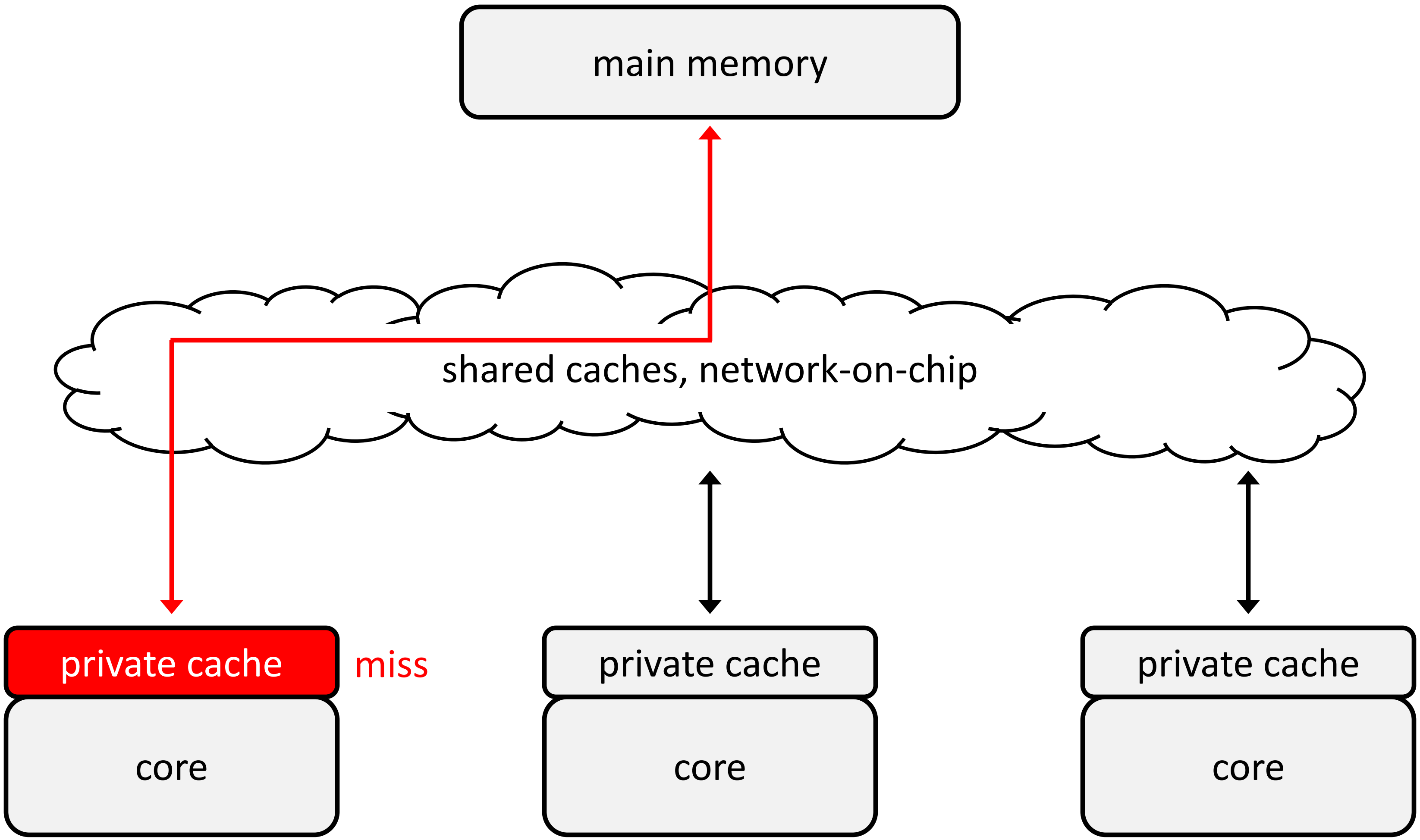
Load Value Approximation: Approaching the Ideal Memory Access Latency

Joshua San Miguel

Natalie Enright Jerger

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING

Chip Multiprocessor



Approximate Data

Many applications can tolerate inexact data values.

- In approximate computing applications, 40% to nearly 100% of memory data footprint can be approximated [Sampson, MICRO 2013].

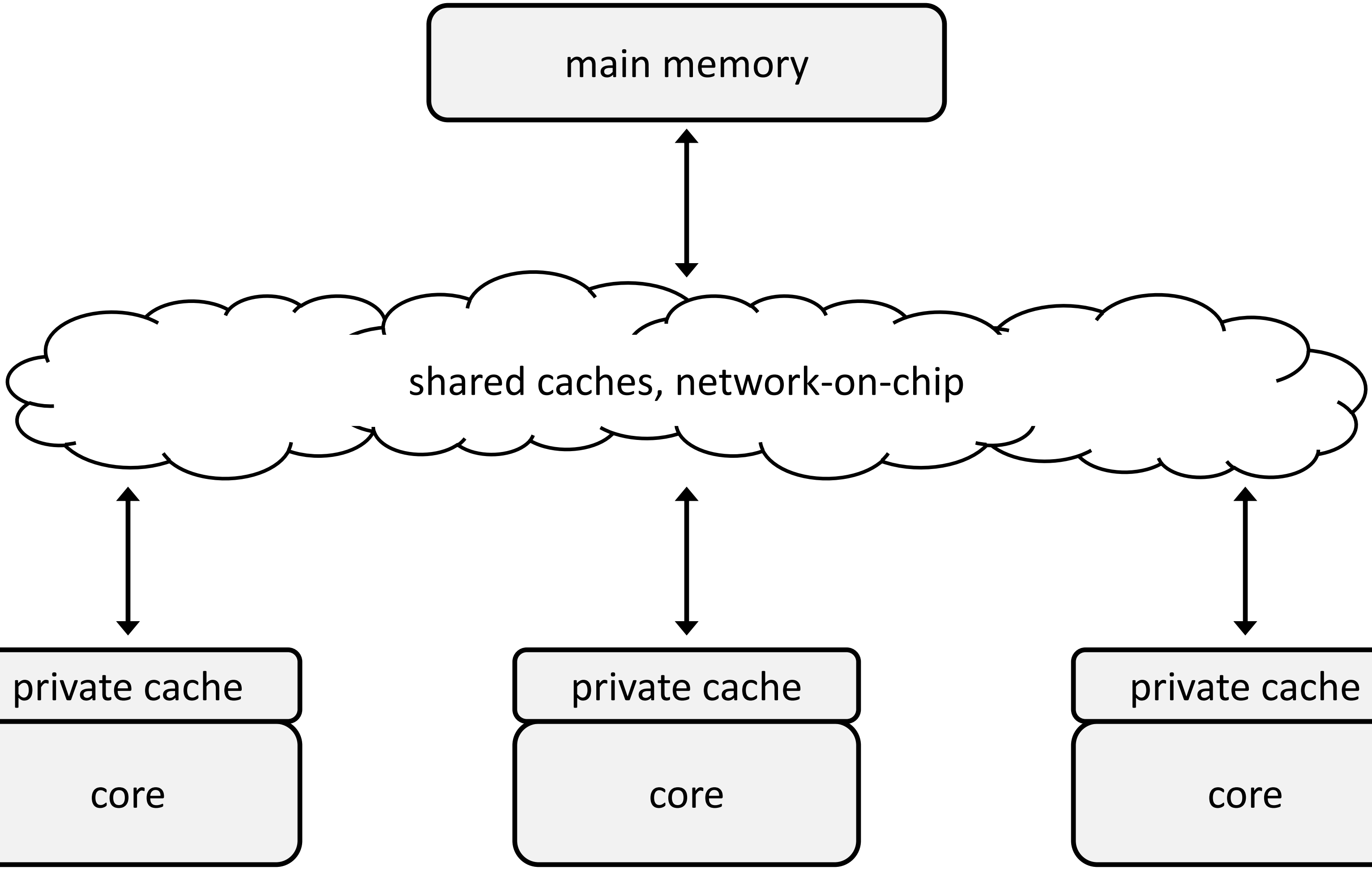
Approximate data storage:

- Reducing SRAM power by lowering supply voltage [Flautner, ISCA 2002].
- Reducing DRAM power by lowering refresh rate [Liu, ASPLOS 2011].
- Improving PCM performance and lifetime by lowering write precision and reusing failed cells [Sampson, MICRO 2013].

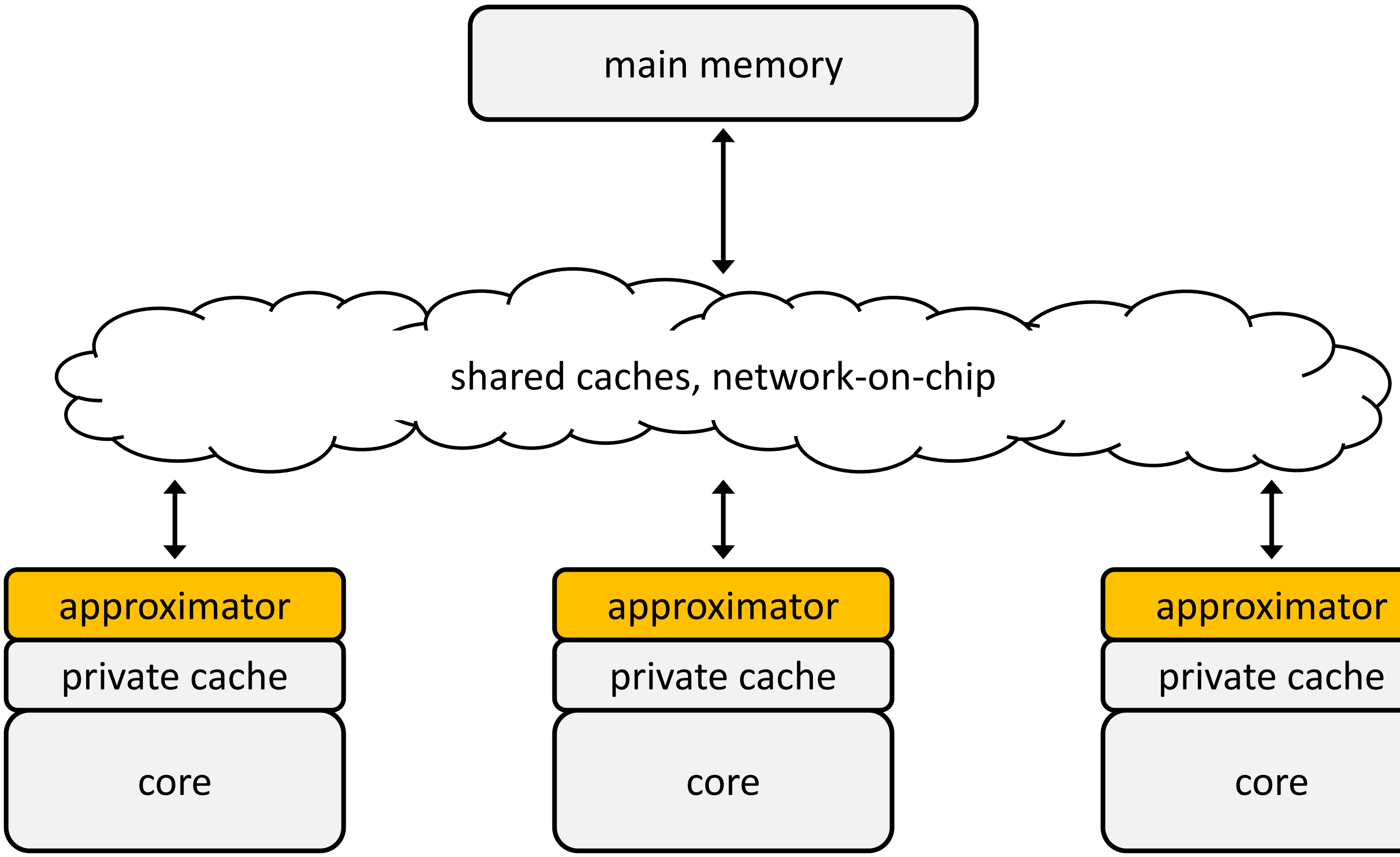
Outline

- Load Value Approximation
- Approximator Design
- Evaluation
- Conclusion

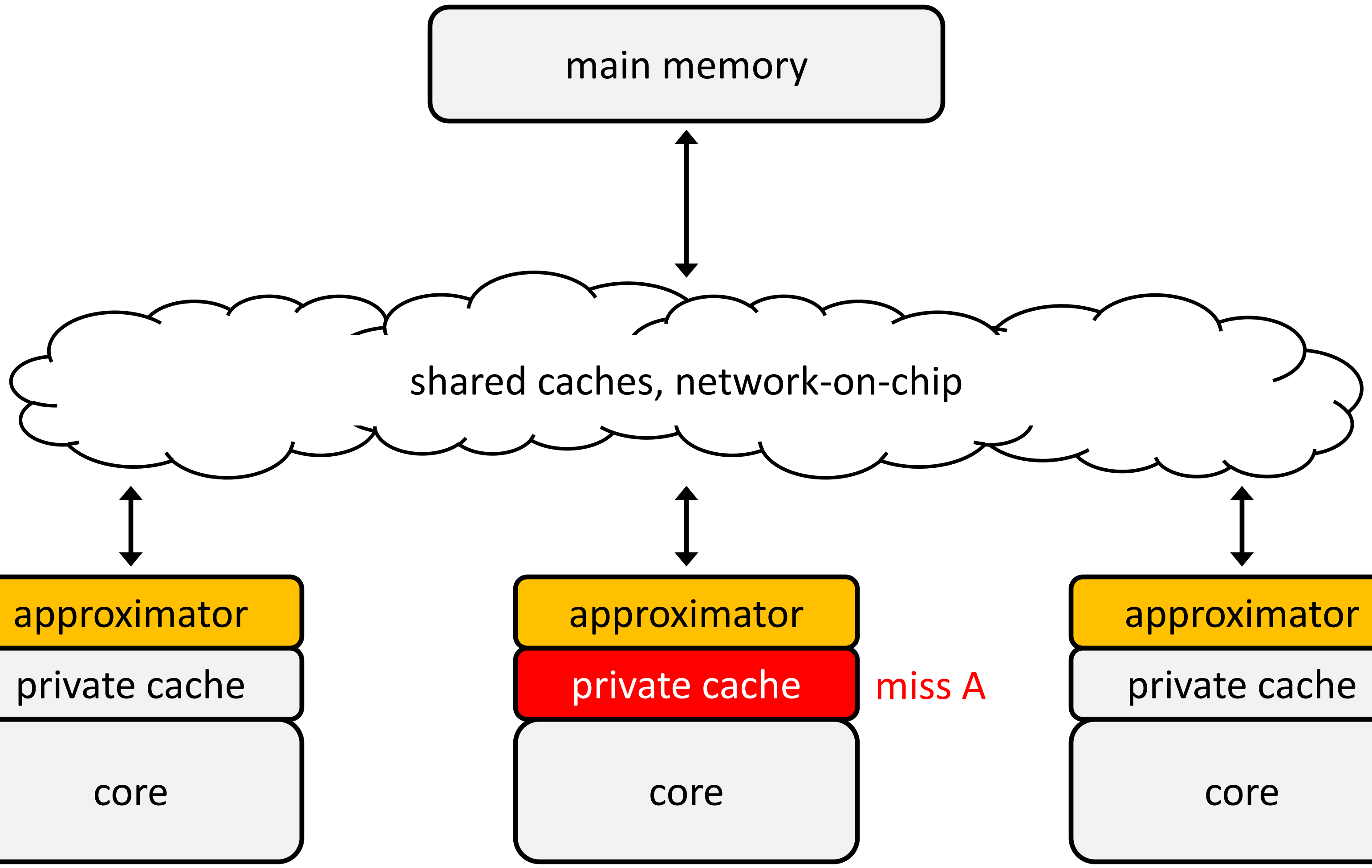
Load Value Approximation



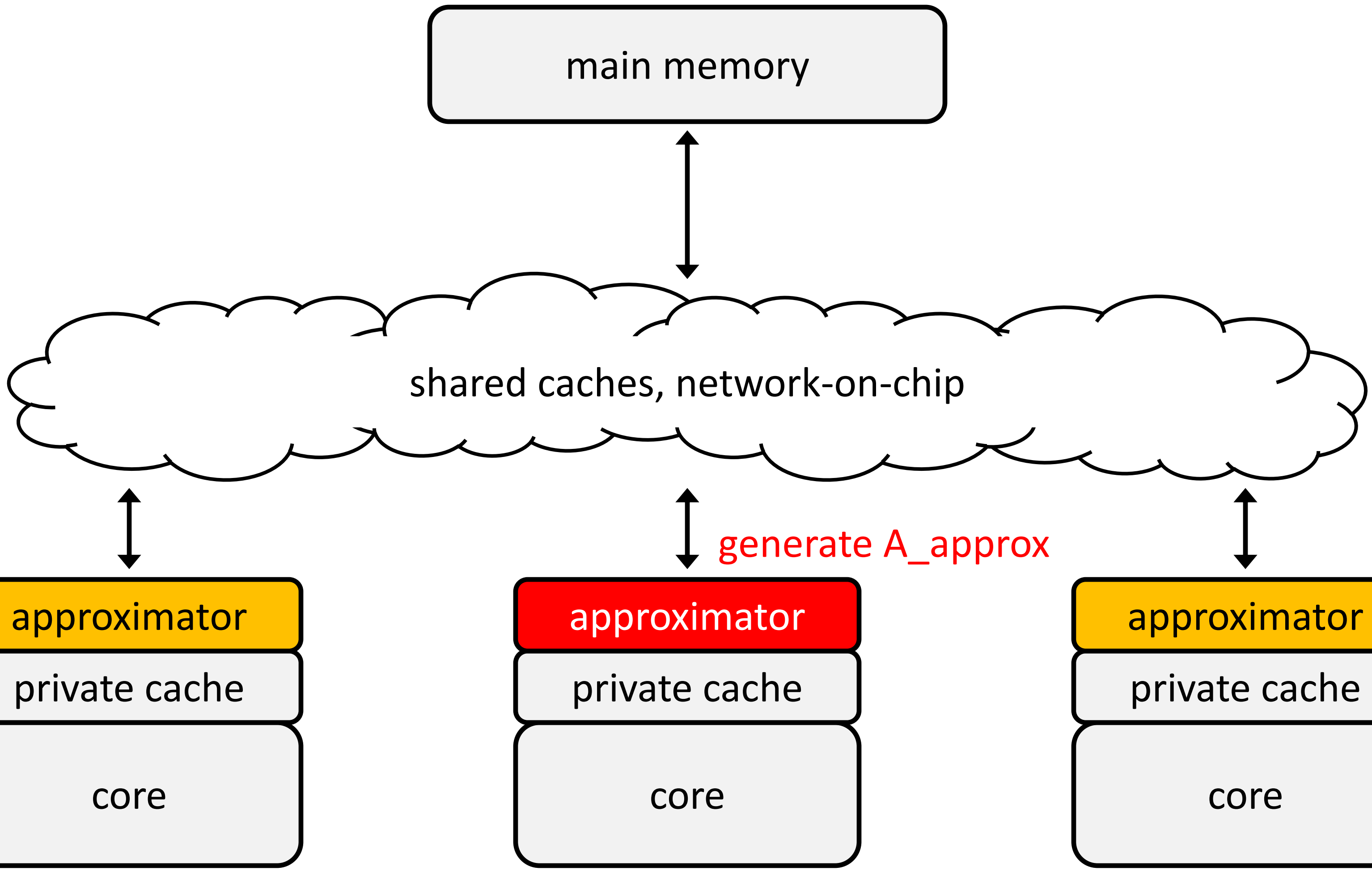
Load Value Approximation



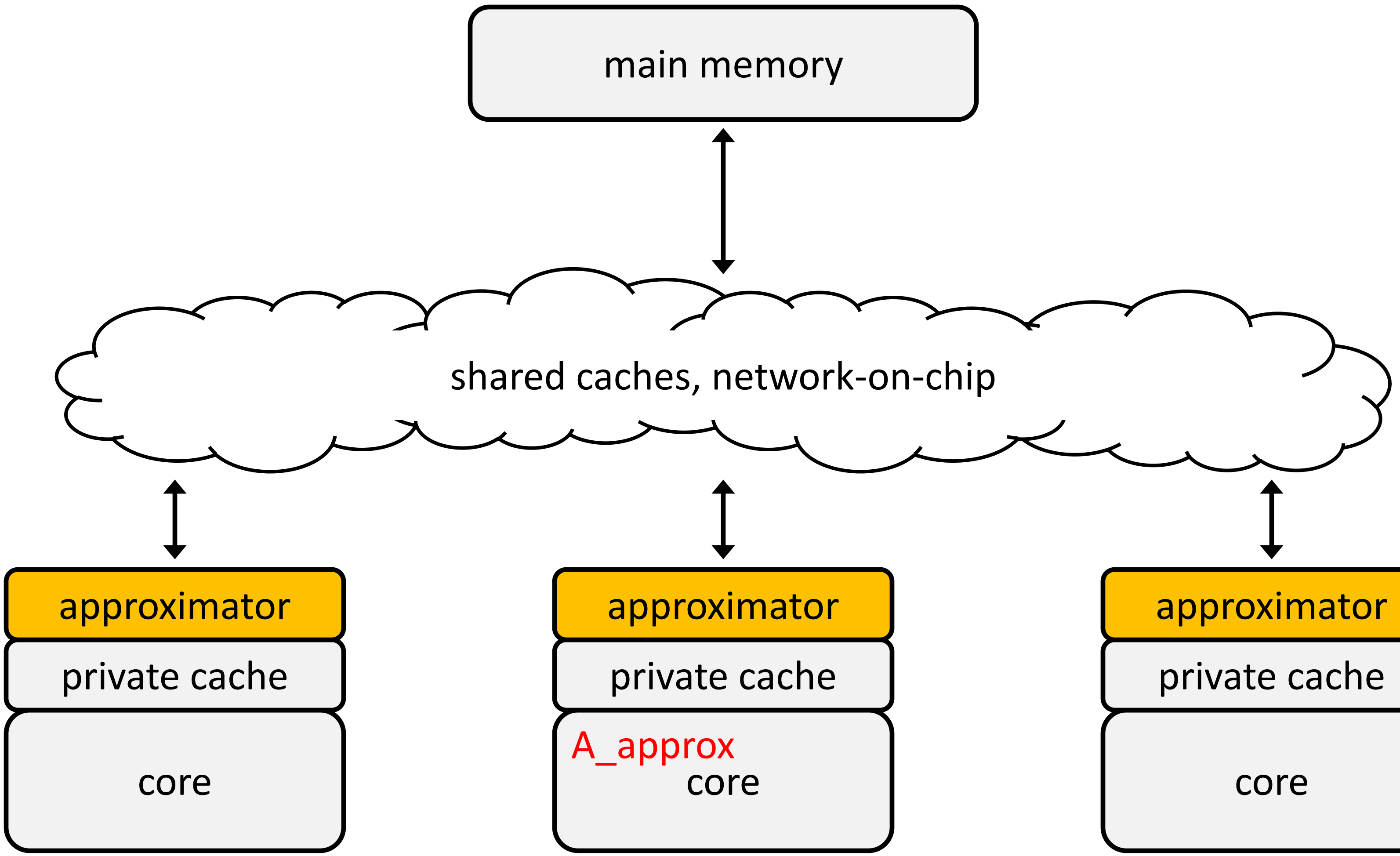
Load Value Approximation



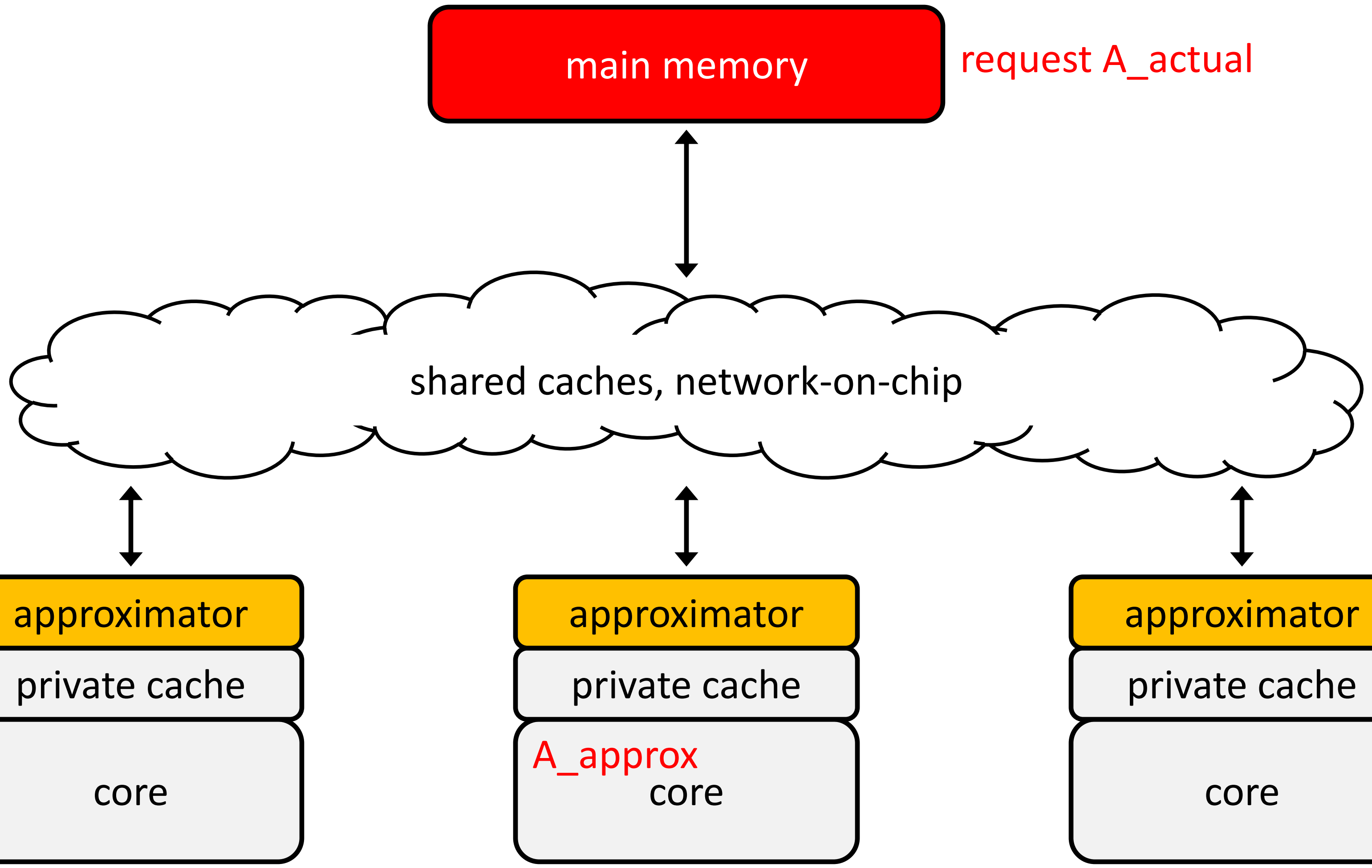
Load Value Approximation



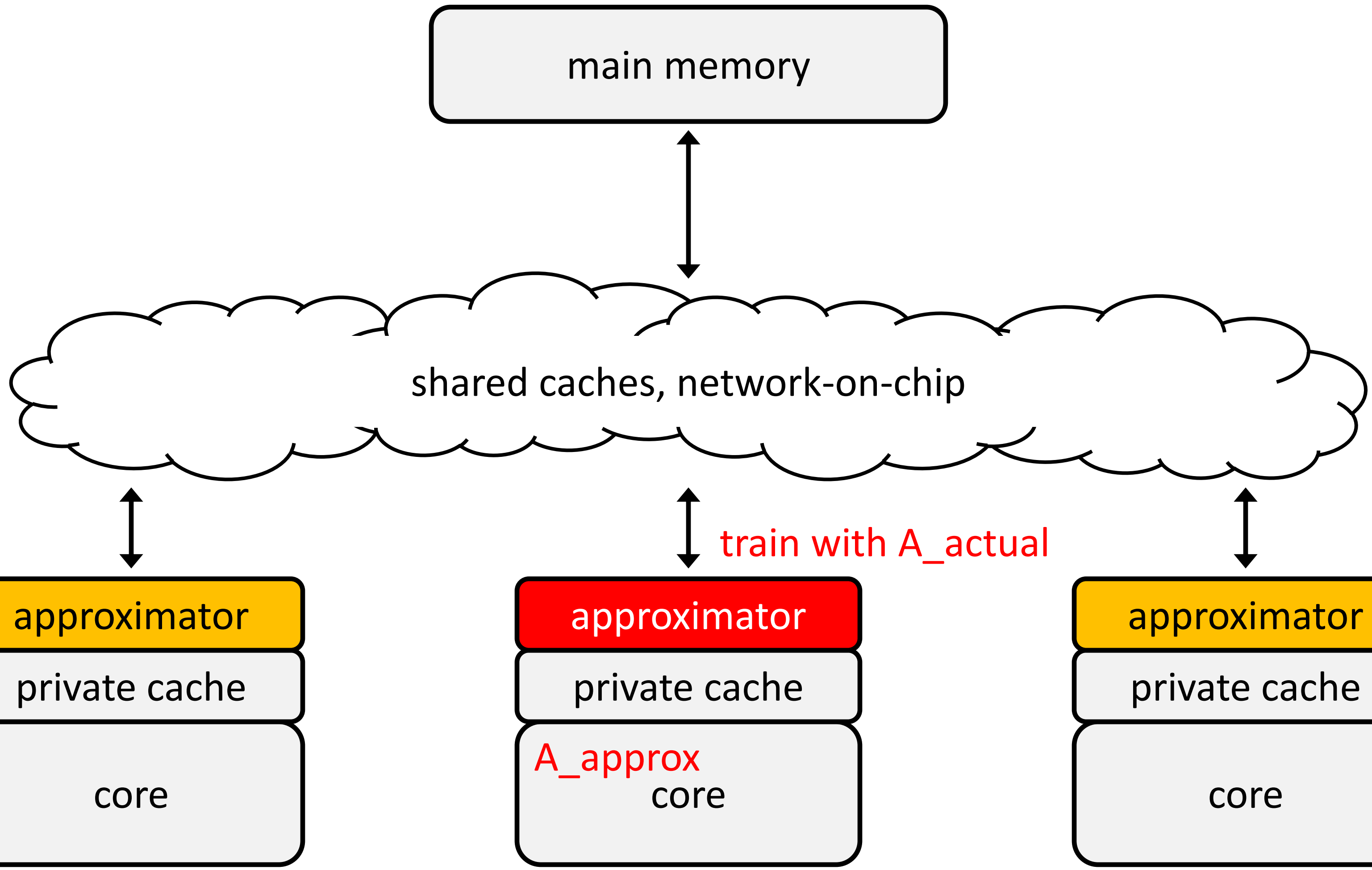
Load Value Approximation



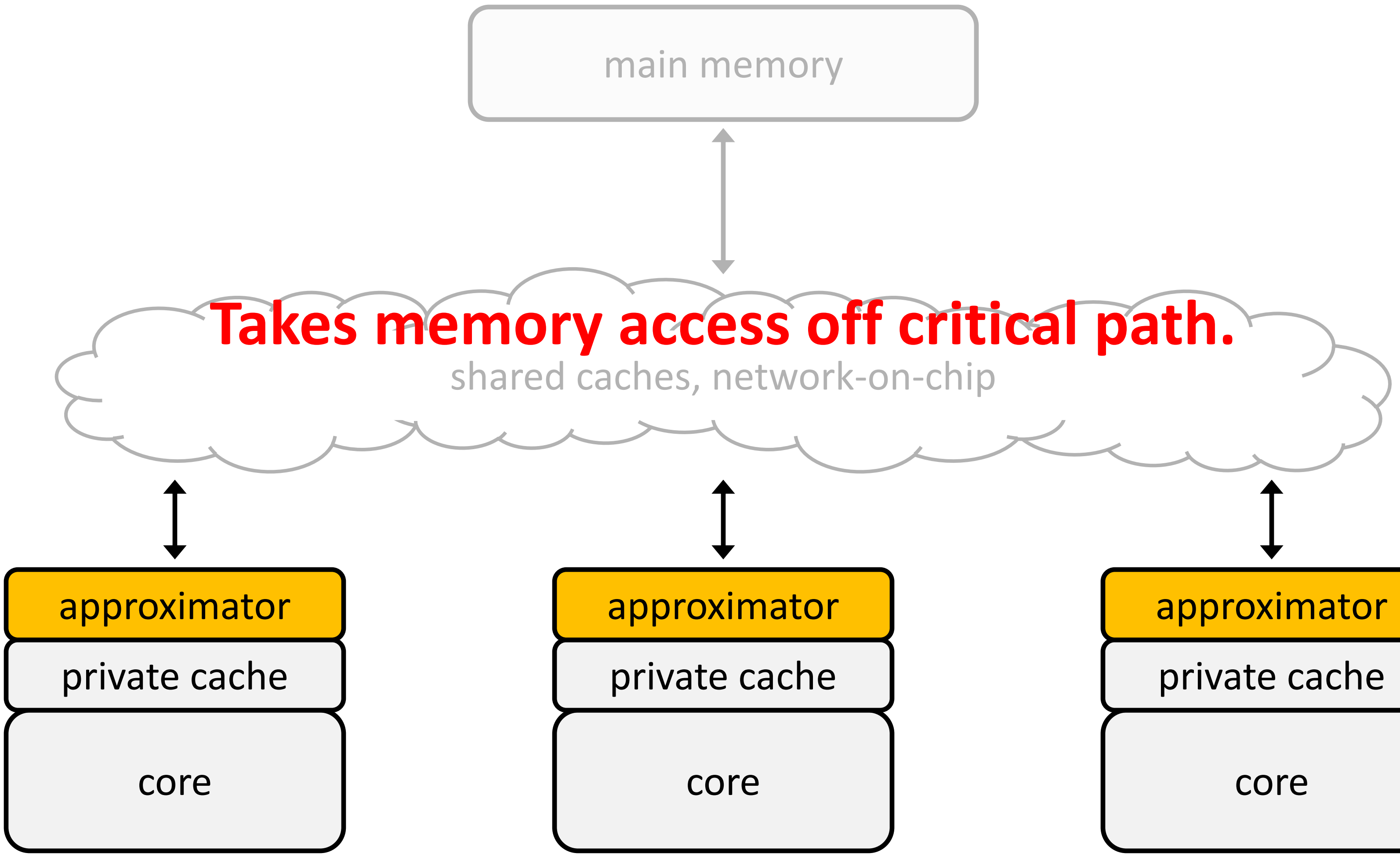
Load Value Approximation



Load Value Approximation

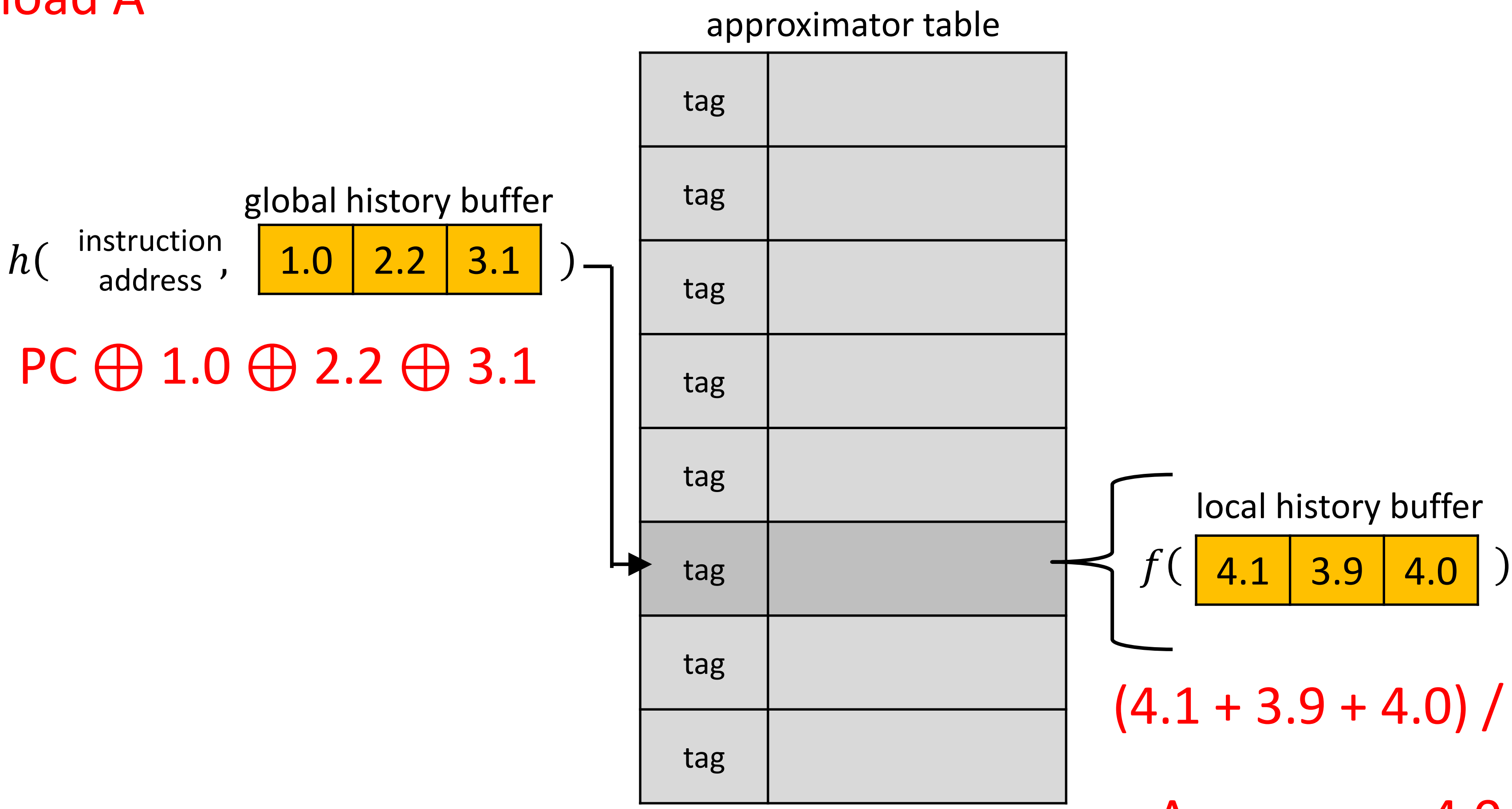


Load Value Approximation



Approximator Design

load A



PC ⊕ 1.0 ⊕ 2.2 ⊕ 3.1

(4.1 + 3.9 + 4.0) / 3

A_approx = 4.0

Approximator Design

Load value approximators overcome the challenges of traditional value predictors:

- No complexity of tracking speculative values.
- No rollbacks.
- High accuracy/coverage with floating-point values.
- More tolerant to value delay.

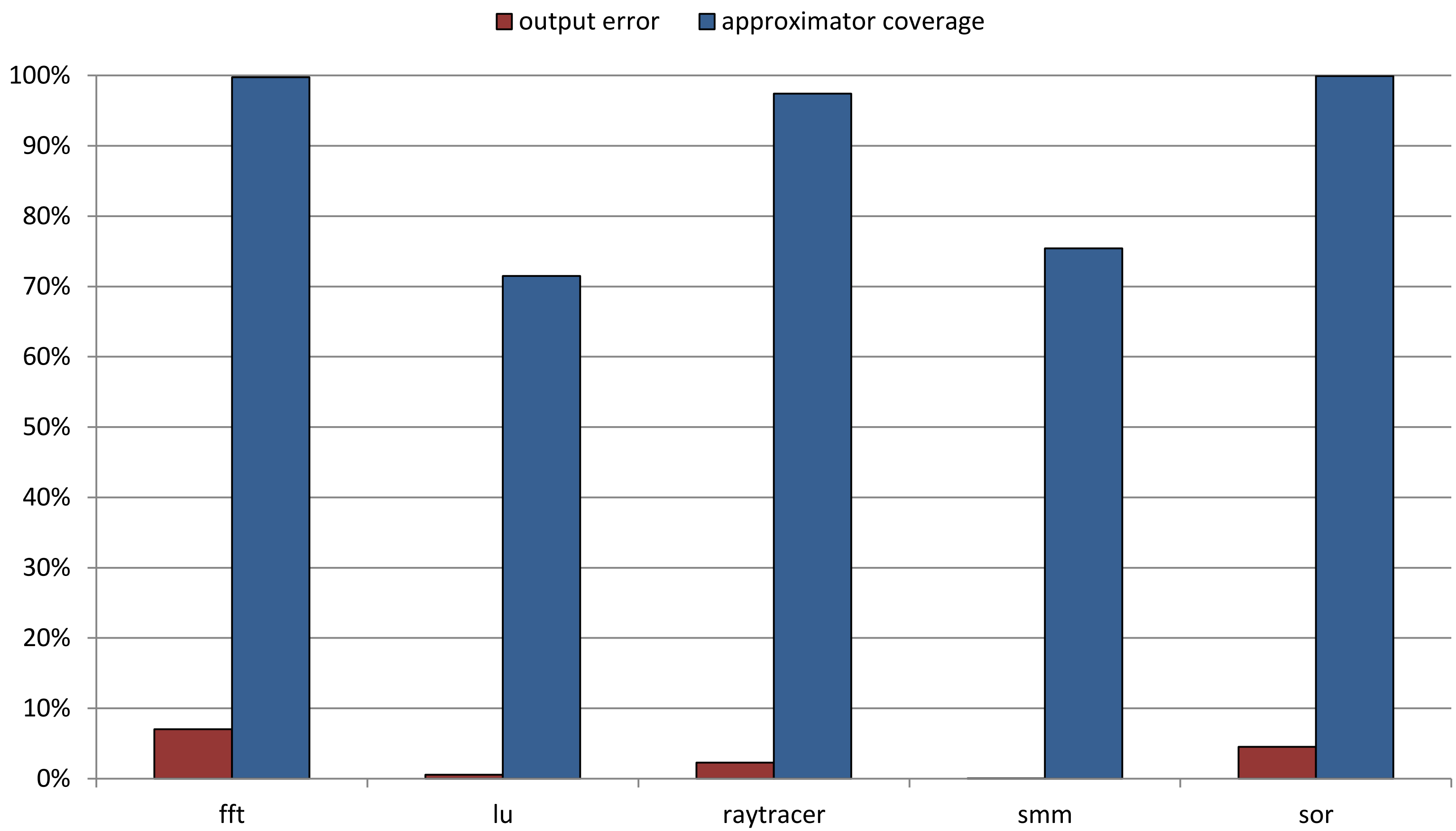
Evaluation

EnerJ framework [Sampson, PLDI 2011]:

- Program annotations to distinguish approximate data from precise data.
- Evaluate final output error and approximator coverage.

benchmark	GHB size	LHB size	approximator size
fft	0	2	49 kB
lu	3	1	32 kB
raytracer	1	1	32 kB
smm	5	1	32 kB
sor	0	2	49 kB

Evaluation



output error
approximator coverage
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
fft
lu
raytracer
smm
sor

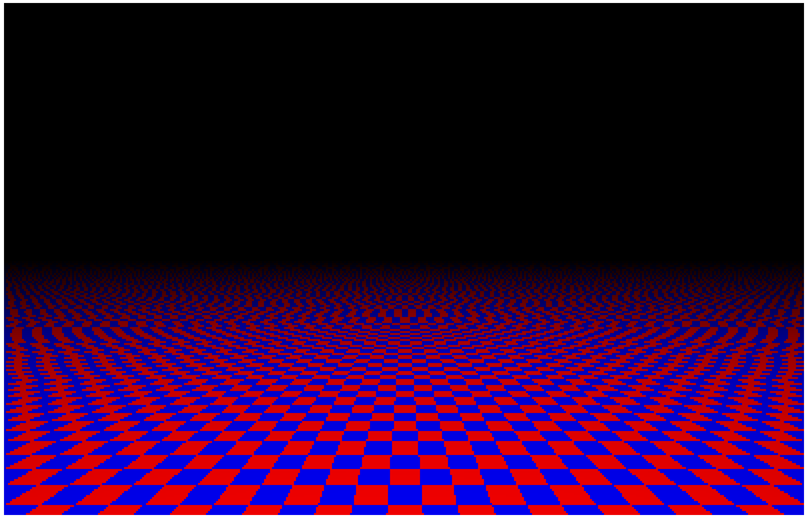
Conclusion

Future work:

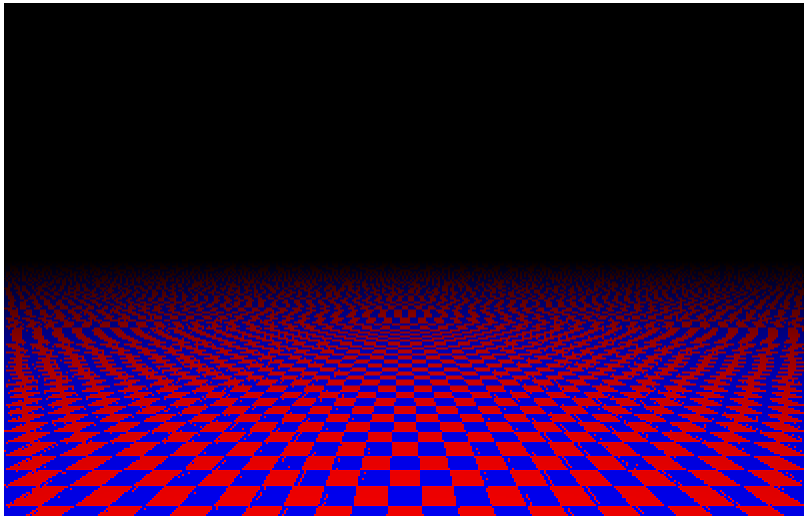
- Further explore approximator design space (dynamic/hybrid schemes, machine learning).
- Measure speedup of load value approximation using full-system simulations.
- Measure power savings (low-power caches/NoCs/memory for approximate data).

Low-error, high-coverage approximators allow us to approach the ideal memory access latency.

Thank you



baseline (precise) - raytracer



load value approximation - raytracer