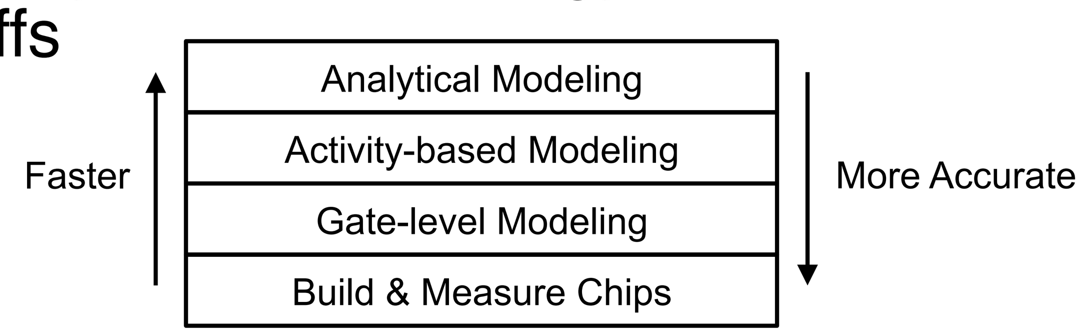
Accelerating Energy Modeling with FPGAs

Yunsup Lee, Donggyu Kim, Hokeun Kim, Michael Soliterman, Ian Juch,
Andrew Gearhart, Adam Izraelevitz, Brian Zimmer, Jonathan Bachrach, Krste Asanovic

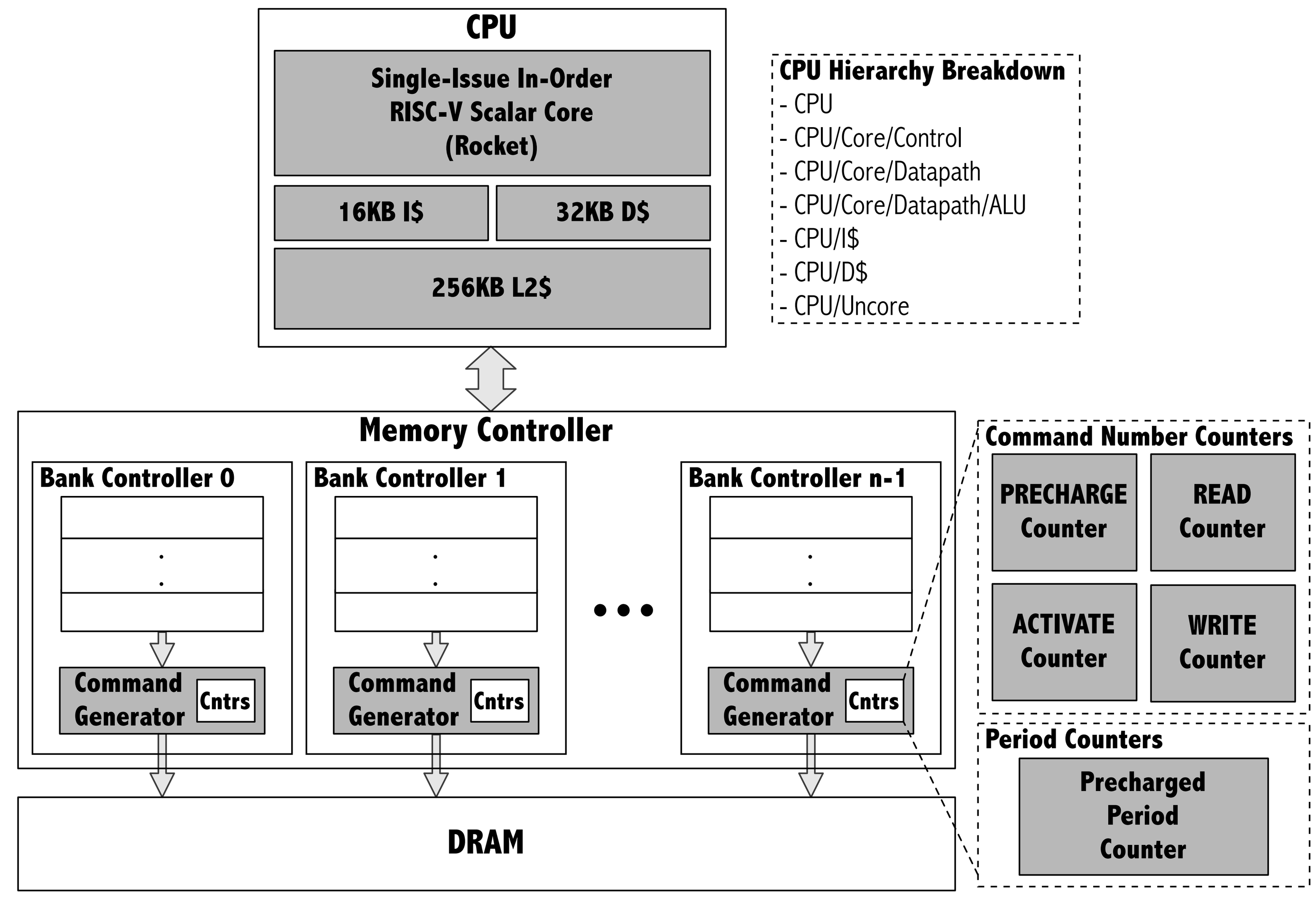
SoC Energy Modeling, Can we have best of both worlds?

- Many ways to model energy with different tradeoffs



- Can we have best of both?
  - As accurate as if we were building and measuring chips
  - As fast and productive as generating and evaluating an analytical model
  - Also, can we get cycle-by-cycle power readings?

Target Machine



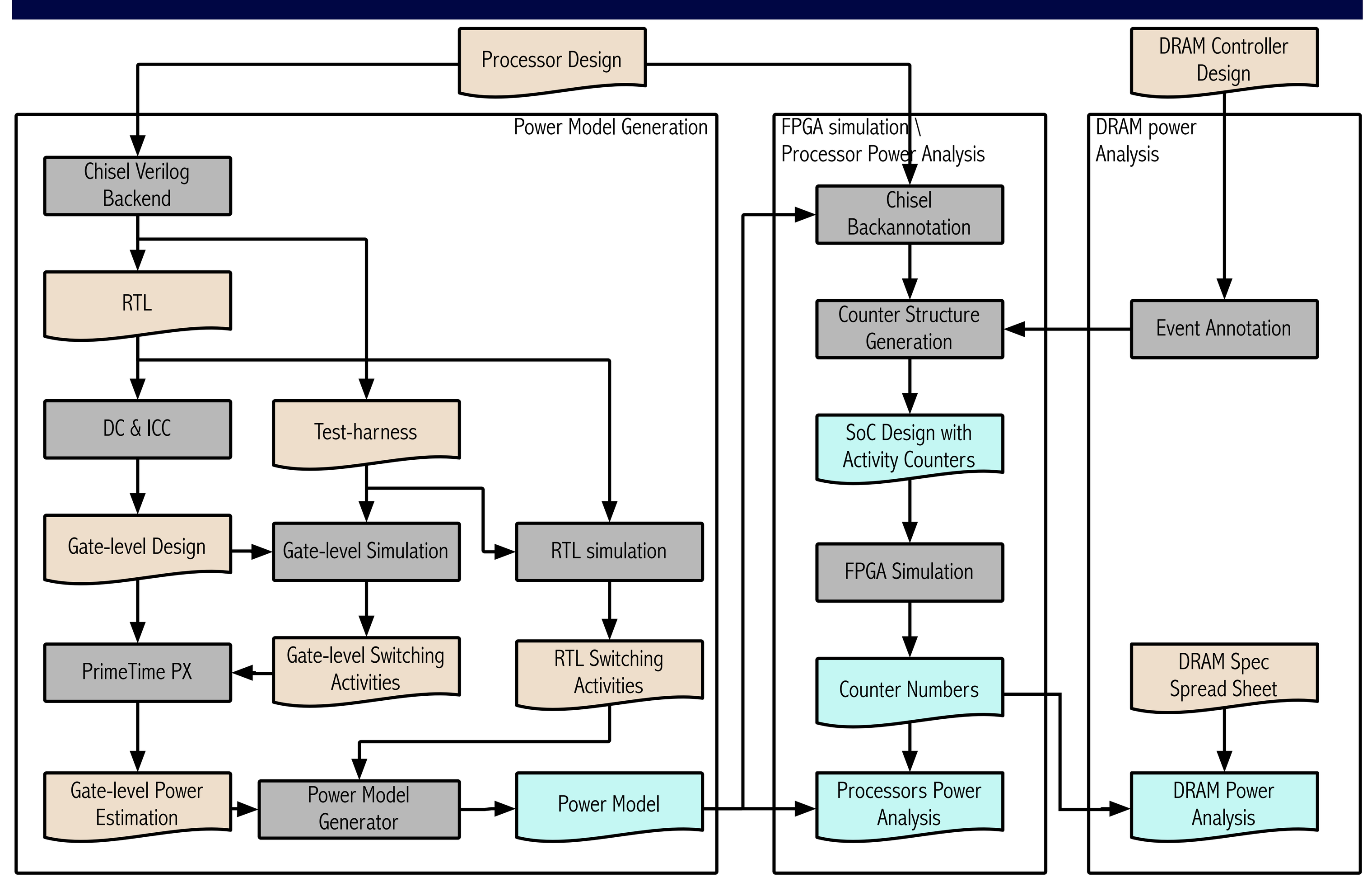
Processor Power Estimation

% Error	dgemm	dhrystone	median	multiply	qsort	spmv	towers	vvadd	average
power(mW)	42.2	46.7	36.9	36.2	37.2	35.2	39.8	36.2	38.8
38.8(mW)	8.1%	16.9%	5.0%	7.2%	4.2%	10.3%	2.5%	7.3%	7.7%
dgemm	0.0%	13.8%	4.0%	4.3%	2.8%	3.4%	6.9%	4.9%	5.0%
dhrystone	11.0%	0.0%	27.3%	20.3%	24.2%	35.8%	20.9%	33.0%	21.5%
median	11.5%	15.3%	0.0%	0.6%	2.1%	5.3%	7.9%	0.4%	5.4%
multiply	21.9%	26.6%	8.4%	0.0%	7.2%	2.5%	17.7%	3.9%	11.0%
qsort	17.7%	17.9%	3.8%	0.0%	0.0%	1.9%	12.5%	1.0%	6.9%
spmv	7.8%	18.2%	7.4%	2.8%	5.4%	0.0%	11.5%	9.8%	7.9%
towers	1.3%	4.1%	2.7%	2.7%	1.8%	16.8%	0.0%	5.9%	4.4%
vvadd	16.0%	19.4%	1.4%	0.0%	5.7%	2.7%	10.3%	0.0%	6.9%
average	10.9%	14.4%	6.9%	3.8%	6.1%	8.5%	11.0%	7.3%	

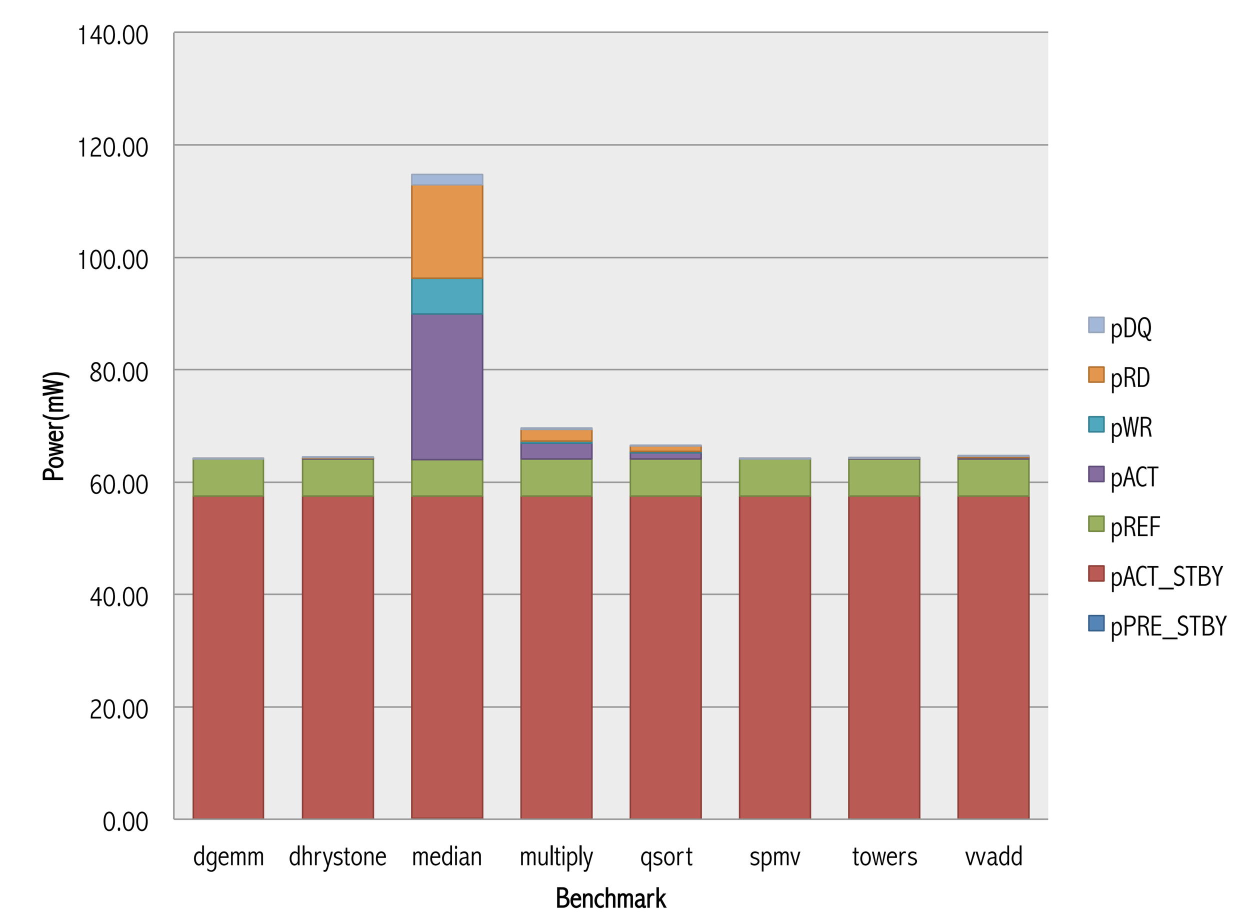
FPS: Fast Power Simulator

- Automatically generates a power model by performing a linear regression of gate-level power on RTL-level switching activity for the processor.
  - Accurate, since gate-level power strongly correlates with power measured on chip
  - Productive, since linear regression is done automatically without any designer knowledge, and is fully integrated with Chisel
  - Fast, since we can collect RTL-level switching activity on a design mapped to an FPGA
- Instantiates counters in the memory controller to track events for DRAM spreadsheet power modeling.

FPS Workflow



DRAM Power Estimation



Linear Power Model

- For given n signals, energy consumption at time t is expressed as follows:

$$e_t = c_0 + c_1 a_t^1 + c_2 a_t^2 + c_3 a_t^3 + \cdots + c_n a_t^n$$

  where c_i and a_t^i are signal i's dynamic power factor and activity factor, respectively.
- Assume that activity factors are identically independent
- To obtain dynamic factors $c = [c_i]$, solve the normal equation by the LMS algorithm:

$$c = \left(A^T A\right)^{-1} A^T e$$

  where $e = [e_t]$ is a power vector, and $A = [a_t^i]$ is a transition matrix
- To pick significant signals for power estimation, the cost function is assigned for each signal
- The LMS algorithm with cost functions:
- Set cost functions zero
- For each iteration k:
  1) Update the dynamic factors: $c^{(k+1)} = c^{(k)} + \rho\left(e_t - \left(c^{(k)}\right)^T a_t\right)$
  2) Update the cost functions

Processor Power Modeling Results

- Total of 18k signals in 185 modules
- Filtered out signals with less than 10% activity
  - ~95% signals were thrown away, ~1k remaining
- 7 hierarchical models were generated per training benchmark
- An example model

Hierarchy	# of Signals	Power (mW)	%	Example Important Signal
Control	254	0.5	1.3	Decoded rs2 Valid Signal
Datapath	244	5.1	12.8	Address of rs1
Datapath/ALU	110	0.4	1.0	Shift Amount
I$	82	6.6	16.5	Output of Data RAM
D$	356	10.2	25.6	Address, Tag, Data Registers
Uncore	43	10.1	25.3	Coherence data port
Everthing else	151	7.0	17.5	Current PC

Conclusions

- There exists a tradeoff between accuracy and latency (speed) in different energy modeling techniques
  - One should think carefully where/when to use which modeling techniques
  - We believe simple analytics models will suffice for modeling memory hierarchies, system interconnect
- FPS can automatically generate an analytical energy model, which is fast and has gate-level energy model accuracy
  - We believe FPS-style models should be used for modeling specialized cores (with accelerators)

Research funded by DARPA Award Number HR0011-12-2-0016. Approved for public release; distribution is unlimited. The content of this paper does not necessarily reflect the position or the policy of the US government and no official endorsement should be inferred.