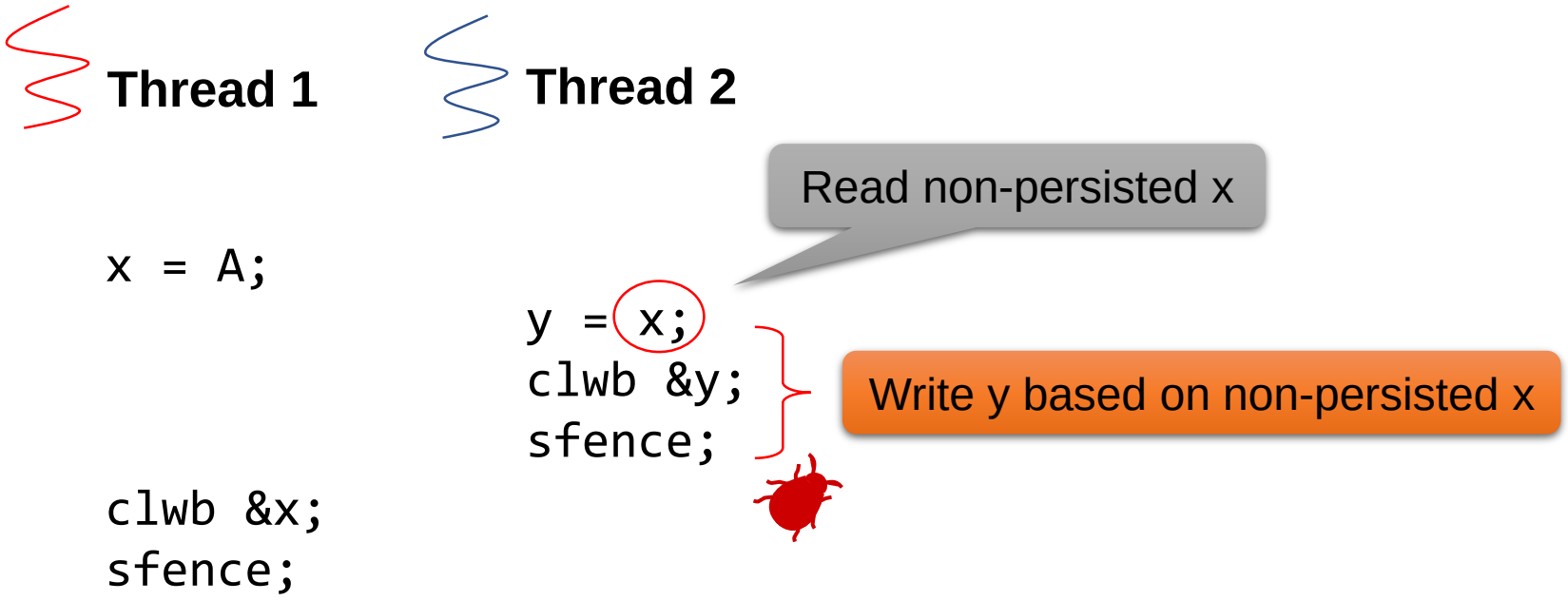
An example of custom recovery code

Our Approach: PMRace

- Two new PM concurrency bug patterns
  - *PM Inter-thread Inconsistency* and *PM Synchronization Inconsistency*
- A fuzzer for PM concurrency bugs
  - Exponential interleaving: PM-aware coverage-guided fuzzing
  - False positive: Post-failure validation
- Found 14 bugs in 5 concurrent PM programs

The Two Bug Patterns

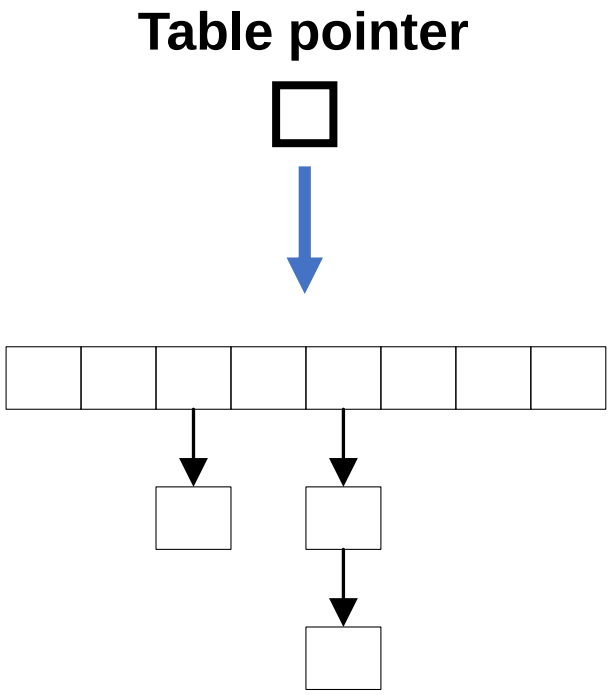
- PM Inter-thread Inconsistency
  - Durable side effects (e.g., PM writes) based on non-persisted data written by other threads



PM Interleaving Concurrency Bug

A PM Inter-thread Inconsistency in P-CLHT

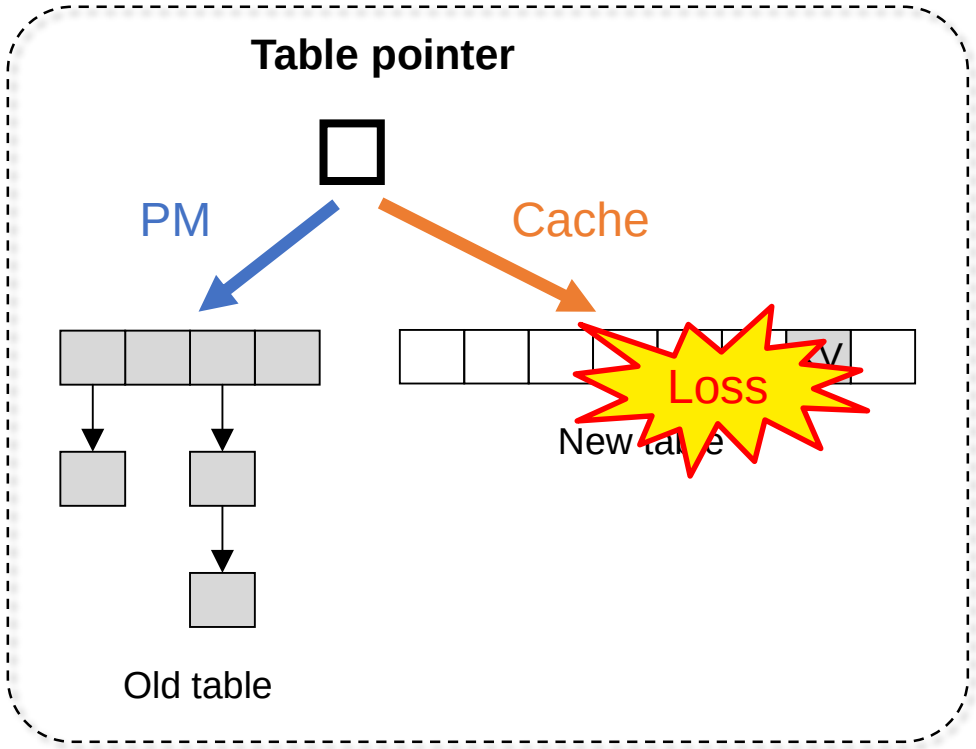
- P-CLHT from RECIPE [SOSP ’19]
  - A chained hash table for PM
  - Lock-free read and bucket locks for write



A PM Inter-thread Inconsistency in P-CLHT

Thread 1: ht_resize_pes

```
// Swap the hash table pointer for resizing
SWAP_U64(h->ht_off, pmemobj_oid(ht_new).off);
```




```
clwb(&h->ht_off); sfence();
```

Thread 2: ht_put

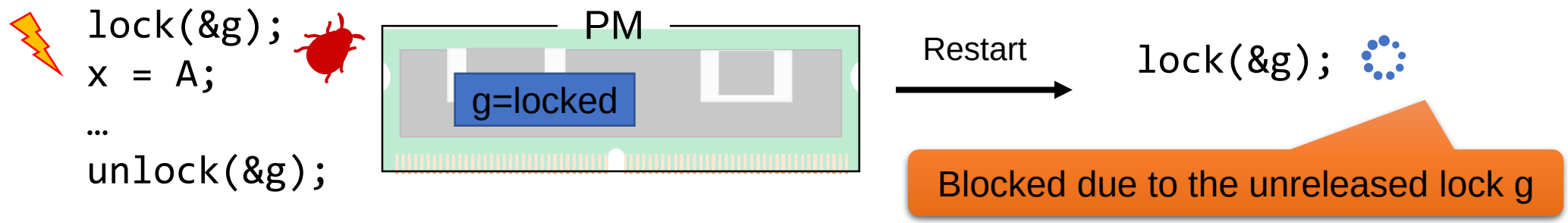
```
// Insert a key-value item
hashtable = clht_ptr_from_off(h->ht_off);
bin = clht_hash(hashtable, key)
bucket = clht_ptr_from_off(hashtable->table_off) + bin;

// Acquire the bucket lock
lock = &bucket->lock;
while (!LOCK_ACQ(lock, hashtable))
{
  ...
}

// Find an empty slot in the bucket
bucket->val[j] = val;
...
clwb(&bucket->val[j]); sfence();
movnt64(&bucket->key[j], key); sfence();
```

** The code is simplified for presentation*

The Two Bug Patterns

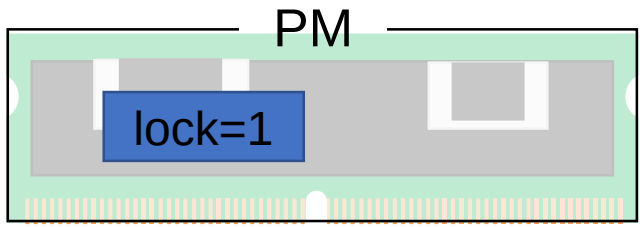
- PM Inter-thread Inconsistency
  - Durable side effects (e.g., PM writes) based on non-persisted data written by other threads

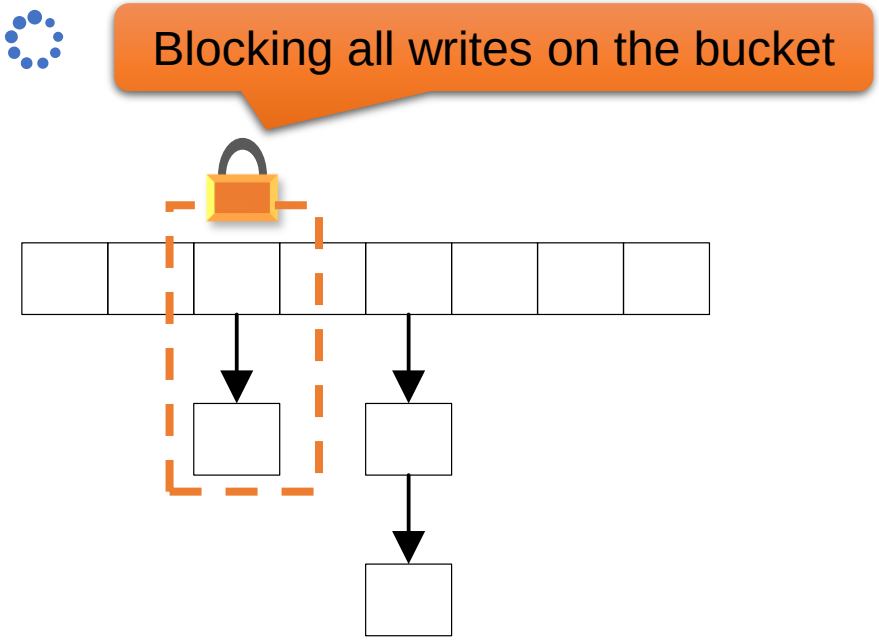
- PM Synchronization Inconsistency
  - Unreleased synchronization data after restarts



PM Execution Context Bug

A PM Synchronization Inconsistency in P-CLHT



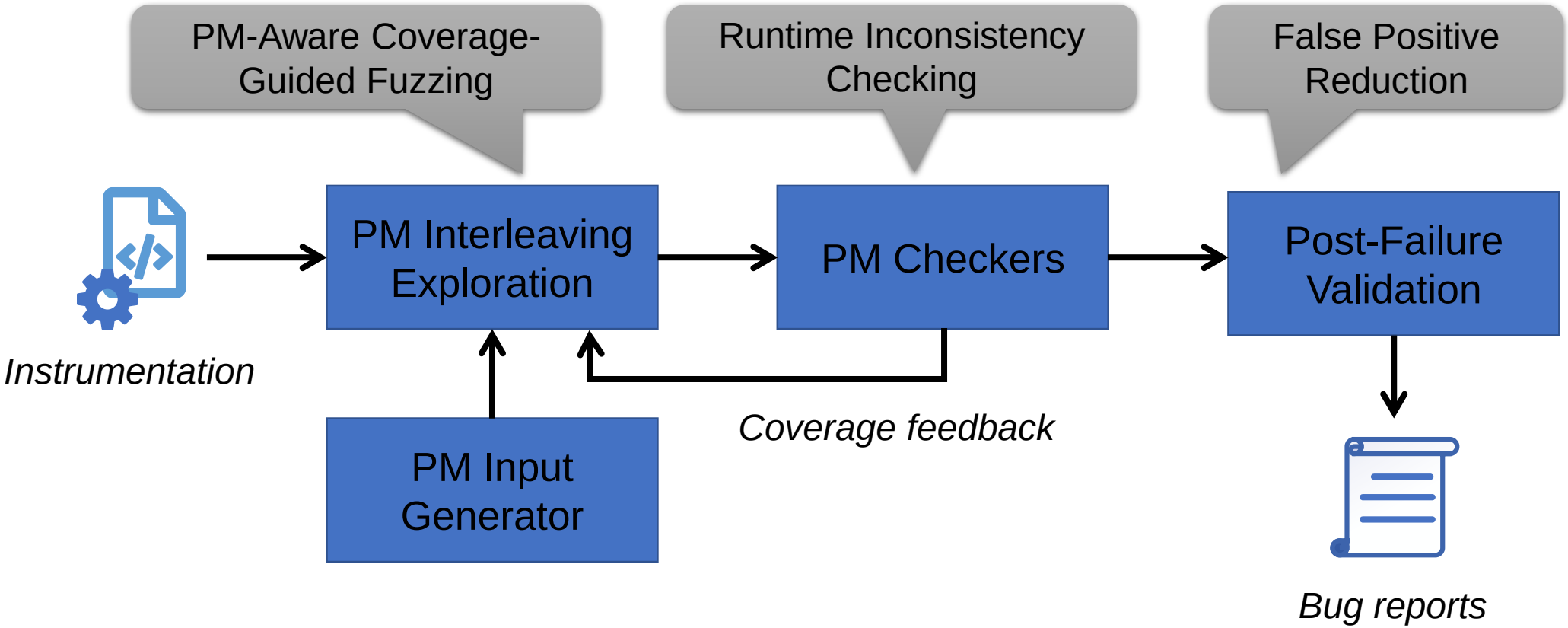


Thread x: ht_put

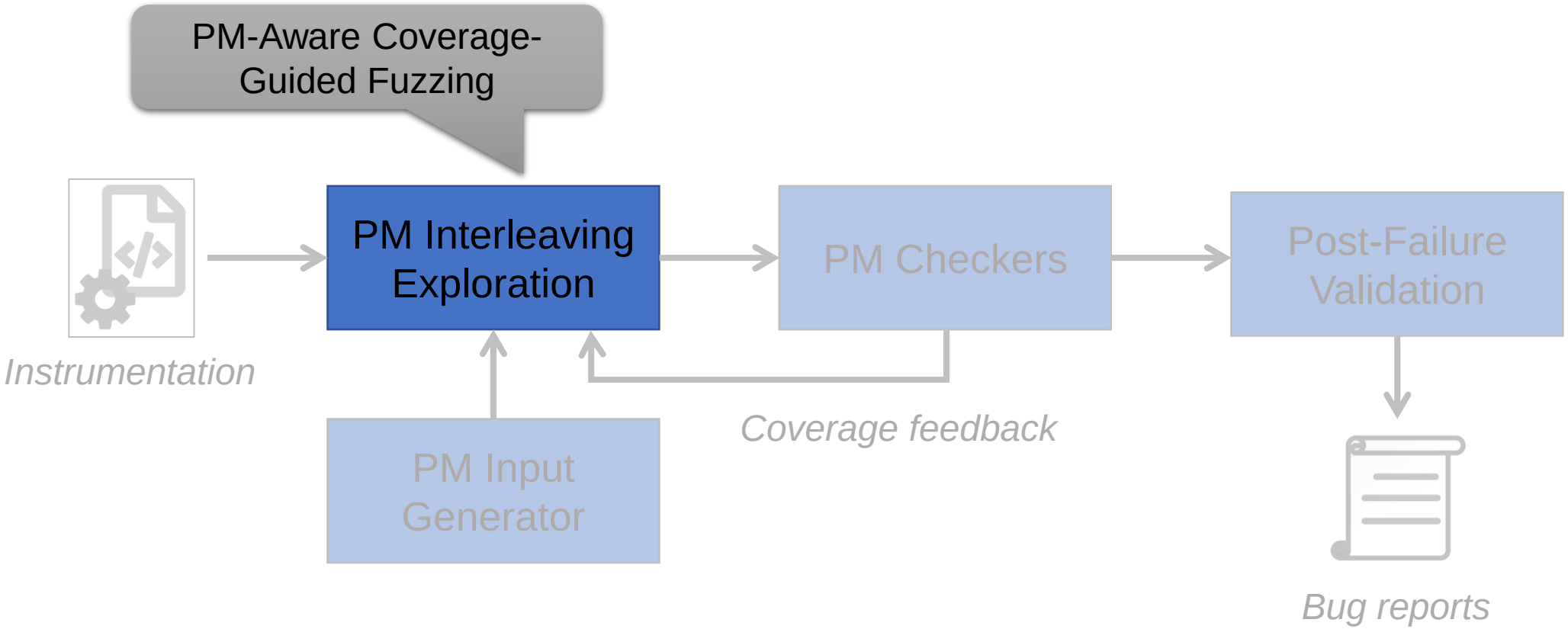
```
// Insert a key-value item
hashtable = clht_ptr_from_off(h->ht_off);
bin = clht_hash(hashtable, key)
bucket = clht_ptr_from_off(hashtable->table_off) + bin;

// Acquire the bucket lock
lock = &bucket->lock;
while (!LOCK_ACQ(lock, hashtable))
{
  ...
}

// Find an empty slot in the bucket
bucket->val[j] = val;
...
clwb(&bucket->val[j]); sfence();
movnt64(&bucket->key[j], key); sfence();

LOCK_RLS(lock);
```

** The code is simplified for presentation*

PMRace Overview

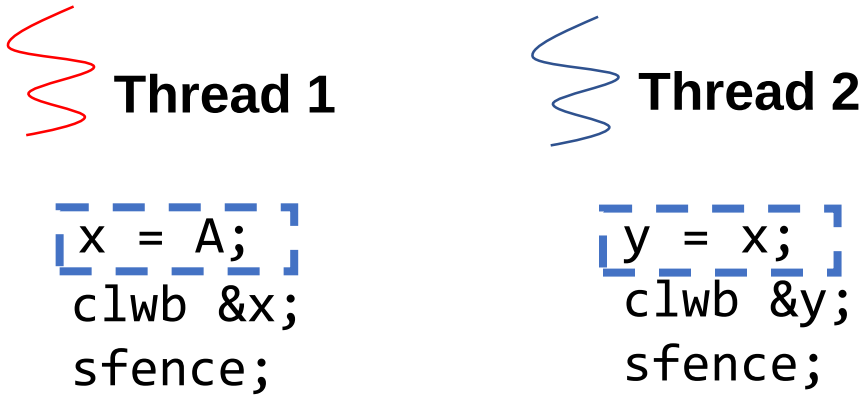


PMRace Overview



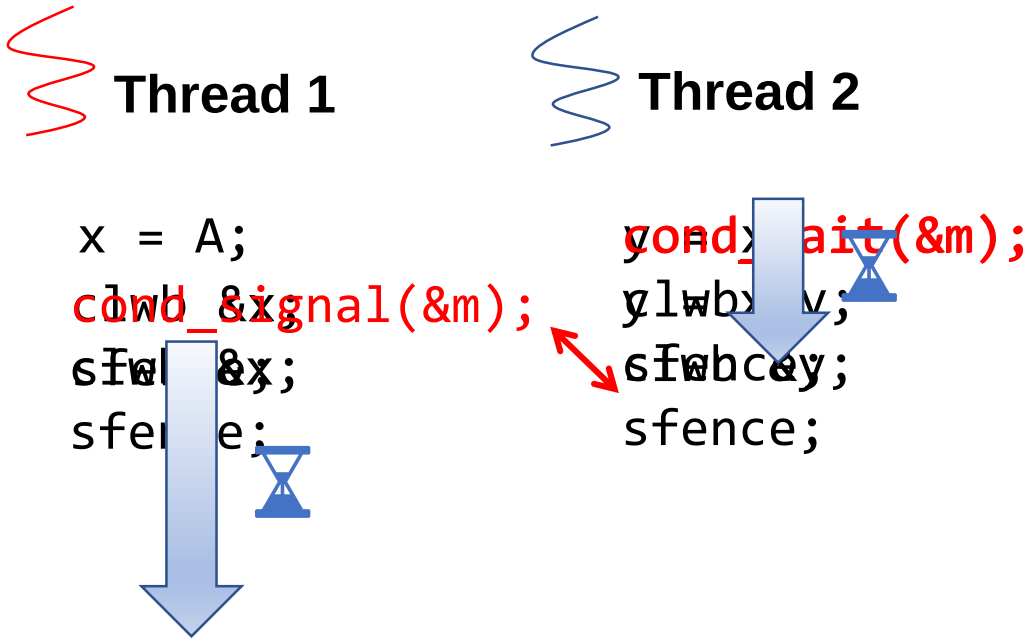
PM-Aware Coverage-Guided Fuzzing

- PM-aware interleaving exploration
  - Driving the execution towards reading non-persisted data
  - Step 1: preemption point selection
    - PM accesses to shared data
    - A priority queue of PM access



PM-Aware Coverage-Guided Fuzzing

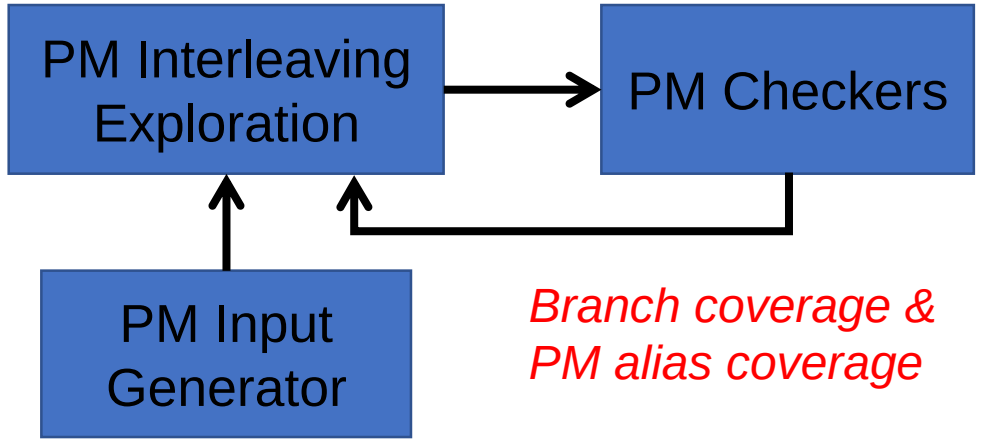
- PM-aware interleaving exploration
  - Driving the execution towards reading non-persisted data
  - Step 1: preemption point selection
  - Step 2: scheduling a group of alias PM accesses (to the same address)
    - `cond_wait` before PM reads
    - `cond_signal` after corresponding PM writes and before flushes
    - Pitfalls and solutions… (refer to the paper)



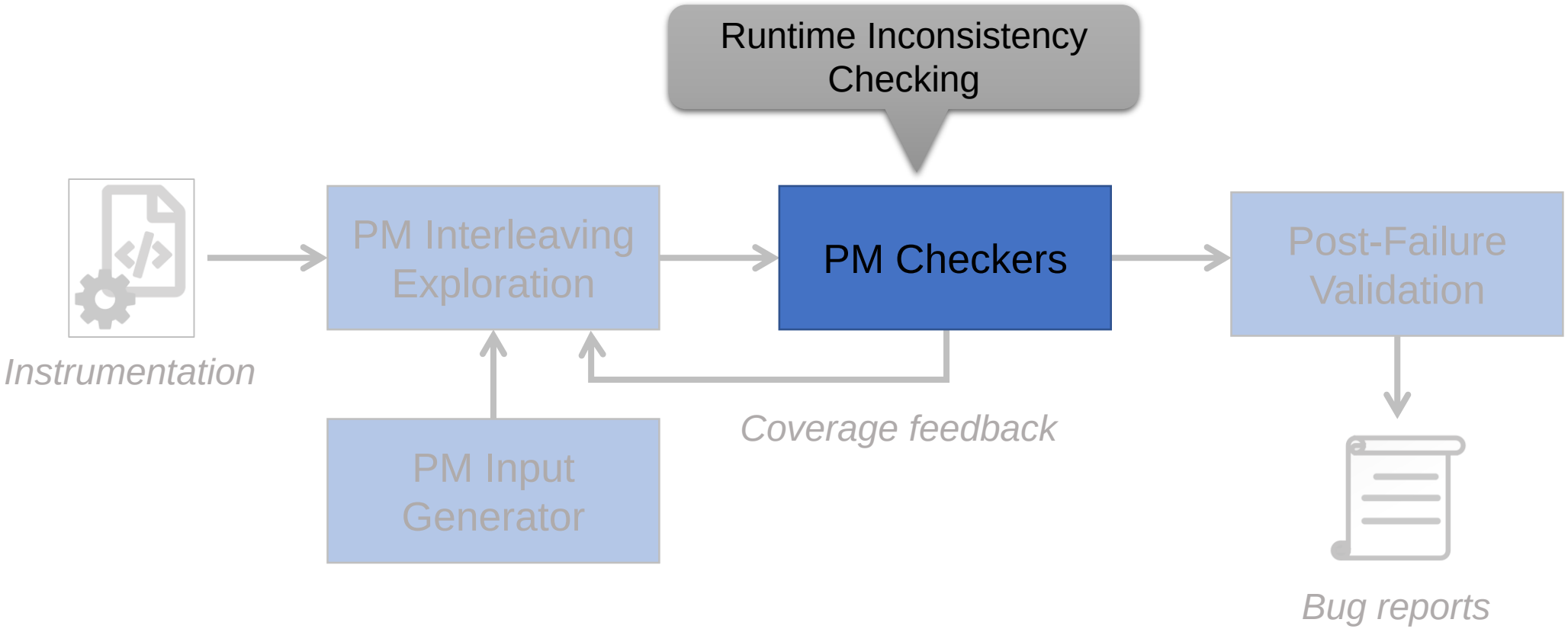
PM-Aware Coverage-Guided Fuzzing

- PM-aware interleaving exploration

- PM alias (pair) coverage
  - Recording visited concurrent PM access to the same address
  - Guiding fuzzing to test “new” interleavings

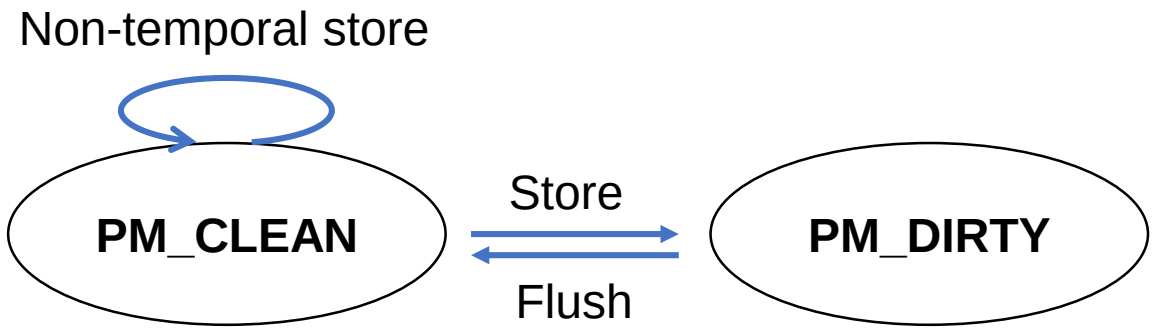


PMRace Overview



PM Inconsistency Checkers

- Persistency state tracking

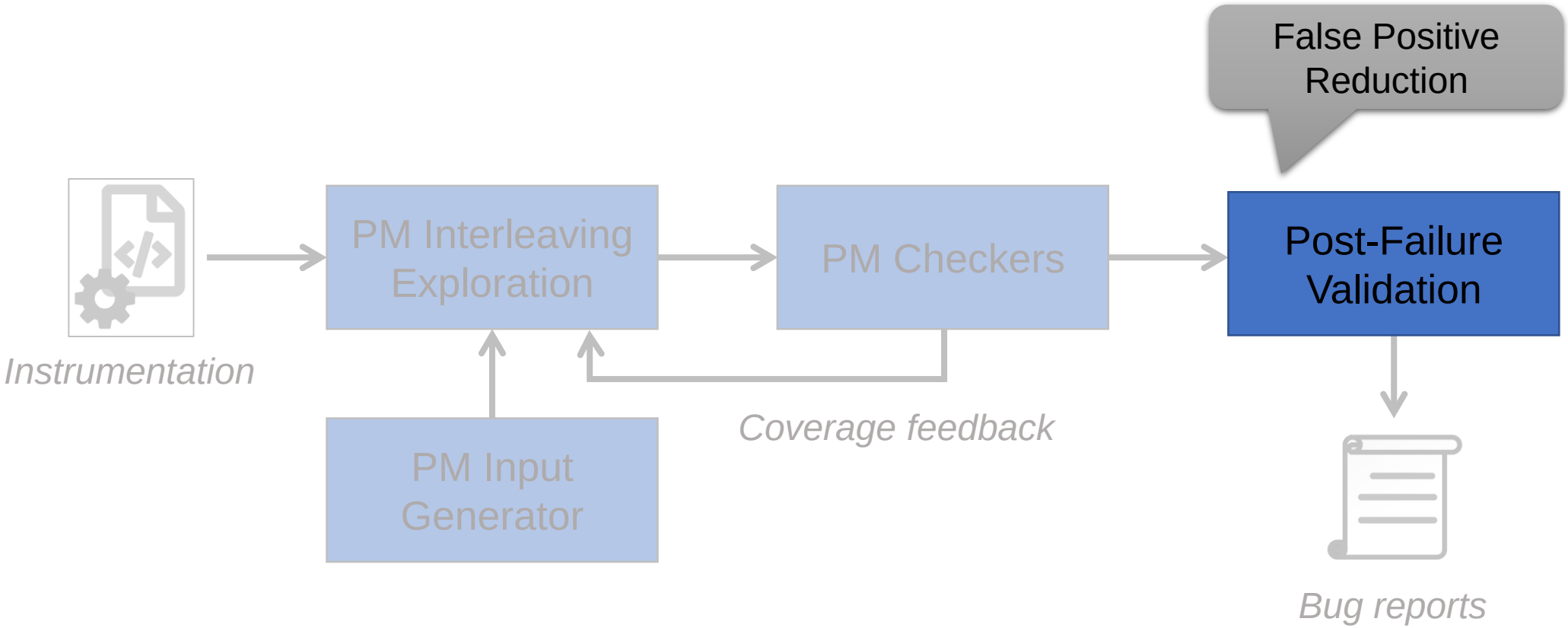


- Runtime PM checkers
  - *PM Inter-thread Inconsistency* when
    1. Reading non-persisted data (PM_DIRTY) and
    2. causing durable side effects (PM writes)
  - *PM Synchronization Inconsistency* when
    1. Updating annotated synchronization data

Data flow

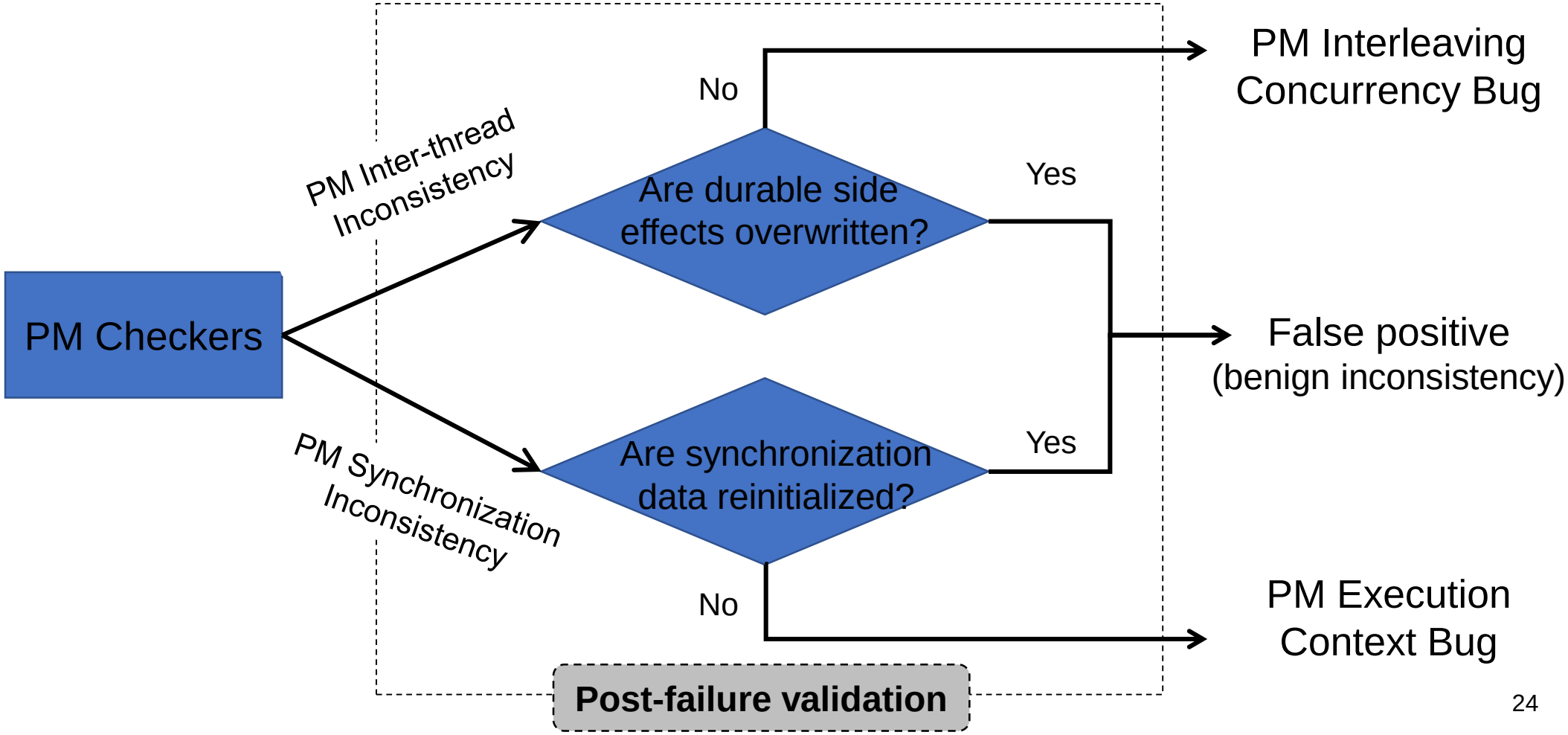
```
1  hook_load(&x);
   y = x;
2  hook_store(&y); 1
   clwb &y;
   sfence;
```

PMRace Overview



Post-Failure Validation

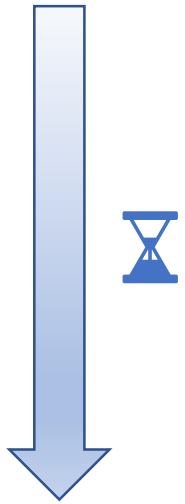
- To identify false positive (benign inconsistency)



An Example of Benign Inconsistency

Thread 1

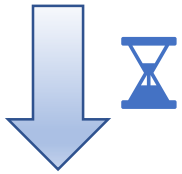
```
x = A;
hook_store(&x);
cond_signal(&m);
```



```
clwb &x;
sfence;
```

Thread 2

```
cond_wait(&m);
```



```
hook_load(&x);
y = x;
hook_store(&y);
clwb &y;
sfence;
```

```
Recover() {
  // Assume y = 0 initially
  if (y != 0 && y != x) {
    // Handle inconsistent x and y
    x = 0;
    y = 0;
  }
}
```

Post-failure validation: benign inconsistency

PM Checker: PM Inter-thread Inconsistency

Evaluation

- System configurations
  - Two 26-core Intel Xeon Gold 6230R CPUs
  - 1.5 TB Intel Optane PM 100 Series, 192 GB DRAM
- Tested 5 open-source concurrent PM programs based on PMDK

Systems	Scope	Concurrency
P-CLHT [SOSP '19]	Static hashing	Lock-based
Clevel Hashing [ATC '20]	PM-optimized hashing	Lock-free
CCEH [FAST '19]	Extendible hashing	Lock-based
FAST-FAIR [FAST '18]	B+-Tree	Lock-based
memcached-pmem	Key-value store	Lock-based

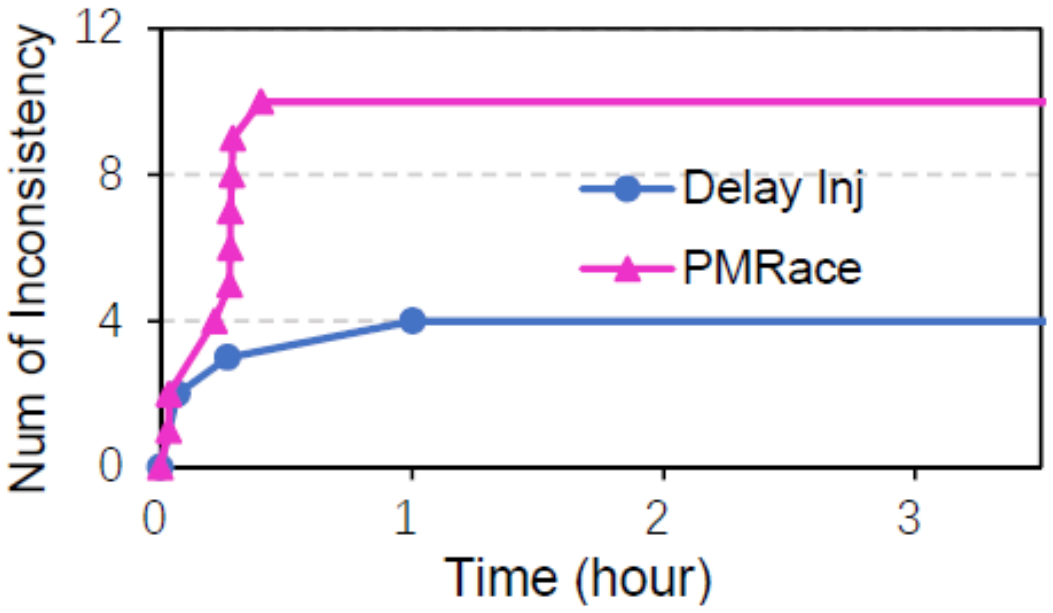
- Comparison
  - **PMRace**: our scheme
  - **Delay Inj**: PMRace with random delay injection for interleaving exploration

14 Bugs

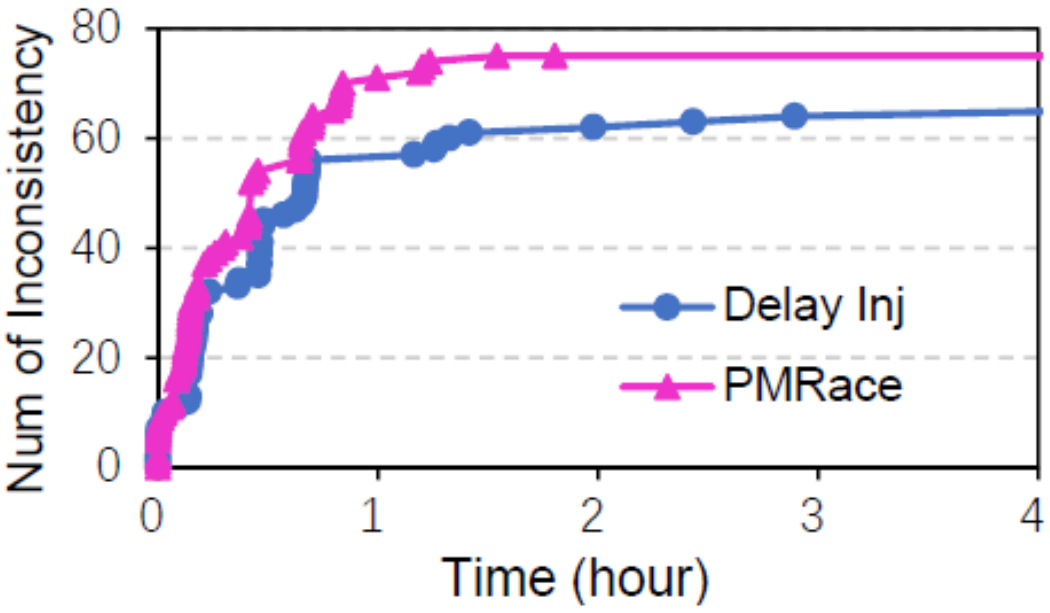
	#	Type	New		
P-CLHT	1	Inter	Y	read unflushed table pointer and insert items	data loss
	2	Sync	Y	do not initialize bucket locks after restarts	hang
	3	Intra	Y	read unflushed table pointer and perform GC	PM leakage
	4	Other	Y	read unflushed keys	redundant PM writes
	5	Other	Y	do not release bucket locks in update	hang
CCEH	6	Sync	Y	do not release segment locks after restarts	hang
	7	Intra	Y	read unflushed capacity and allocate segments	PM leakage
FAST-FAIR	8	Inter	Y	read unflushed pointer and insert data	data loss
memcached-pmem	9	Inter	Y	read unflushed value and write value	inconsistent data
	10	Inter	Y	read unflushed value and write value	inconsistent data
	11	Inter	N	read unflushed "prev" and write "slabs_clsid"	inconsistent data
	12	Inter	N	read unflushed "prev" and write "it_flags" or value	inconsistent index
	13	Inter	N	read unflushed "it_flags" and write value	inconsistent data
	14	Inter	N	read unflushed "slabs_clsid" and write "slabs_clsid" of others	inconsistent index

Inter: PM Inter-thread Inconsistency Intra: PM Intra-thread Inconsistency Sync: PM Synchronization Inconsistency

The Time to Identify PM Inter-thread Inconsistency



(1) P-CLHT



(2) FAST-FAIR

- PMRace efficiently triggers reading non-persisted data

Inconsistencies and False Positives

Systems	Inconsistencies (pre-failure)			False Positives (post-failure)		Bug
	Inter-Cand	Inter	Sync	Inter	Sync	
P-CLHT	35	10	4	0	3	2
clevel hashing	6	2	0	0	0	0
CCEH	15	0	1	0	0	1
FAST-FAIR	179	69	0	3	0	1
memcached-pmem	266	79	0	62	0	6
Total	501	160	5	65	3	10

- Durable side effects refine inconsistencies
- Post-failure validation reduces false positives
- Limitation: false positives still exist due to lazy recover mechanisms…

Conclusion

- PM-specific concurrency bugs are hard to detect and unexplored
- We identify two new PM concurrency bug patterns
- **PMRace**: the first tool to detect PM concurrency bugs
  - PM-aware coverage-guided fuzzing to speedup interleaving searching
  - Post-failure validation to reduce false positives
- Found 14 bugs in 5 concurrent PM programs
- Open-source at https://github.com/yhuacode/pmrace

Thanks!

Email: chenzy@hust.edu.cn

Homepage: https://chenzhangyu.github.io