Performance Assessment of Data Prefetchers in High Error Rate Technologies

Nikos Foutris, Dimitris Gizopoulos, Athanasios Chatzidimitriou
University of Athens, Dept. of Informatics & Telecom., Greece

John Kalamatianos, Vilas Sridharan
AMD, USA

AMD

Clereco
Cross-Layer Early Reliability Evaluation for the Computing cOntinuum

SEVENTH FRAMEWORK PROGRAMME

cal@di

Outline

Background

Motivation

Scope

Performance Impact

Performance Variability

Conclusions

Background

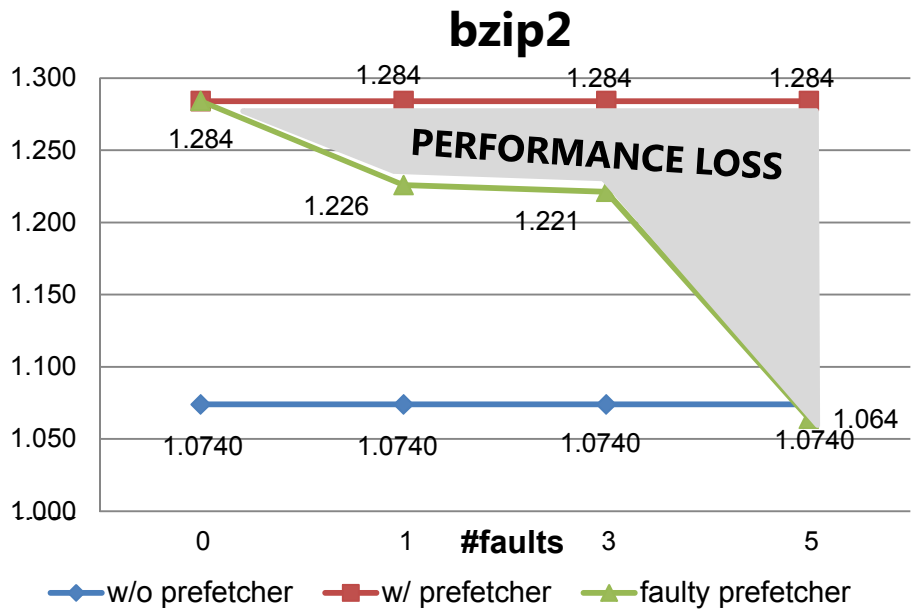
- Hide latency of memory accesses
  - cache memory
  - **prefetcher**
- Extreme integration levels
- Near-threshold voltage operation
- **High SRAM cell failure probability**
  - pfail 10^{-6} to 10^{-4} (single-bit failure)
  - 5,000-bit SRAM @ 12nm → 7.25E-1 (cum. prob. of 5-faults)
- **Measure prefetcher's reliability**

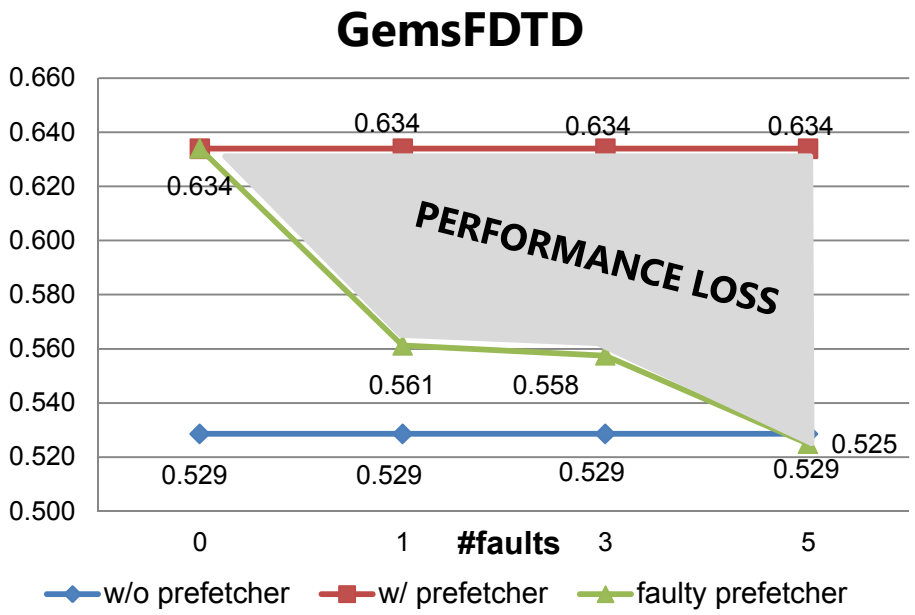
Background (2)

- **Stride data prefetcher**
  - prefetch data based on memory access patterns
- **Program correctness unaffected***
- **Impact performance**:
  - reduce prefetch coverage
  - reduce prefetch accuracy
  - degrade prefetch timeliness

*N. Foutris, D. Gizopoulos, J. Kalamatianos, V. Sridharan, "Assessing the impact of hard faults in performance components of modern microprocessors," ICCD, 2013.

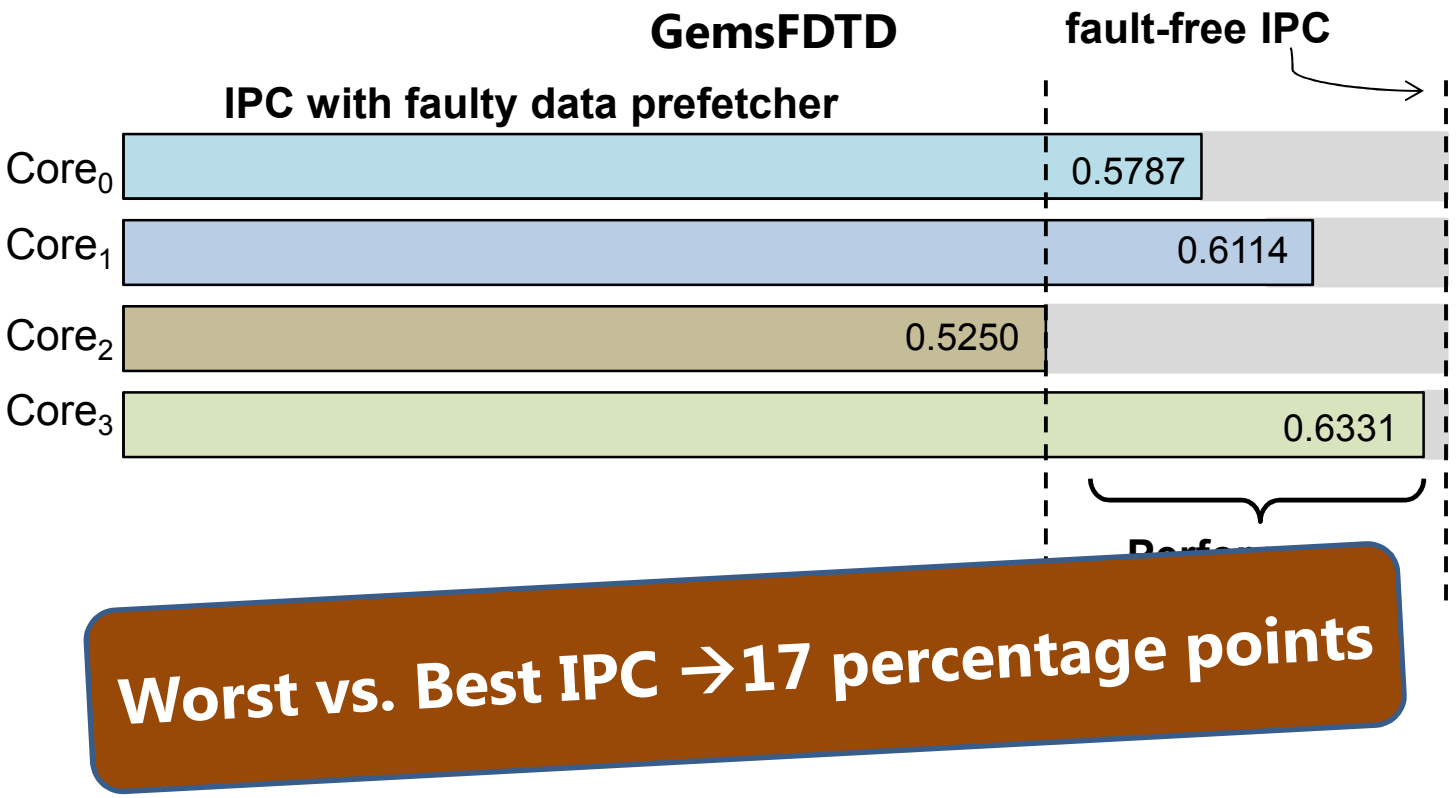
Motivation





severely impacts performance (more than 17%)

Motivation (2)



Worst vs. Best IPC →17 percentage points

Scope

- x86-64 architecture
  - **L1 stride data prefetcher**
- Reliability evaluation
  - **multi-bit permanent fault**
  - prefetch table, prefetch input & request queue
- Performance degradation
- Performance variability

Experimental Setup

- **PTLsim** x86 architectural simulator
  - L1 data cache stride prefetcher

Parameter	Setting
Prefetch input queue	8 entries
Prefetch table	64 entries, direct-mapped, PC-indexed
Confidence size	3 bits (threshold: 2)
Stride size	5 bits
Prefetch distance	Address + Stride, Address +2*Stride
Prefetch request queue	8 entries

- 29 SPEC CPU2006 (SimPoint-based)
- 120M committed instr. (20M warm-up)

Experimental Setup (2)

- **single, triple, quintuple-faults**
- tag, load address, stride, confidence, LRU, valid, prefetch request queue, prefetch input queue
- **26,100** fault simulation experiments
  - 99% confidence level, 3% error margin

Workload profiling

- SPEC CPU2006 benchmarks
- Stream of memory access patterns
- **IPC impact** from the data prefetcher
  - on average **6.85% IPC speed-up**
- Prefetch-**friendly** vs. Prefetch-**unfriendly**
  - IPC change **greater than** average speedup
  - IPC change **less than** average speedup

Prefetch-unfriendly

unfriendly benchmarks	IPC (%) Speed-up
perlbench	2.58
gcc	1.89
mcf	0.16
milc	
groma	
cactusAD	
namd	0.34
gobmk	0.60
povray	0.42
calculix	4.73

unfriendly benchmarks	IPC (%) Speed-up
hmmer	0.62
sjeng	0.07
astar	3.01
sphinx3	0.67
xalancbmk	3.80
average	1.780

Average IPC speedup 1.780%
(overall speedup 6.85%)

Low impact on IPC

*milc's IPC i

Prefetch-friendly

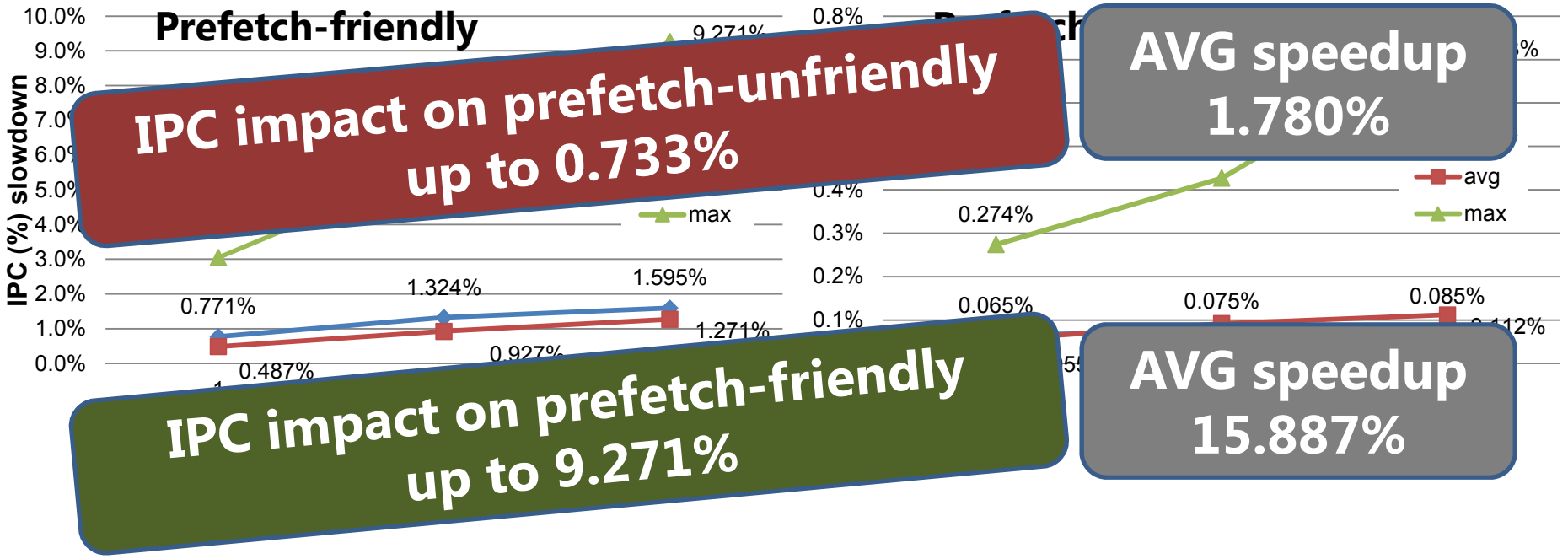
friendly benchmarks	IPC (%) Speed-up
bzip2	19.99
[illegible]	[illegible]
[illegible]	7.55
dealII	9.61
soplex	9.50
GemsFDTD	19.66
libquantum	17.20
tonto	15.91
wrf	34.59

Average IPC speedup 15.887% (overall speedup 6.85%)

High impact on IPC

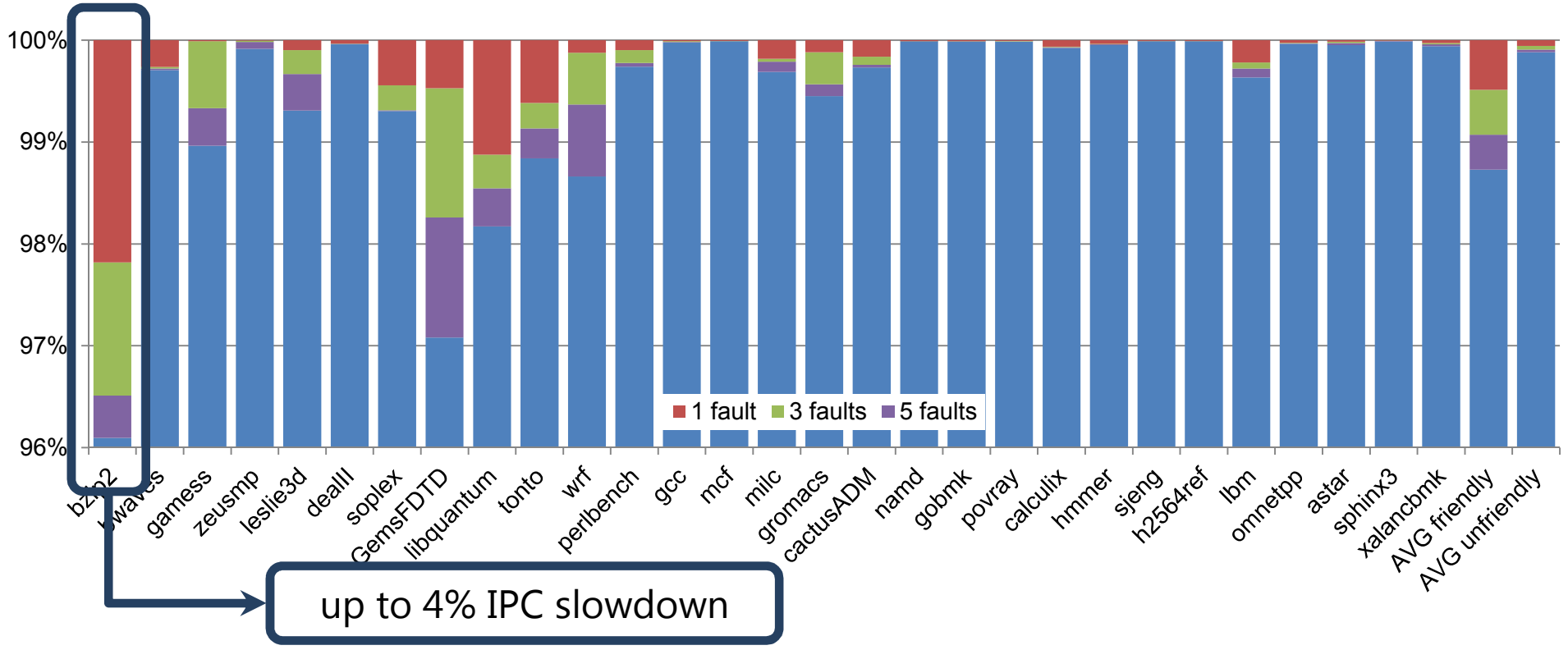
Performance impact

- **IPC impact** due to **multi-bit faults** on **PT**



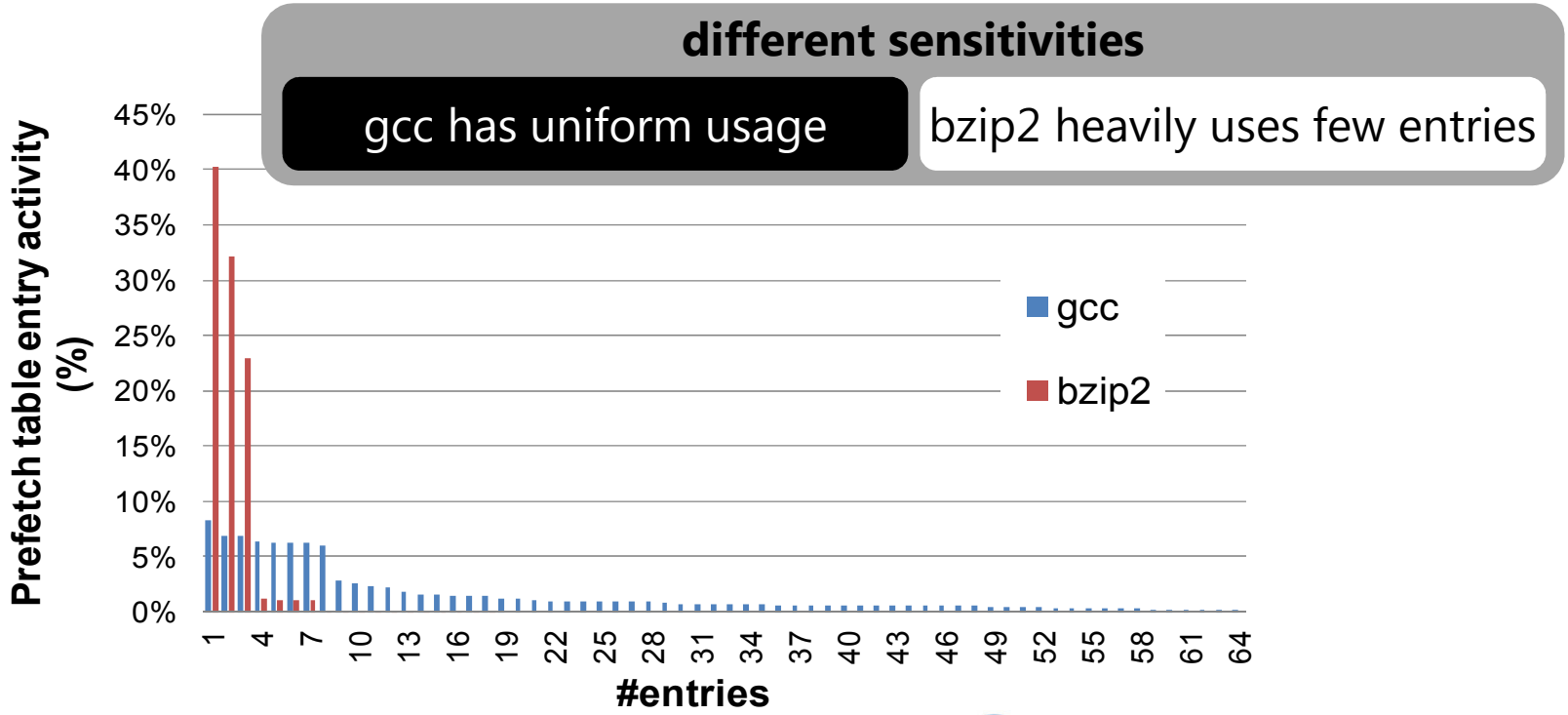
Performance impact (2)

- Average IPC (%) slowdown



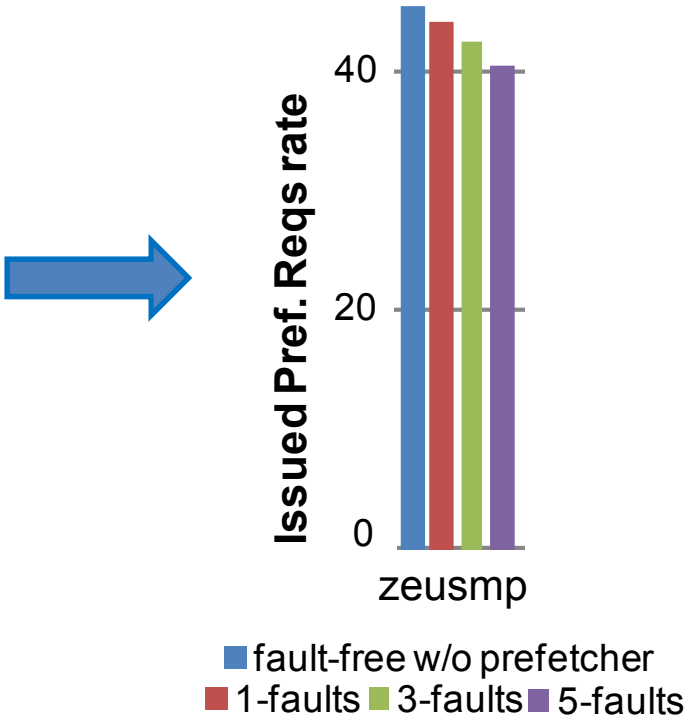
Performance impact (3)

- **Distribution of prefetch addresses** across prefetch table entries

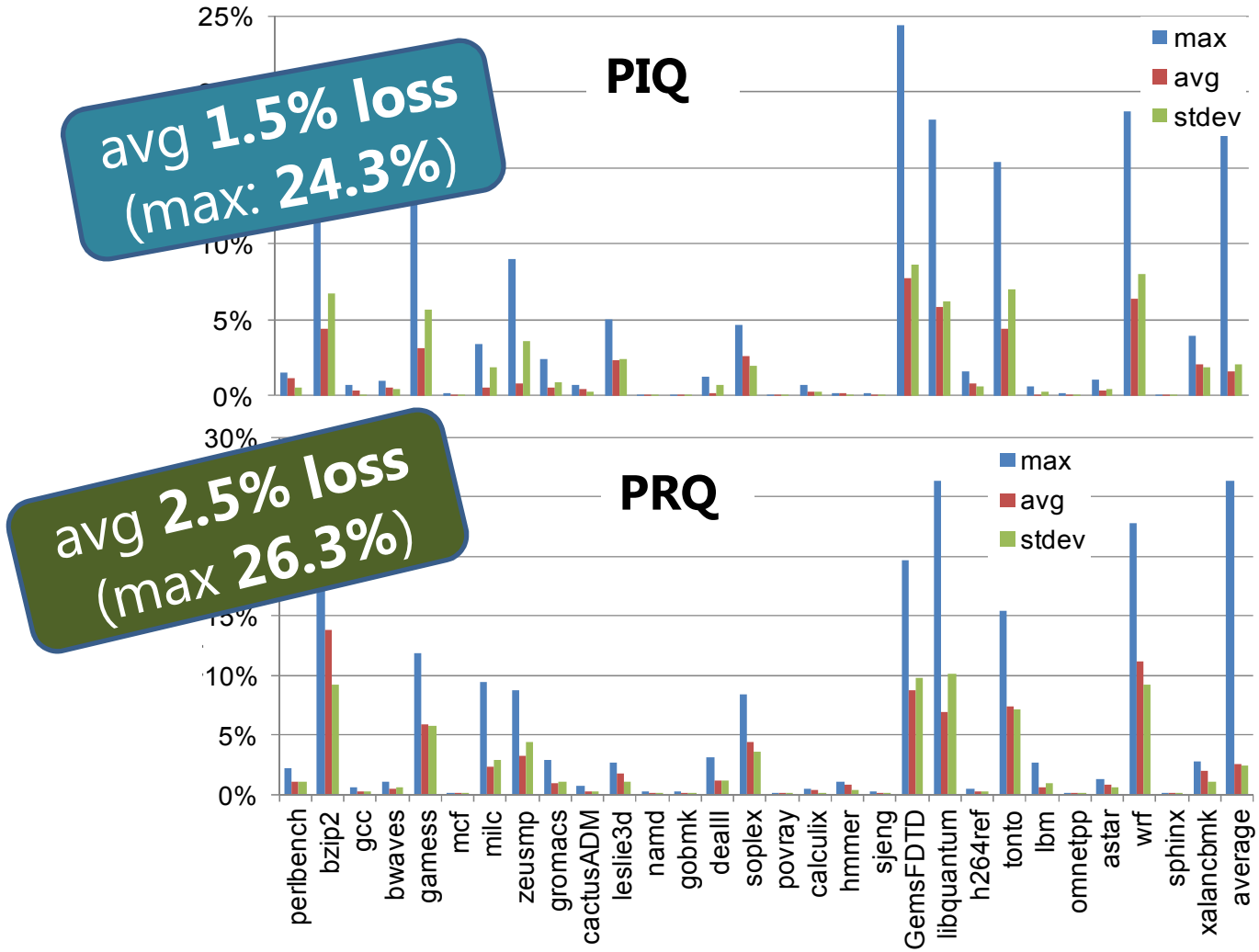


Performance impact (4)

- **Prefetch request rate**
  - drops from 22 to 20 per 1,000 committed instr.
- **L1 miss rate**
  - increases from 26 to 27 per 1,000 committed instr.



Performance impact (5)



Performance impact (6)

- Fault location determines the extent of the performance impact
- Prefetch Input Queue
  - entries **uniformly** utilized
    - top to bottom: 35% …7%
- Prefetch Request Queue
  - High utilization of **top entries**
    - top 3: 95% activity

Findings summary

- Performance **severely degraded** from faulty L1 cache stride data prefetcher:
  - up to 24% due to PIQ faults
  - up to 26% due to PRQ faults
  - up to 18% due to PT faults
- **Prefetch throttling, Cache pollution**
- Benchmarks with **high repeatable** address patterns are the **most susceptible to IPC loss**

Performance variability

- **Large variation** in IPC under the presence of **different single faults**
- All cores are affected by **the same number** of faults:
  - up to 7% for the PT
  - up to 24% for the PIQ
  - up to 26% for the PRQ

Performance variability (2)

- Each core is affected by **different number** of faults (multi-bit faults):
  - up to 18% for PT
  - up to 24% for PIQ
  - up to 26% for PRQ

Performance variability (3)

- Difference between best- and worst-case is **not an outlier**
- **Standard deviation**:
  - up to 4.5% for PT
  - up to 2.0% for PIQ
  - up to 2.4% for PRQ

Conclusions

- L1 cache stride data **prefetcher**
- Multi-bit permanent faults
- Severely impacts performance
  - **up to 26% slowdown**
- Increases inter-core performance variability
  - **stdev more than 4.5%**
- **Protect prefetcher**

Thank You!

?