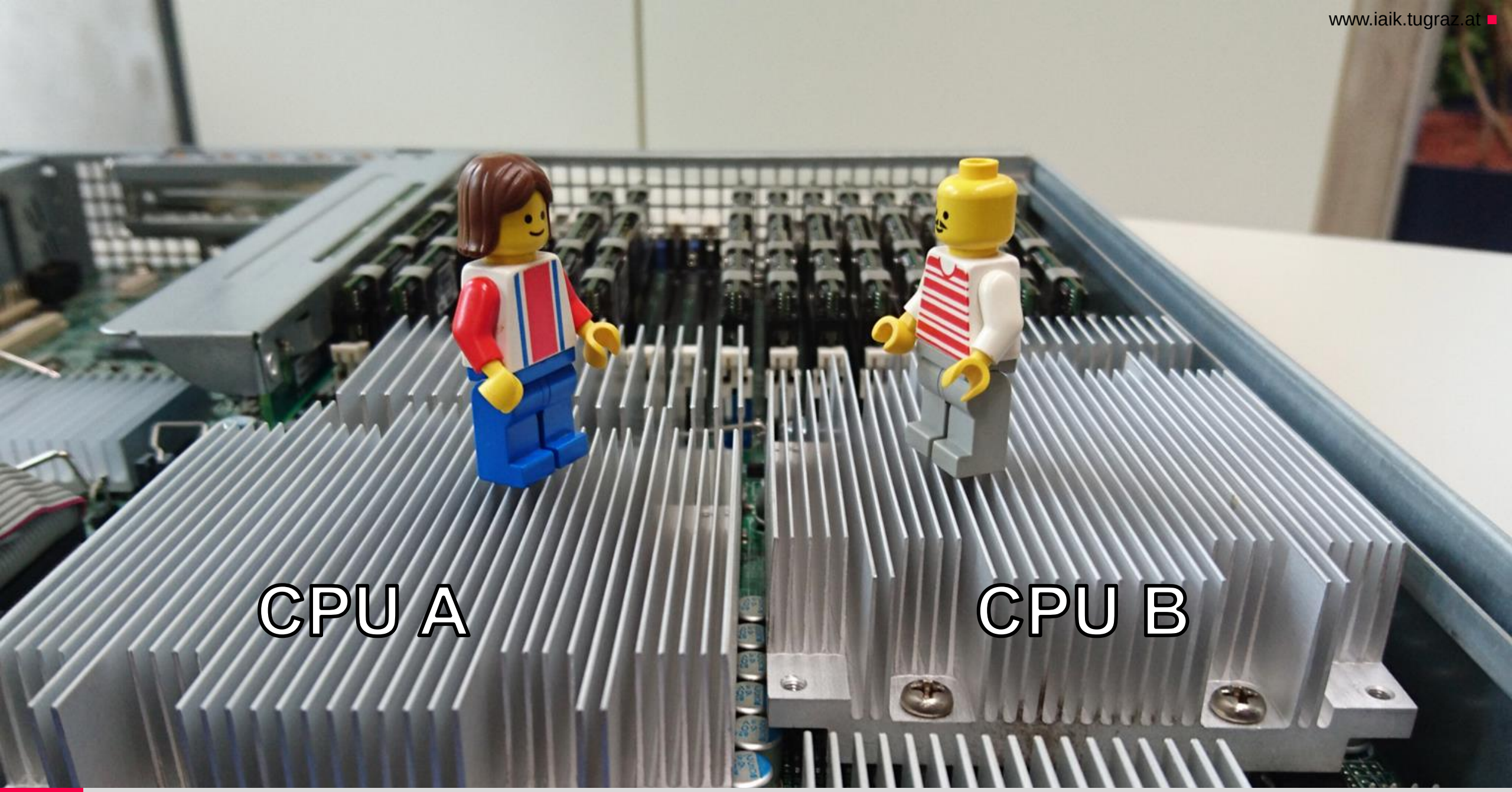
DRAMA: Exploiting DRAM Addressing for Cross-CPU Attacks

Peter Pessl, Daniel Gruss, Clémentine Maurice, Michael Schwarz, Stefan Mangard
IAIK, Graz University of Technology, Austria

Usenix Security 2016, August 11


CPU A
CPU B

Setting – Cloud Servers

- Multi-CPU (multi-socket) systems
- Multiple tenants
  - separate VMs
  - dedicated CPUs → no shared cache
- No shared memory
  - no cross-VM memory deduplication

- Previously
  - slow covert channel (< 1 kbps)
  - no side channel

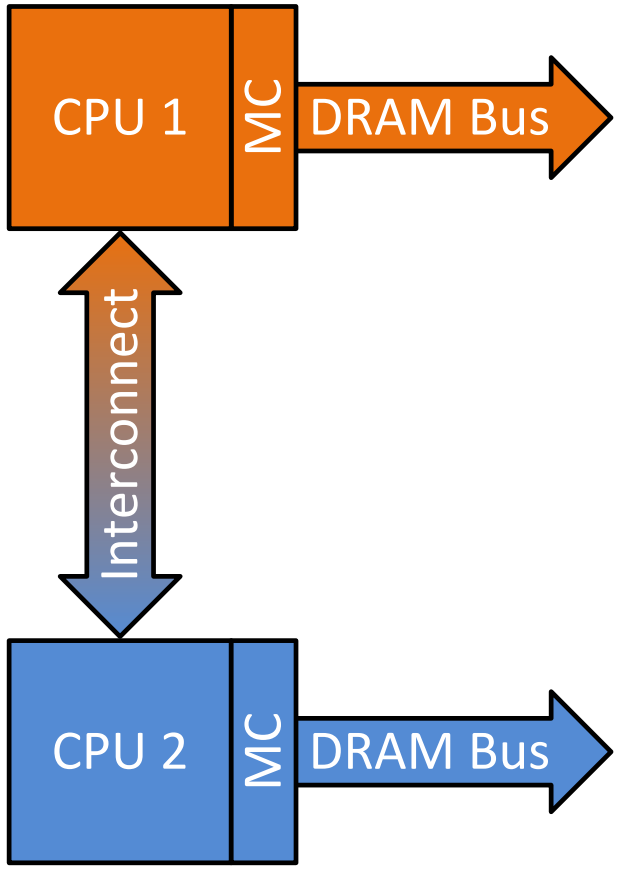
Overview

- Cross-CPU attacks using **DRAM a**ddressing (**DRAMA**)
  - fast covert channel (up to 2 Mbps)
  - first side-channel attack
- Reverse-engineered DRAM addressing
  - two approaches
- Improving existing attacks

DRAM Organization

Hierarchy of

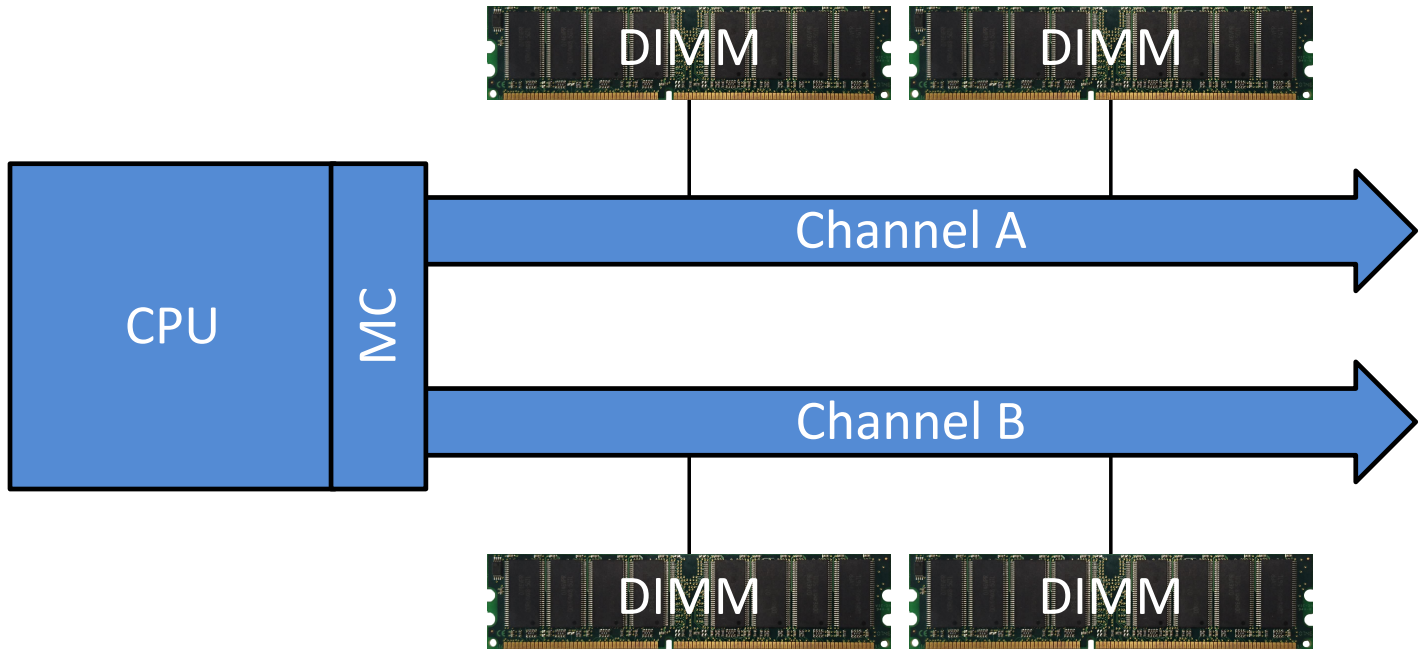
- CPUs



DRAM Organization

Hierarchy of

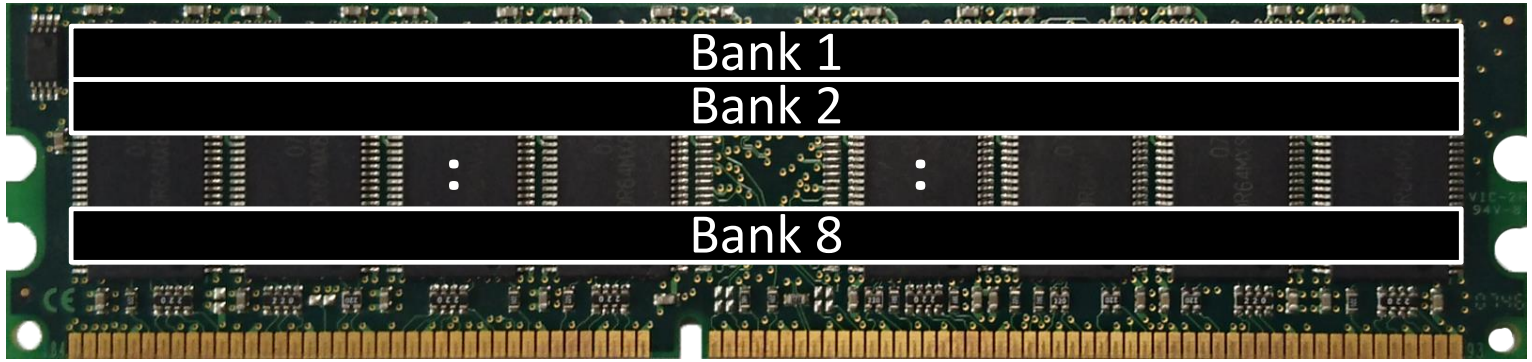
- CPUs
- Channels
- DIMMs



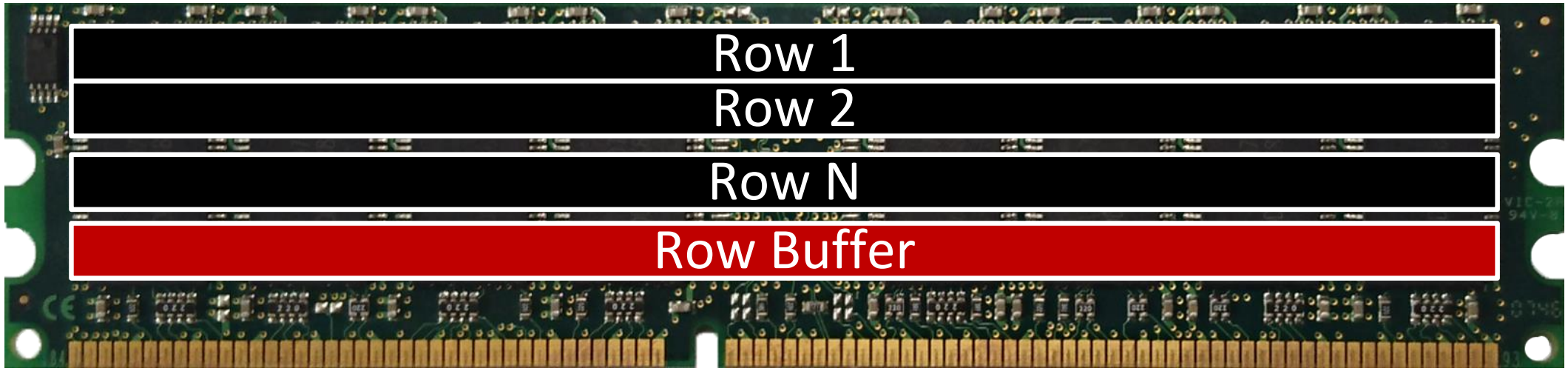
DRAM Organization

Hierarchy of

- CPUs
- Channels
- DIMMs
- Ranks
- Banks

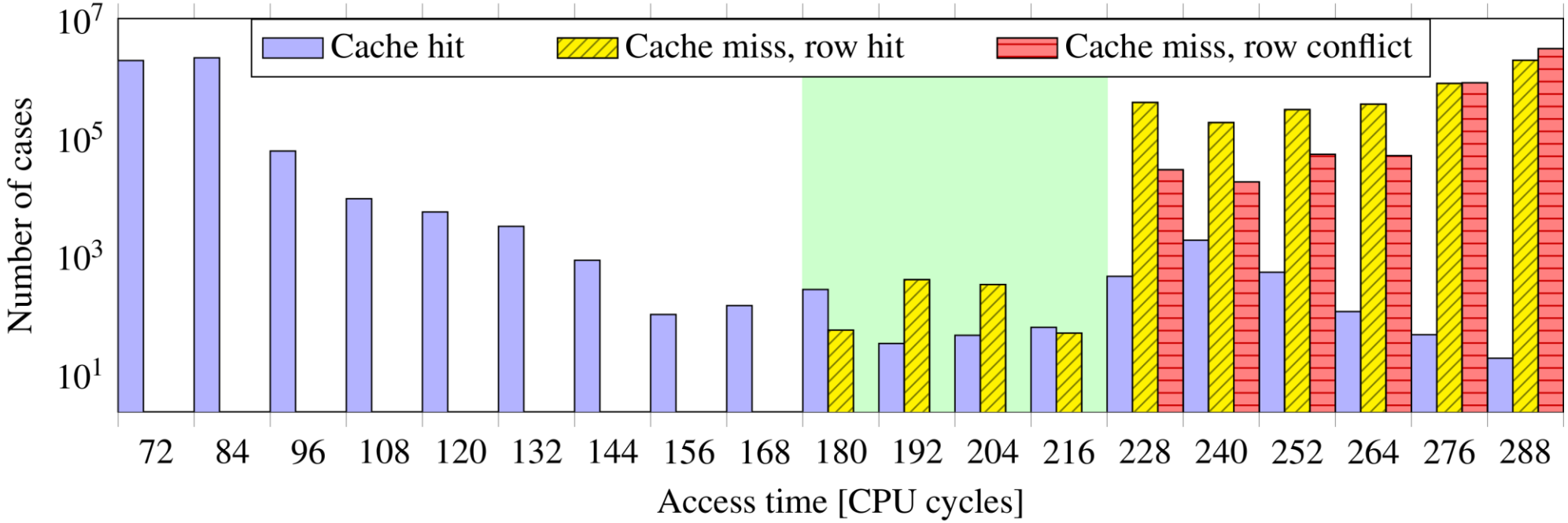


DRAM Banks



- Memory array
  - rows of columns
- Row Buffer
  - buffers one entire row (8 KB)

The Row Buffer



- Behavior similar to a cache
  - row hits → fast access
  - row conflicts → slow access

Reverse Engineering

of DRAM Addressing

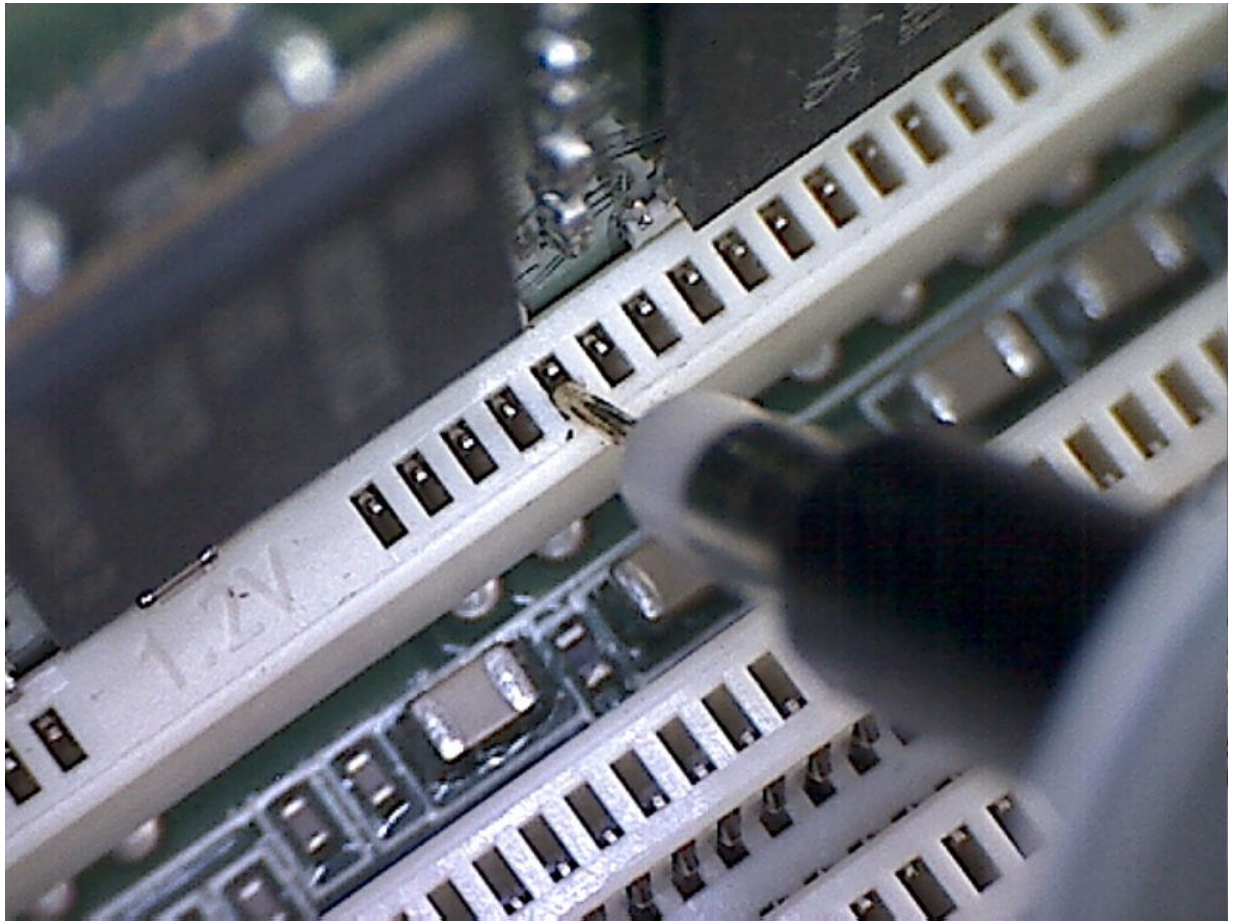
Reverse-Engineering DRAM Addressing

- Mapping to banks using physical-address bits
- „Complex“ addressing functions
  - distribute traffic to channels/banks
  - undisclosed (Intel)

- Two approaches to reverse engineer
- Presumption: linear functions (XORs)

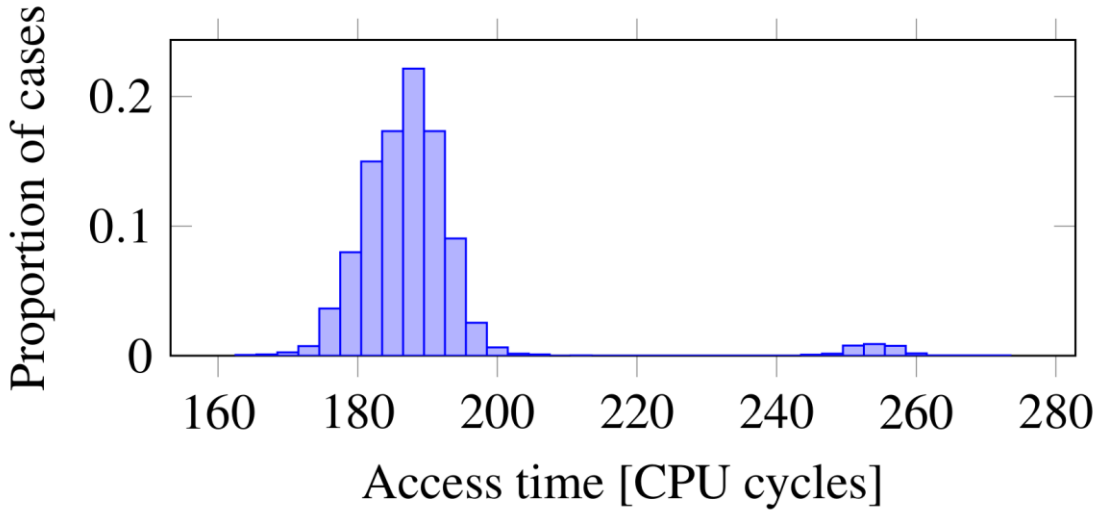
Approach 1: Probing the Memory Bus

- Probing of control signals
  - CS, BA, …
  - measure voltage with Osci.
  - recover logic value
- Repeated access to address
  - until value is determined
- Function reconstruction
  - linear algebra over bits



Approach 2: Fully Automated SW-based

- Exploit timing differences
- Measuring phase
  - build sets of same-bank addresses
  - alternating access to two addresses
  - measure avg. access time



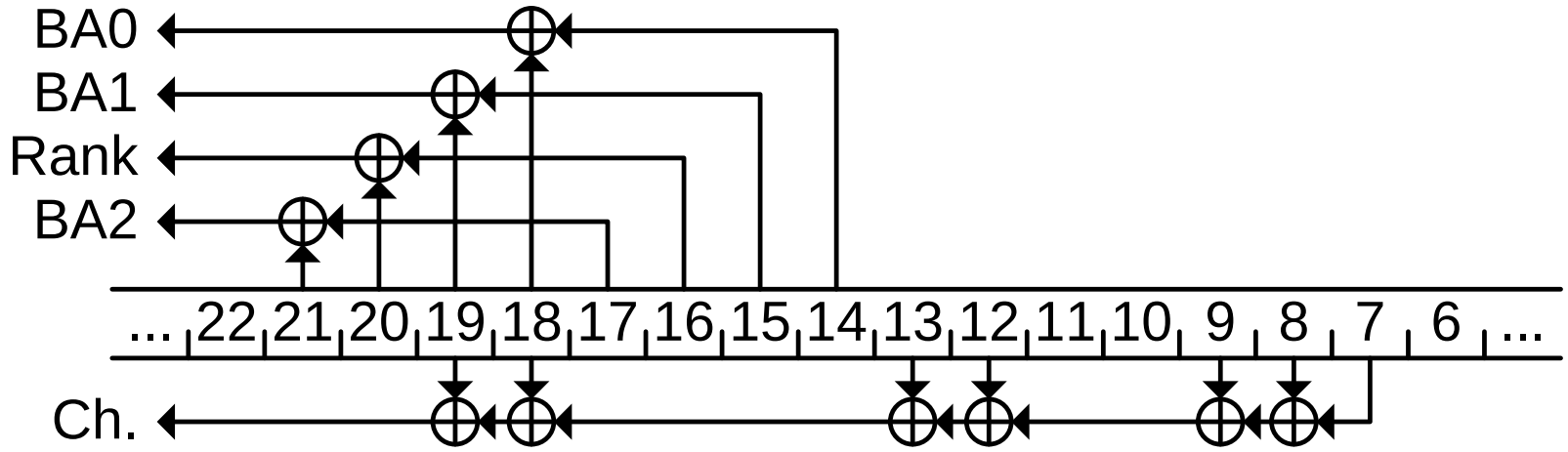
- Reconstruction phase
  - exhaustive search over linear functions with up to *n* set coefficients
- Total time: seconds

Comparison

- Probing
  - recover function labels
  - find a ground truth
  - equipment and access to internals of machine

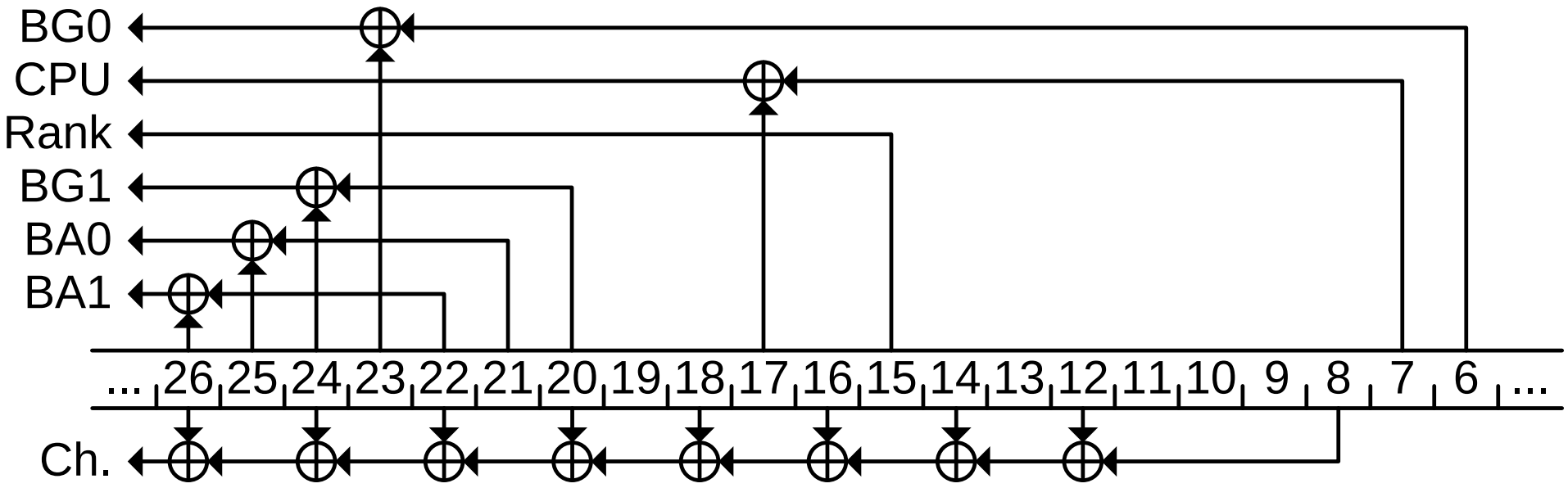
- SW-based
  - fully automated
  - ability to run remotely, sandboxed, and on mobile devices

Some Results - Desktop



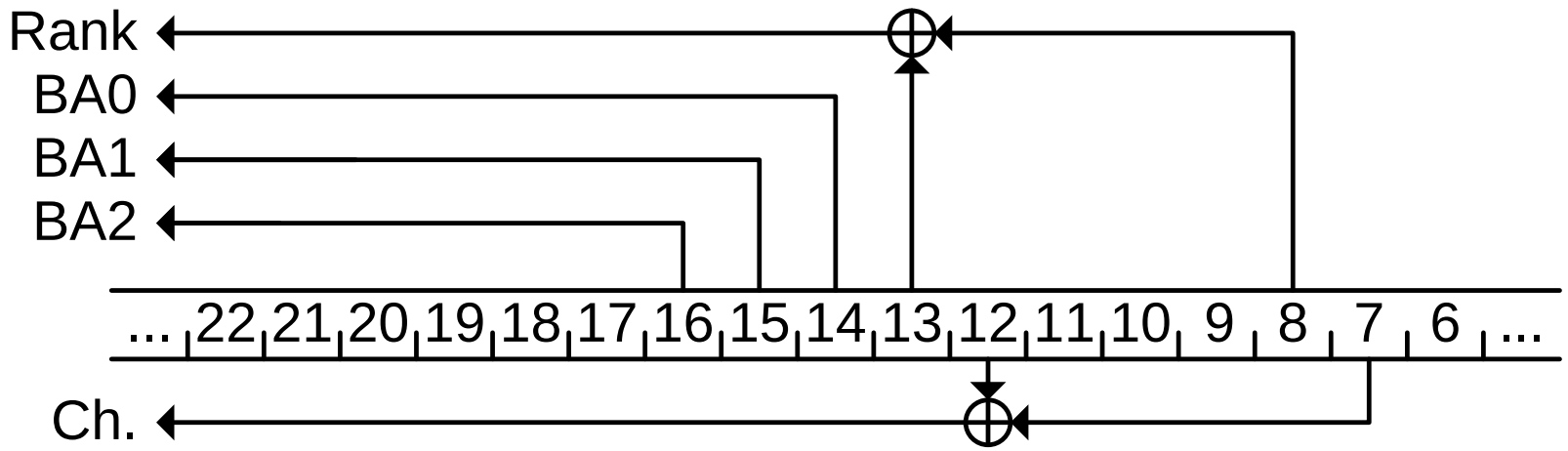
Intel Haswell (desktop system) – DDR3

Some Results – Server System



Dual-CPU Intel Haswell-EP – DDR4

Some Results – Mobile

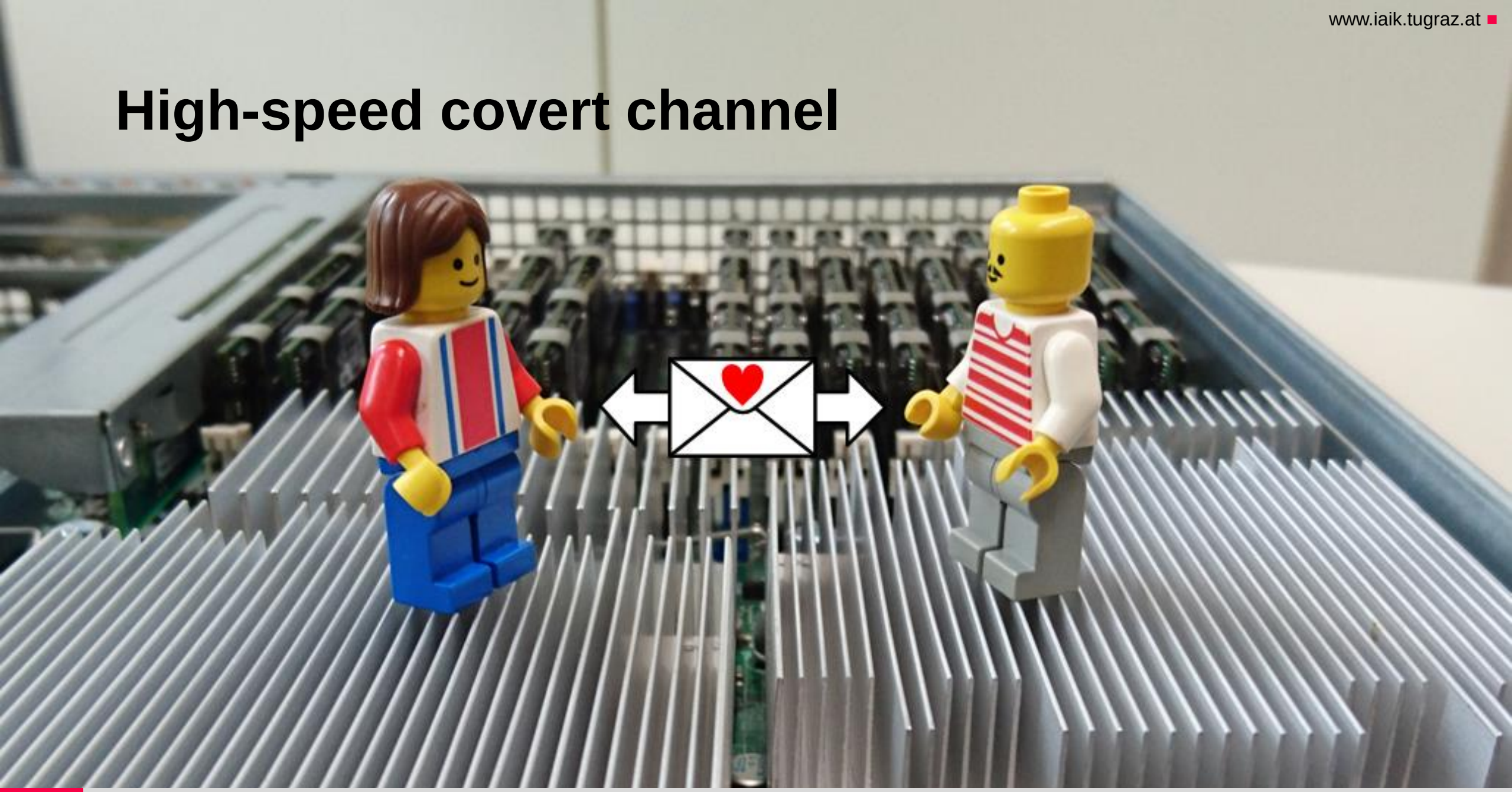


Samsung Exynos 7420 (Galaxy S6) – LPDDR4

Cross-CPU Attacks

…and how it continues with Romeo and Juliet

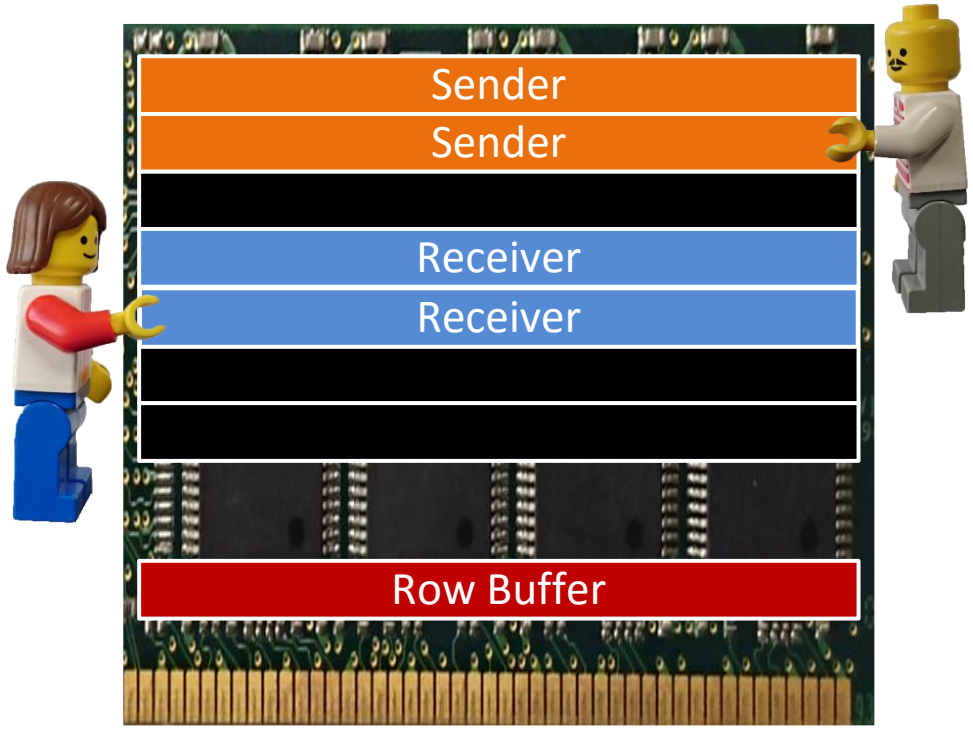
High-speed covert channel



Concept

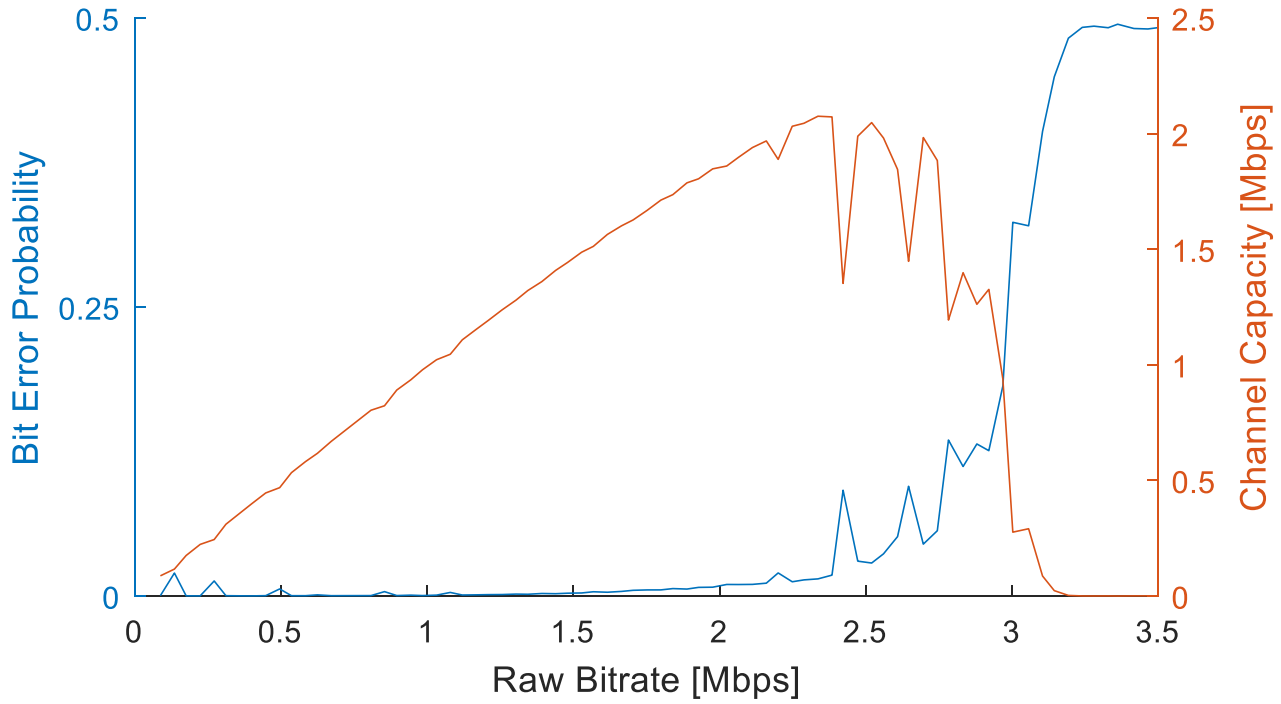
- Occupy different rows in the same bank

- Sender
  - send 1: continuously access row
  - send 0: don‘t do anything
- Receiver
  - access row and measure avg. time
  - infer sent bits based on time



Implementation

- Each bank is a channel
  - use up to 8 banks in parallel
  - multithreading
- Performance:
  - desktop: 2.1 Mbps
  - multi-CPU server: 1.2 Mbps

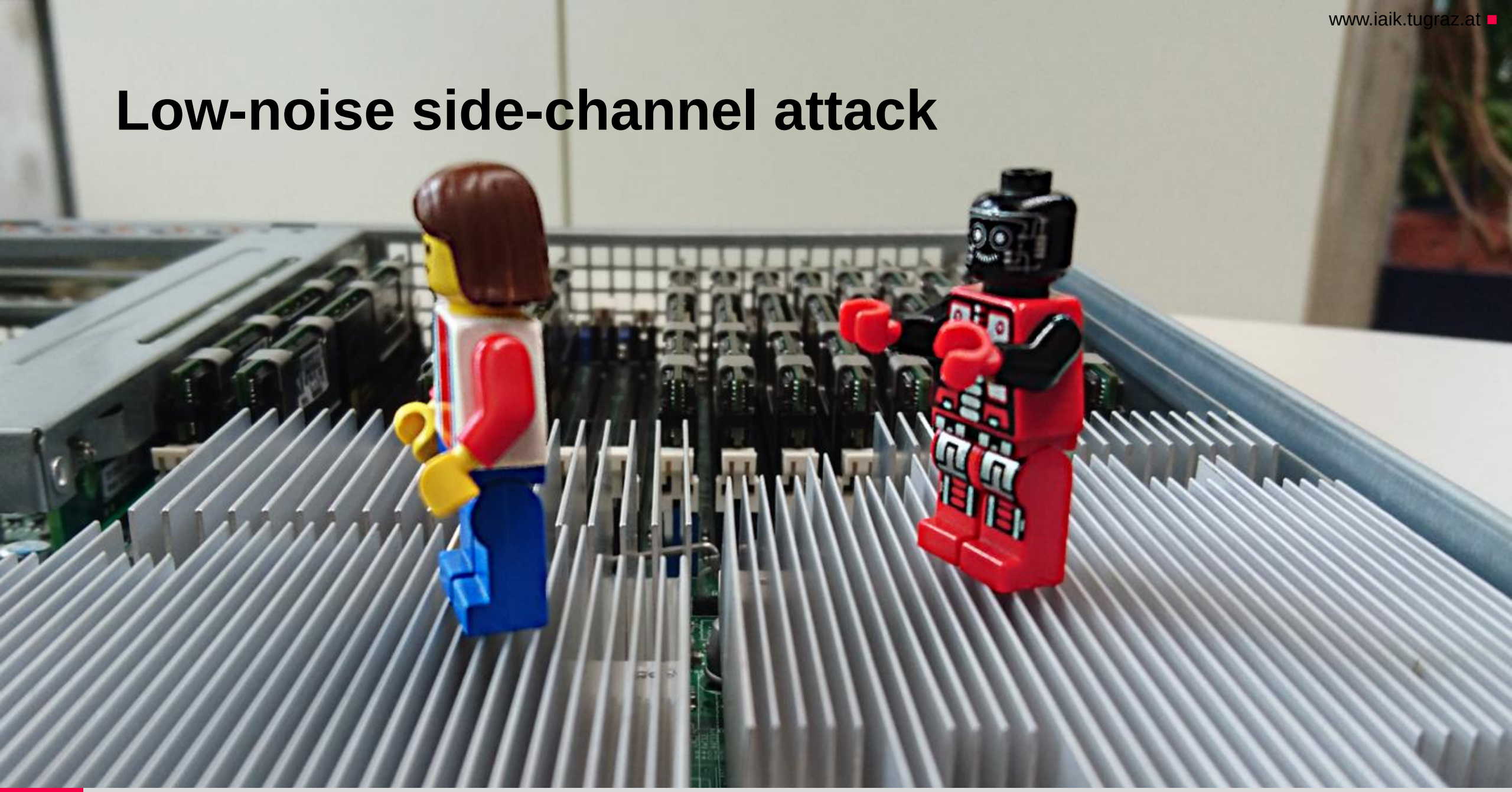


Intel Haswell (desktop system)

Performance Comparison

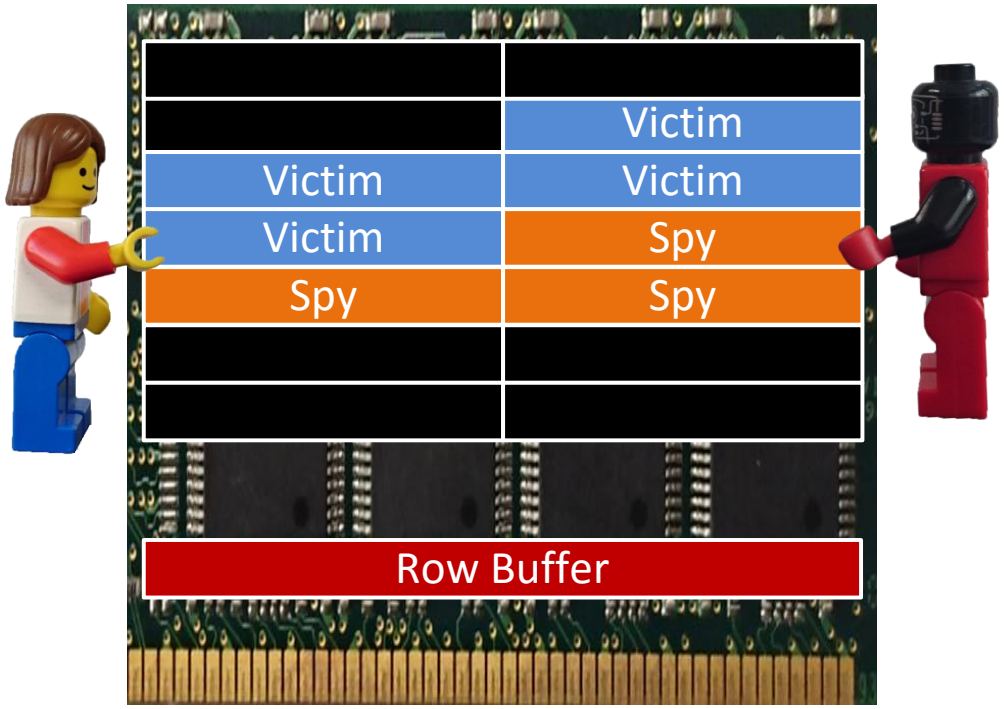
	Performance	Cross-CPU	No Shared Memory
Ours	**2.1 Mbps**	✓	✓
Prime+Probe [2]	536 Kbps	✗	✓
Flush+Reload [2]	2.3 Mbps	✗	✗
Flush+Flush [2]	3.8 Mbps	✗	✗
Memory Bus Contention [3]	746 bps	✓	✓
Deduplication [4]	90 bps	✓	✗

Low-noise side-channel attack

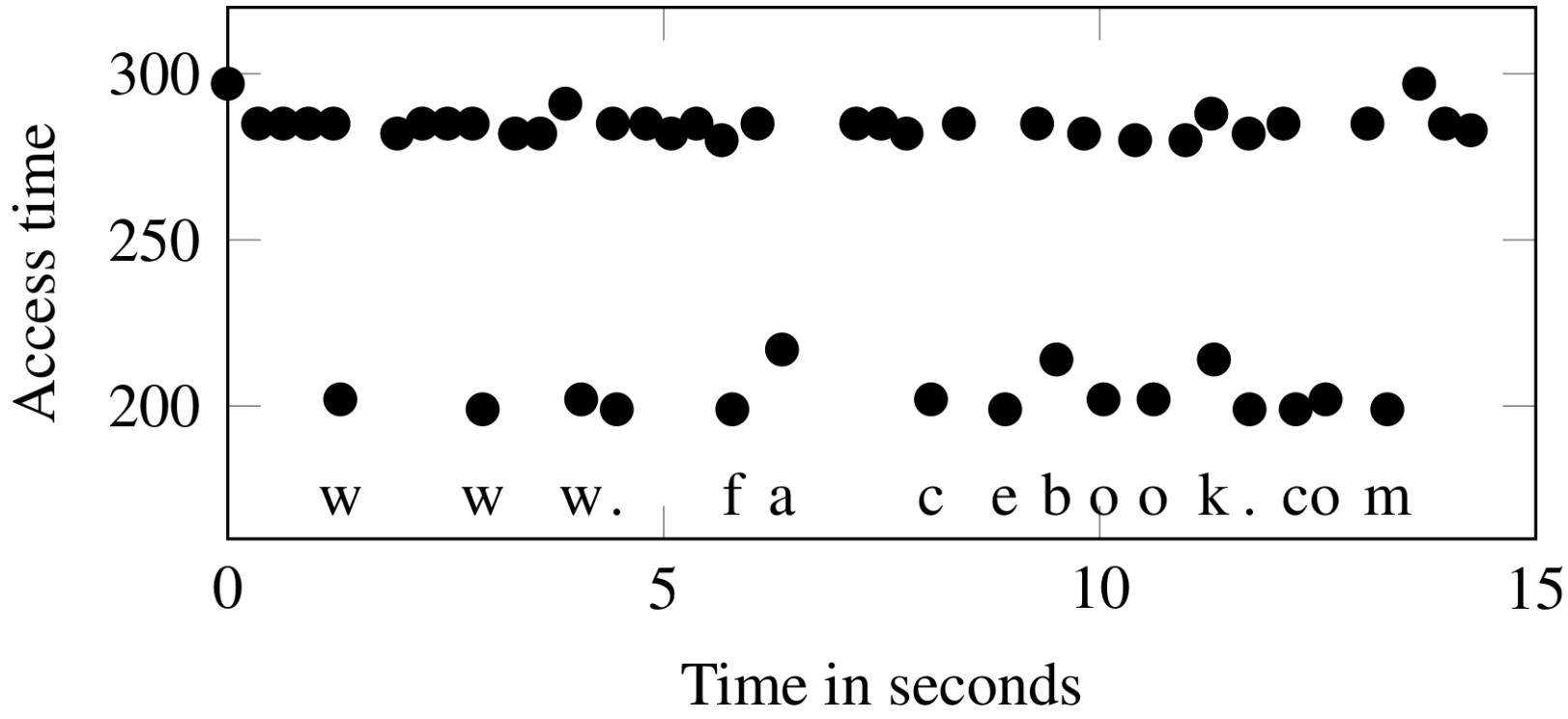


Spying on Memory Accesses

- Memory in the same row/bank
  - row size 8 KB / page size 4 KB
- Spy activates conflict row
- Victim computes and possibly accesses shared row
- Spy accesses shared row
  - fast→ row hit → victim access



Example



Keystrokes in Firefox address bar

Implementation

- high spatial accuracy (down to 512 B)
- very low number of false positives
  - monitor single events

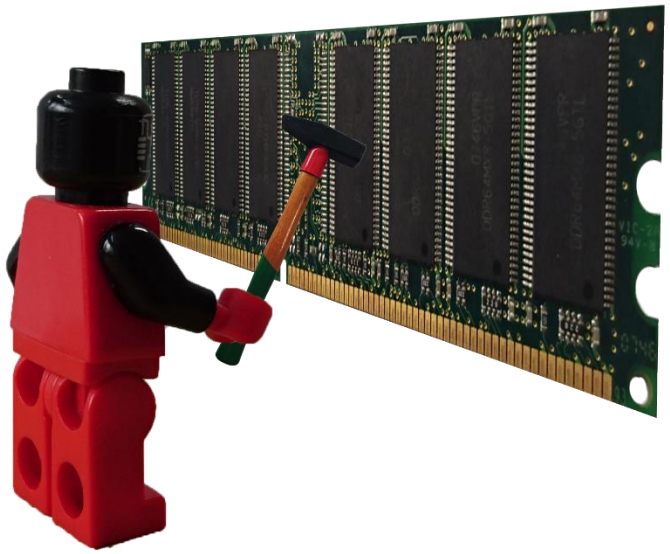
- Finding addresses: template attack [1]
  - automatic location of vulnerable addresses
  - scan large fraction of memory (4 KB pages)

Countermeasures to DRAMA

- Restrictions of
  - `rdtsc`
  - `clflush`
- Multi-CPU: separating DRAM for tenants
  - only access to CPU-local memory
  - degradation into single-CPU system
- Detection via high number of cache misses / row conflicts

Improving Attacks - Rowhammer

- Rowhammer
  - inducing bit flips in DRAM
  - by quickly switching rows
  - requires addressing functions



- First documented bit flips on DDR4
  - Jan. 2016

The End

… of Romeo and Juliet

Source code for reverse-engineering tool and side-channel attack at

`https://github.com/IAIK/drama`

DRAMA: Exploiting DRAM Addressing for Cross-CPU Attacks

Peter Pessl, Daniel Gruss, Clémentine Maurice, Michael Schwarz, Stefan Mangard
IAIK, Graz University of Technology, Austria

Usenix Security 2016, August 11

Bibliography

[1] Gruss, Spreitzer, Mangard. Cache Template Attacks: Automating Attacks on Inclusive Last-Level Caches. In Usenix Security 2015

[2] Gruss, Maurice, Wagner, Mangard. Flush+Flush: A Fast and Stealthy Cache Attack. In DIMVA'16

[3] Wu, Xu, Wang. Whispers in the Hyper-space: High-bandwidth and Reliable Covert Channel Attacks Inside the Cloud. In Usenix Security 2012

[4] Xiao, Xu, Huang, Wang. Security implications of memory deduplication in a virtualized environment. In DSN'13