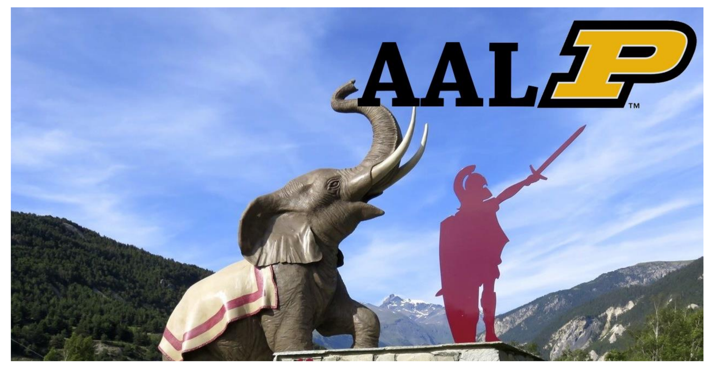
A Quantitative Evaluation of Contemporary GPU Simulation Methodology

Akshay Jain, Mahmoud Khairy, Timothy G. Rogers
Purdue University, Department of Electrical and Computer Engineering
akshayj@alumni.purdue.edu, abdallm@purdue.edu, timrogers@purdue.edu

SIGMETRICS 2018



Accelerator Architecture Lab at Purdue

Introduction

- We quantify error in the popular GPU architecture simulator, GPGPU-Sim (1100+ citations).
- We demonstrate that the simulator's accuracy is highly dependent on the workload type.

Workload Characteristics

Type	IPC	L1 Miss	L2 Miss	Mem Util
cache-sensitive	Low	High	Mod	Low
memory-sensitive	Mod	High	High	High
compute-intensive	High	Mod	Mod	Mod
compute-balanced	Mod	Mod	Mod	Mod

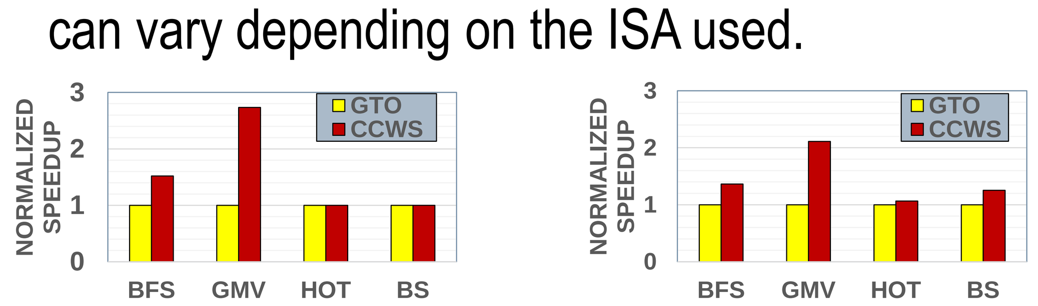
- Simulator supports 2 ISA representations of each app: a modern virtual ISA (vISA) and the machine ISA (mISA) for the decade old GT200 architecture.
- vISA has infinite registers, is generally less optimized and more abstract.
- vISA is called PTX and the GT200 mISA is PTXPlus

Code demonstrating difference between parameter loads, and higher code density of PTXPlus. Code segment taken from invert_mapping in kmeans [20]

PTX	PTXPlus
mov.u16 %rh1, %ctaid.x;	mov.u16 $r0.hi, %ntid.x;
mov.u16 %rh2, %ntid.x;	cvt.u32.u16 $r1, $r0.lo;
mul.wide.u16 %r1, %rh1, %rh2;	mad.wide.u16 $r1, %ctaid.x, $r0.hi, $r1;
cvt.u32.u16 %r2, %tid.x;	
add.u32 %r3, %r2, %r1;	
ld.param.s32 %r4, [__cudaparm..npoints];	set.le.s32.s32 $p0\|$o127, s[0x0020], $r1;
setp.le.s32 %p1, %r4, %r3;	
@%p1 bra $Lt_0_2050;	@$p0.ne retp;
ld.param.s32 %r5, [__cudaparm..nfeatures];	
mov.u32 %r6, 0;	
setp.le.s32 %p2, %r5, %r6;	set.le.s32.s32 $p0\|$o127, s[0x0024], $r124;
@%p2 bra $Lt_0_2562;	@$p0.ne retp;
...	
$Lt_0_2562:	
$Lt_0_2050:	
exit;	

Case Study

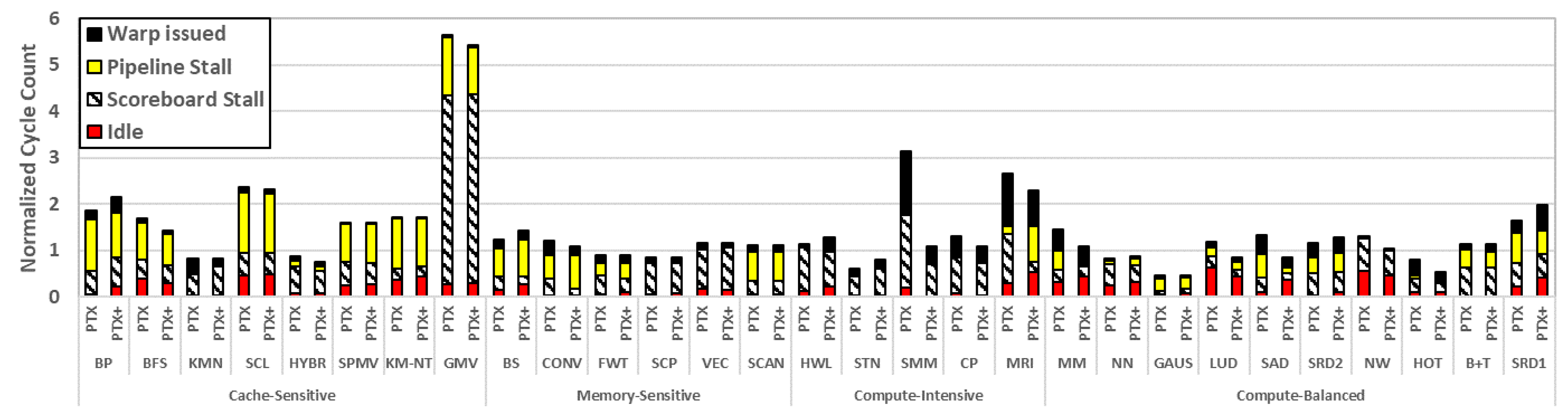
- Performance predictions for an architectural study can vary depending on the ISA used.



(a) Speedup using vISA (b) Speedup using mISA

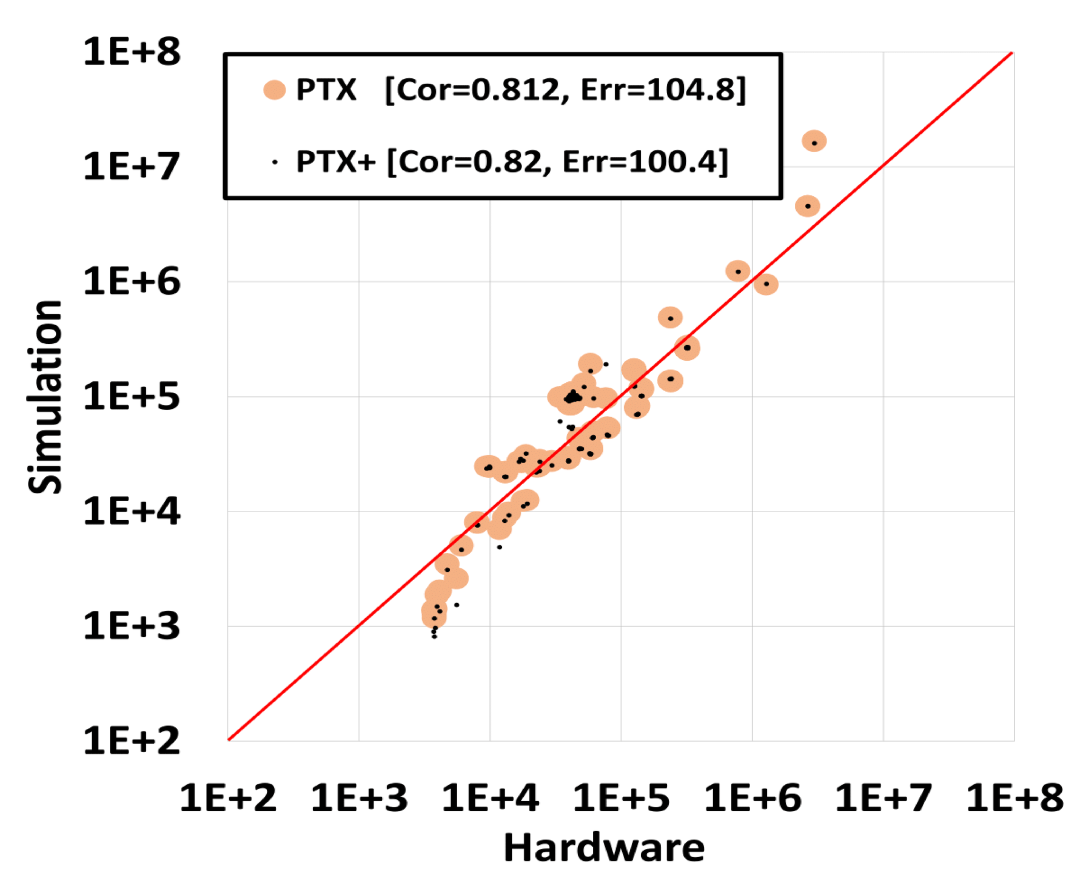
- mISA models the memory traffic due to register spilling (as it is register allocated).
- We see performance improvement for HOT and BS only with the mISA.

Experimental Results

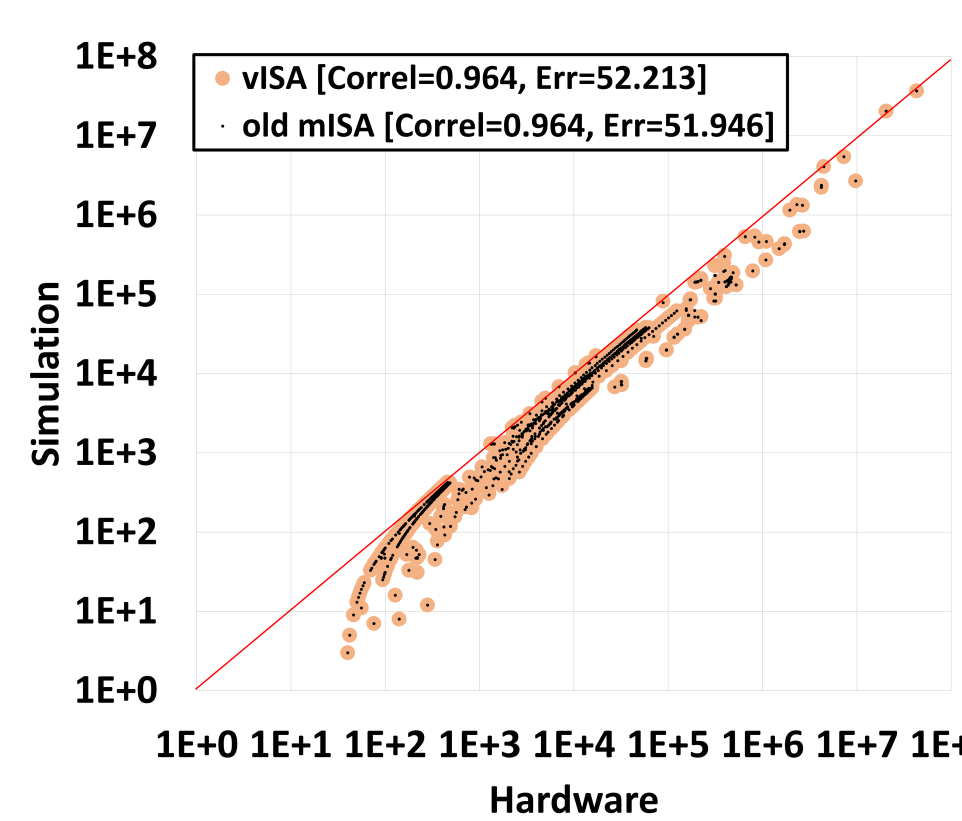


Execution cycles in simulation relative to hardware. Figure shows a breakdown of stalls and useful execution from GPGPU-Sim.

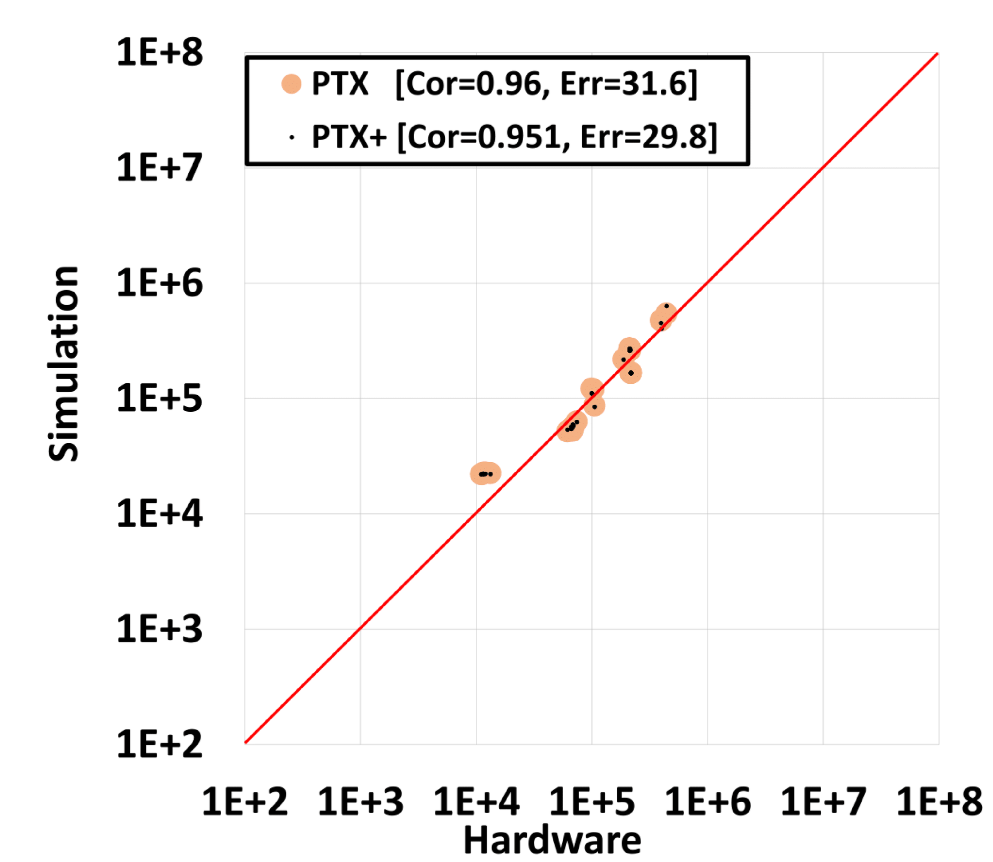
Cache Sensitive



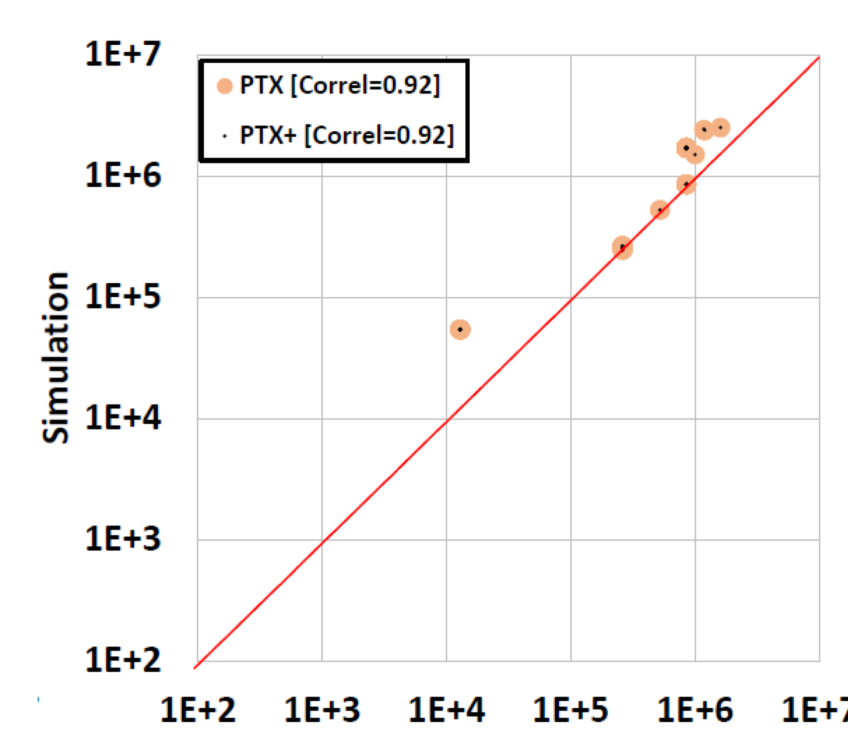
- Low Correlation, high error
- Over estimation and high error due to inaccurate caching model.
- Cache system modeling needs to be improved.



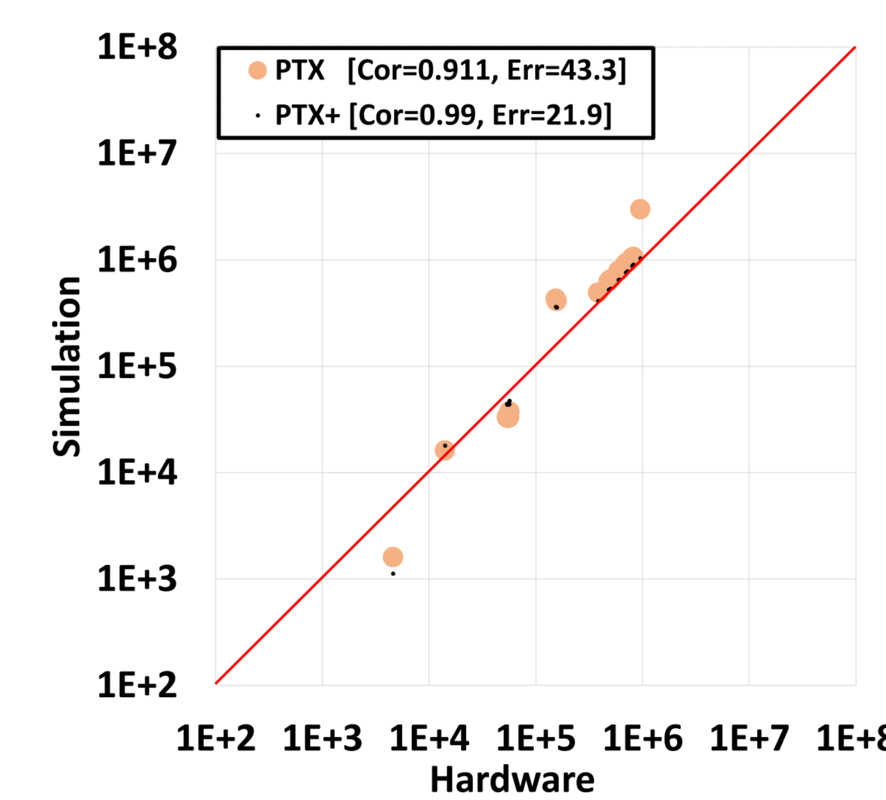
Memory Sensitive

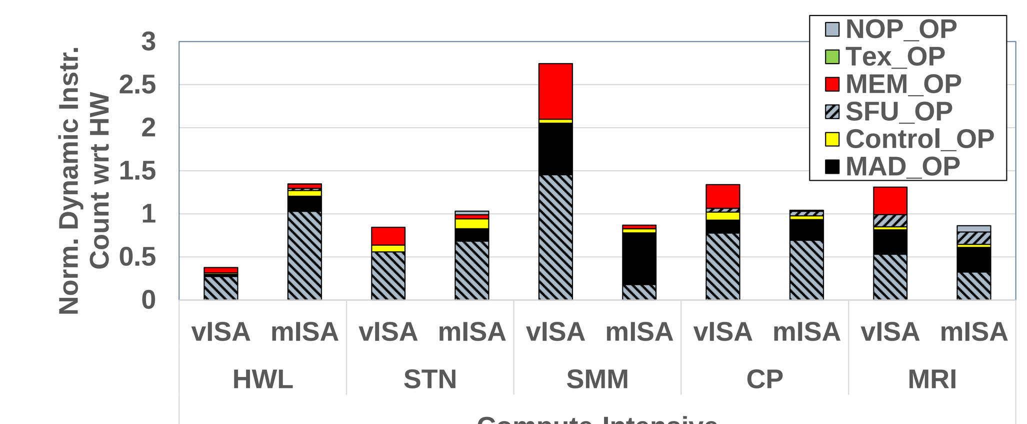


- High Correlation, low error.
- DRAM reads for these apps have high correlation with low error.
- Improving DRAM modeling can decrease the error further.



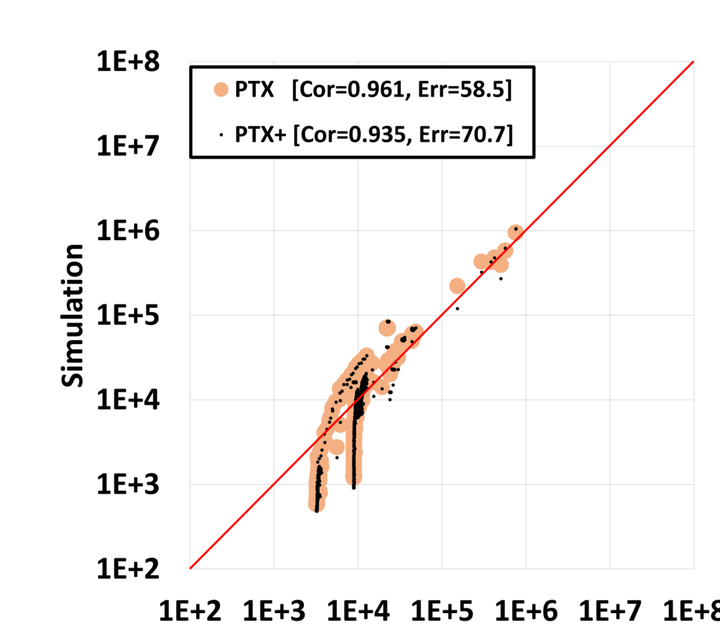
Compute Intensive





- High correlation and low error with mISA.
- mISA (even if old) is better than vISA.
- mISA has high correlation wrt dynamic inst breakdown.
- Supporting new mISA can improve correlation further.

Compute Balanced



- Better results with vISA in certain apps.
- vISA is able to hide the deficiencies of the old mISA.
- Z5 kernel in SRAD1 uses one instruction in vISA for integer division, which gets expanded into 33 instructions in the old mISA