Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs

Yannan Nellie Wu[1] , Joel S. Emer[1,2] , Vivienne Sze[1]

[1] MIT [2] NVIDIA

Accelergy Overview

- An architecture-level energy estimator
- Flexibly characterizes various basic building blocks of different technologies
- Succinctly models diverse and complicated designs
- Improves estimation accuracy via fine-grained classification of operations
- Achieves 95% accuracy in evaluating a deep neural network (DNN) accelerator – Eyeriss [ISSCC 2016]

Energy Consumption Concerns

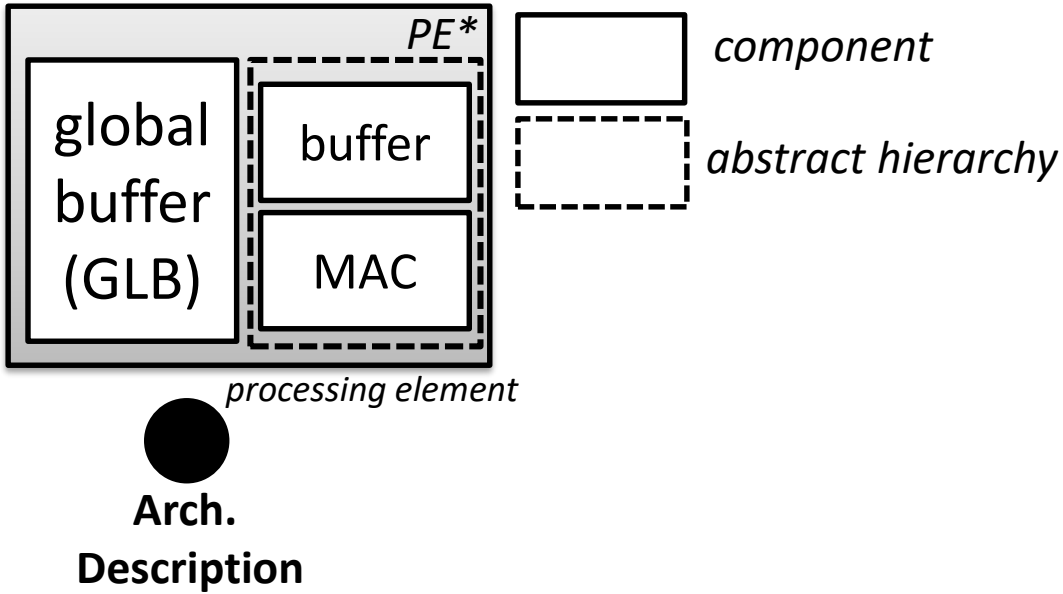
Data and computation-intensive applications are power hungry

Object Detection → **DNN Accelerator**

Database Processing → **Database Accelerator**

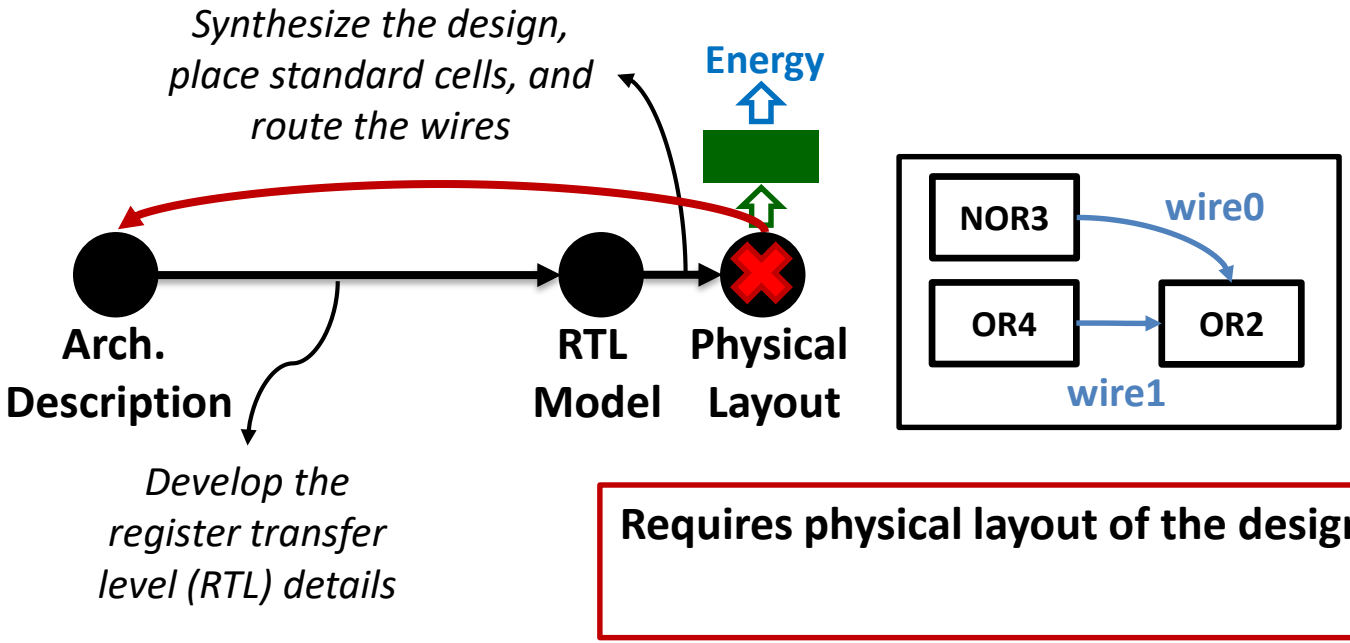
We must quickly evaluate energy efficiency of arbitrary potential designs in the large design space

Energy Estimation and Design Exploration



Energy Estimation and Design Exploration

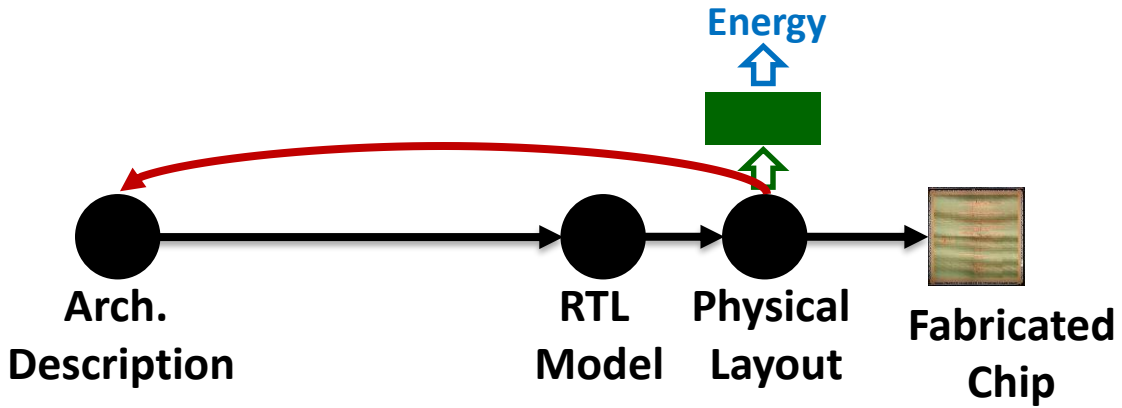
- **Physical-Level Energy Estimator (Synopsys Prime Power, Cadence Joules)**



Requires physical layout of the design

Energy Estimation and Design Exploration

- **Physical-Level Energy Estimator** (Synopsys Prime Power, Cadence Joules)



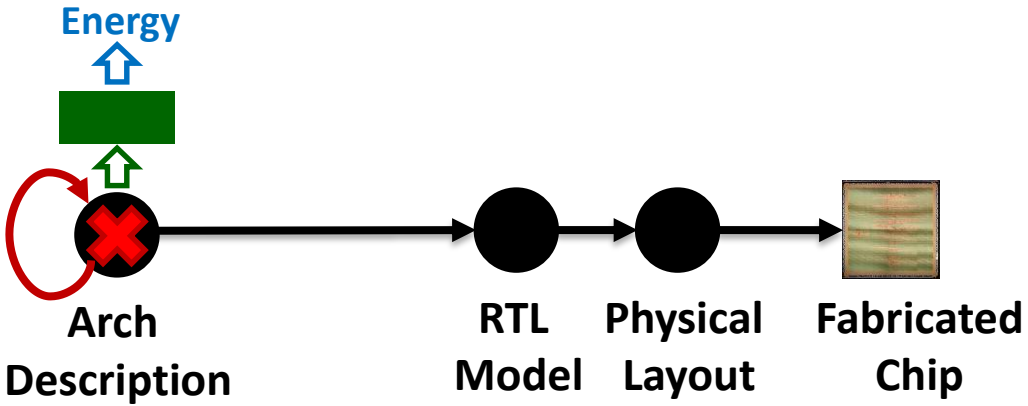
Requires physical layout of the design
Slow design space exploration

Accelergy Overview

- **An architecture-level energy estimator**
- Flexibly characterizes various basic building blocks in the design
- Succinctly models diverse and complicated designs
- Improves estimation accuracy via fine-grained classification of operations
- Achieves 95% accuracy in evaluating a deep neural network (DNN) accelerator – Eyeriss [ISSCC 2016]

Energy Estimation and Design Exploration

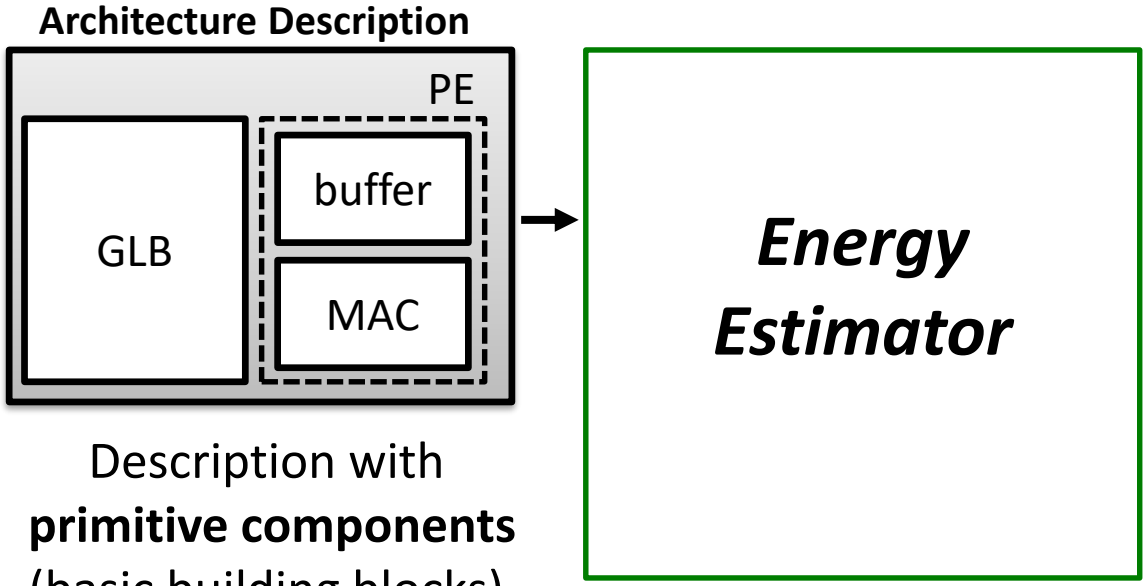
- **Architecture-Level Energy Estimators**



Only requires architecture-level design
Fast design space exploration

Existing Architecture-Level Energy Estimators

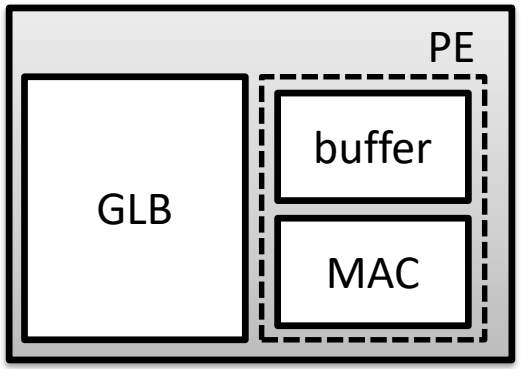
- **Design-Specific Accelerator Estimators: Aladdin**[ISCA2014], **fixed-cost**[Asilomar2017]



Existing Architecture-Level Energy Estimators

- **Design-Specific Accelerator Estimators: Aladdin**[ISCA2014], **fixed-cost**[Asilomar2017]

Architecture Description



Energy Estimator

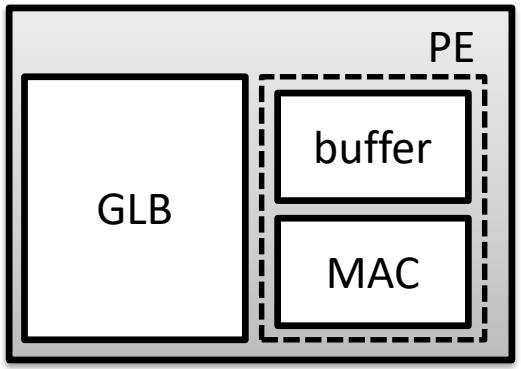
Energy Reference Table (ERT)

Comp.	Action	Energy
GLB	access()	100pJ
buffer	access()	10pJ
MAC	compute()	5pJ

Existing Architecture-Level Energy Estimators

- **Design-Specific Accelerator Estimators: Aladdin**[ISCA2014], **fixed-cost**[Asilomar2017]

Architecture Description



Comp.	Action	Counts
GLB	access()	10
buffer	access()	800
MAC	compute()	400

Action Counts

Energy Estimator

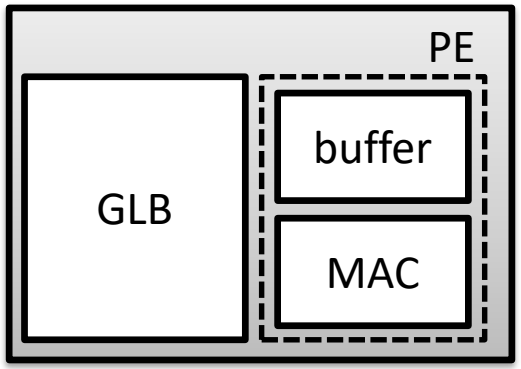
Energy Reference Table (ERT)

Comp.	Action	Energy
GLB	access()	100pJ
buffer	access()	10pJ
MAC	compute()	5pJ

Existing Architecture-Level Energy Estimators

- **Design-Specific Accelerator Estimators: Aladdin**[ISCA2014], **fixed-cost**[Asilomar2017]

Architecture Description



Comp.	Action	Counts
GLB	access()	10
buffer	access()	800
MAC	compute()	400

Action Counts

Energy Estimator

Energy Reference Table (ERT)

Comp.	Action	Energy
GLB	access()	100pJ
buffer	access()	10pJ
MAC	compute()	5pJ

Energy Calculator

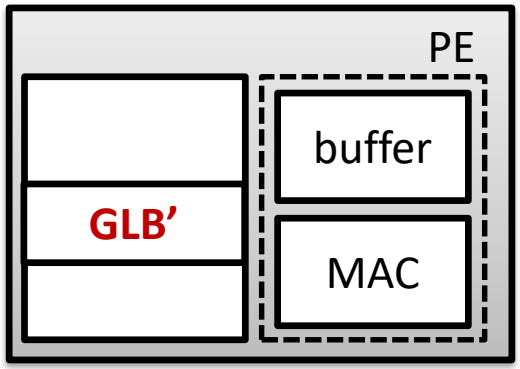
Energy Estimations

Name	Energy
GLB	1000pJ
buffer	8000pJ
MAC	2000pJ

Existing Architecture-Level Energy Estimators

- **Design-Specific Accelerator Estimators: Aladdin**[ISCA2014], **fixed-cost**[Asilomar2017]

Architecture Description



Comp.	Action	Counts
GLB'	access()	10
buffer	access()	800
MAC	compute()	400

Action Counts

Energy Estimator

Energy Reference Table (ERT)

Comp.	Action	Energy
GLB	access()	100pJ
buffer	access()	10pJ
MAC	compute()	5pJ

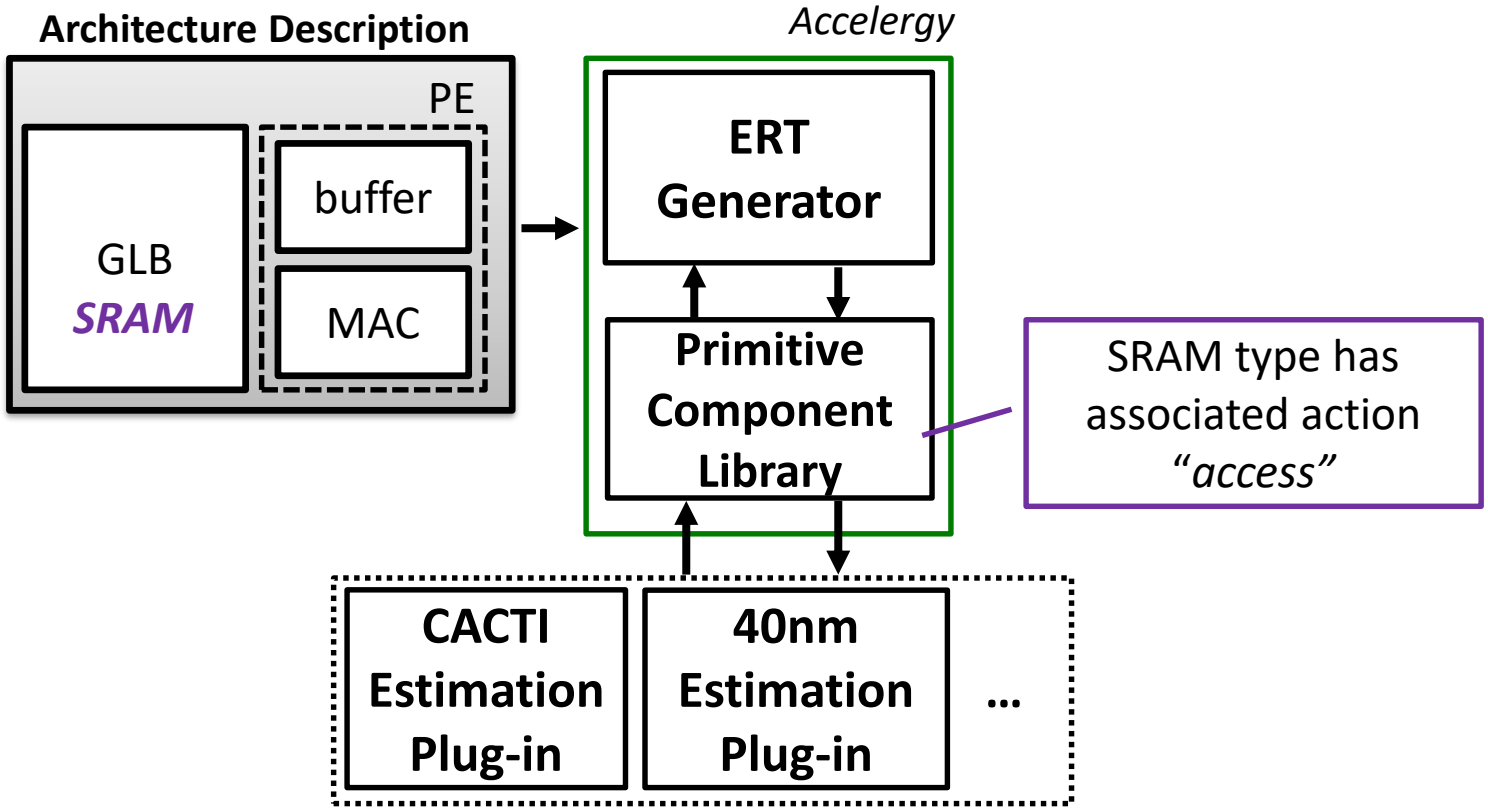
Energy Calculator

Not generalizable to other designs

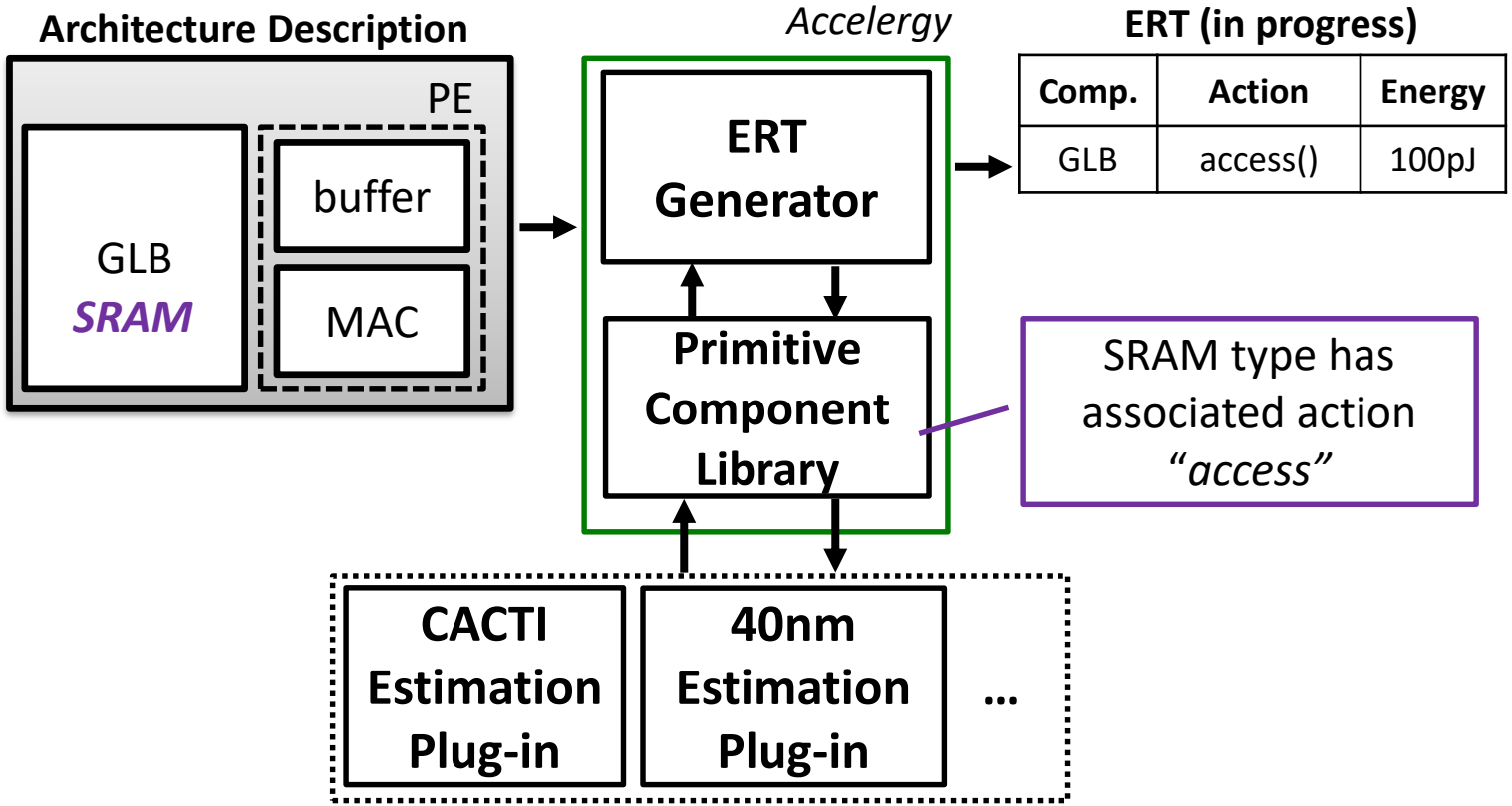
Accelergy Overview

- An architecture-level energy estimator
- **Flexibly characterizes various primitive components of different technologies**
- Succinctly models diverse and complicated designs
- Improves estimation accuracy via fine-grained classification of operations
- Achieves 95% accuracy in evaluating a deep neural network (DNN) accelerator – Eyeriss [ISSCC 2016]

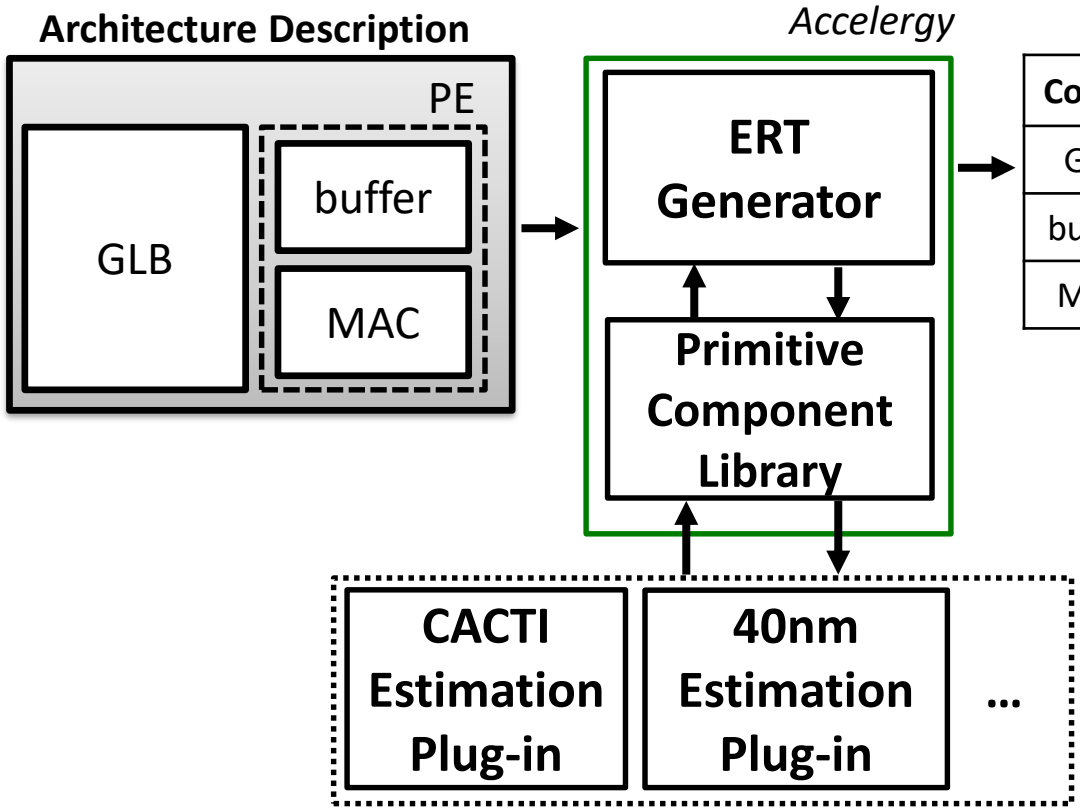
Accelergy: Flexibly Model Various Primitive Components



Accelergy: Flexibly Model Various Primitive Components



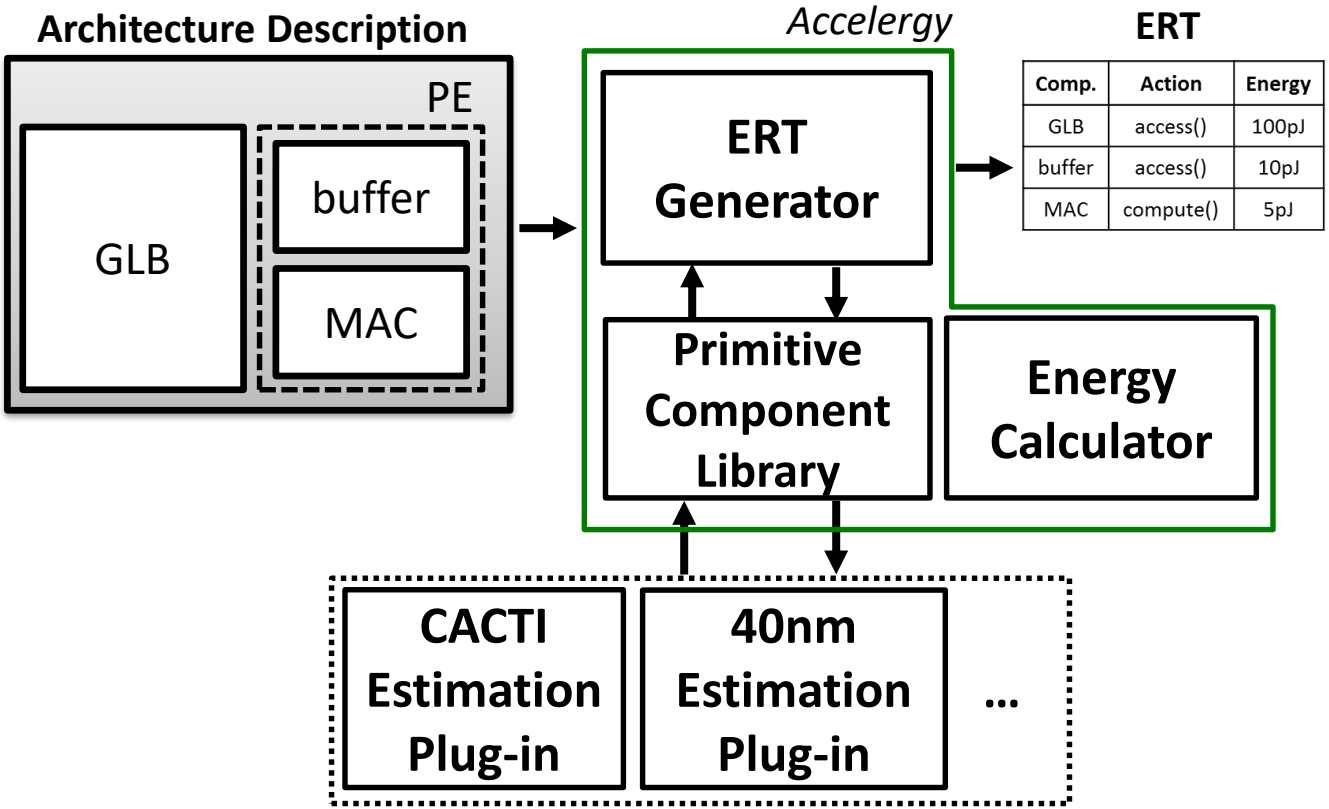
Accelergy: Flexibly Model Various Primitive Components



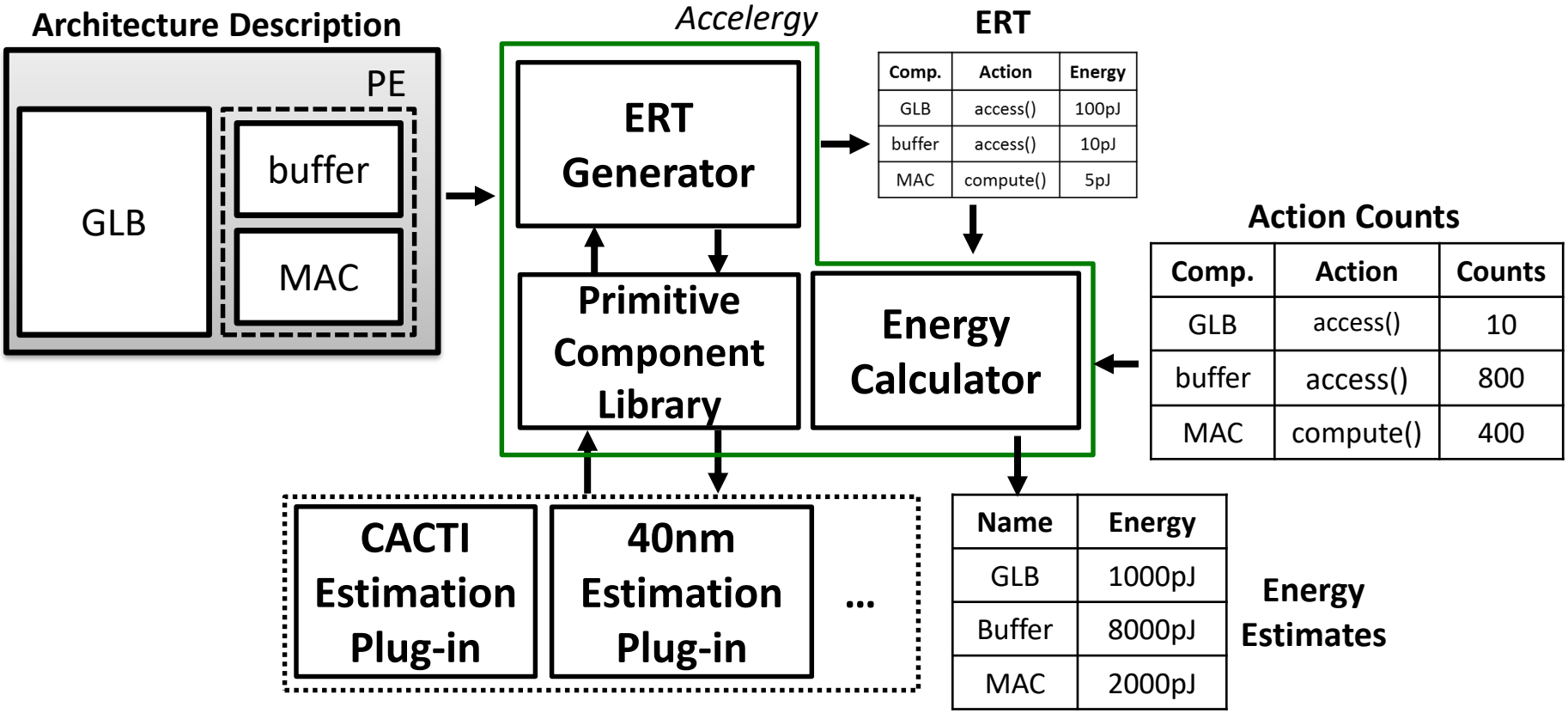
ERT

Comp.	Action	Energy
GLB	access()	100pJ
buffer	access()	10pJ
MAC	compute()	5pJ

Accelergy: Flexibly Model Various Primitive Components



Accelergy: Flexibly Model Various Primitive Components



Accelergy: Flexibly Model Various Primitive Components

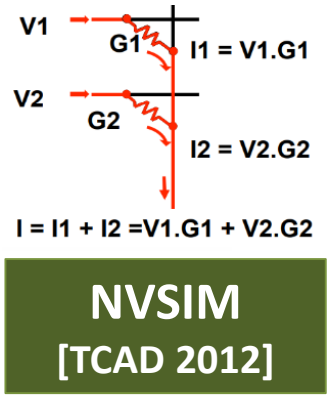
Use energy estimation plug-ins to characterize primitive components

CACTI Estimation Plug-in

40nm Estimation Plug-in

Traditional open-source plug-ins*

Proprietary plug-ins

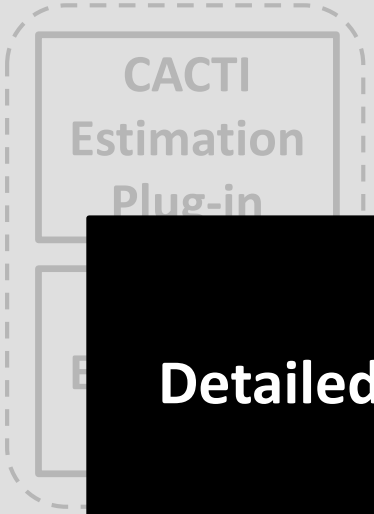


Emerging technology plug-ins

**available at* http://accelergy.mit.edu

Accelergy: Flexibly Model Various Primitive Components

Use energy estimation plug-ins to characterize primitive components



Traditional open-source plug-ins*

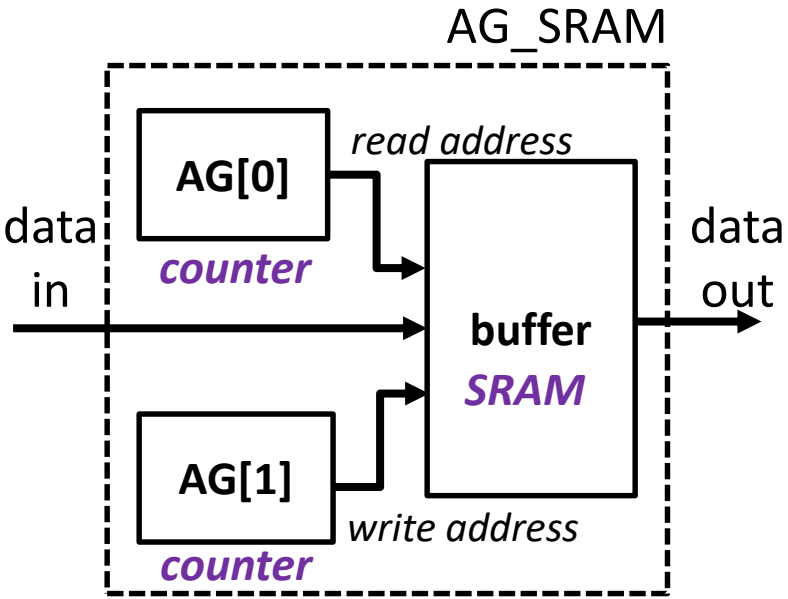
plug-ins

Detailed plug-in interface in open-source repo

**available at* http://accelergy.mit.edu

Modeling Complicated Designs

- **Practical architecture designs involve much more details**
  - **Example: storage units with local address generators (AGs)**

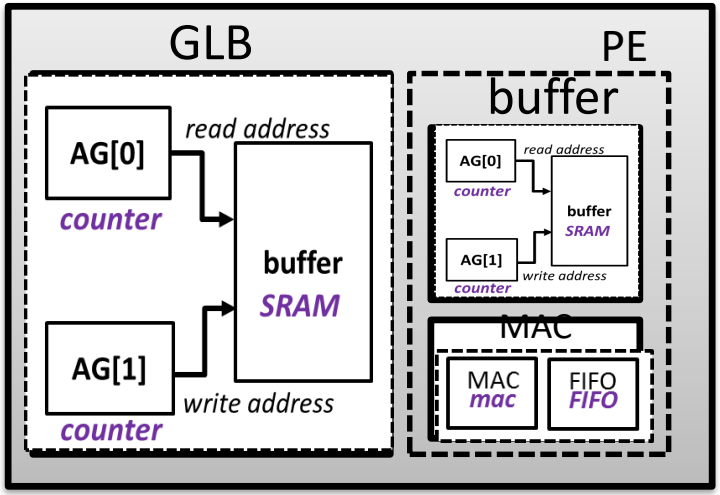


- AG_SRAM is an abstract hierarchy
- Buffer is of SRAM type
- AGs is of counter type

Modeling Complicated Designs

- **Practical architecture designs involve much more details**
  - **Example: storage units with local address generators (AG)**

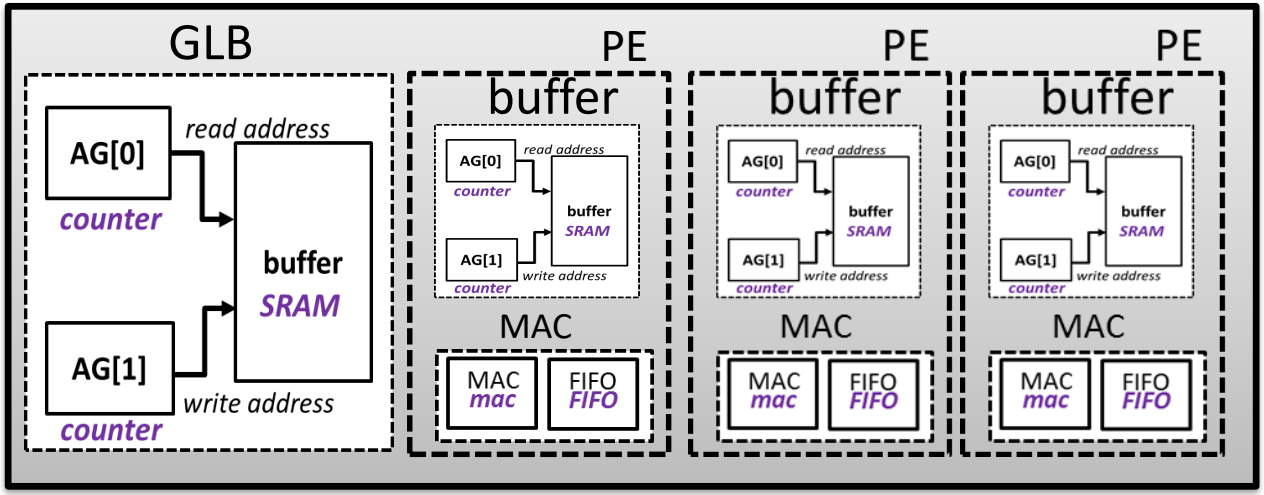
Let's construct a more practical design!



Modeling Complicated Designs

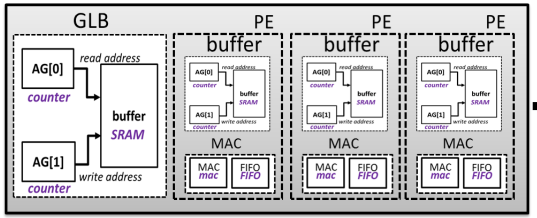
- **Practical architecture designs involve much more details**
  - **Example: storage units with local address generators (AG)**

Let's construct a more practical design!



Modeling Complicated Designs

Architecture Description



- Architecture description is tedious
- Hard to make modifications

Accelergy

ERT Generator

Primitive Component Library

Energy Calculator

CACTI Estimation Plug-in

40nm Estimation Plug-in

…

- Action counts are even more tedious
- Small modification requires new action counts

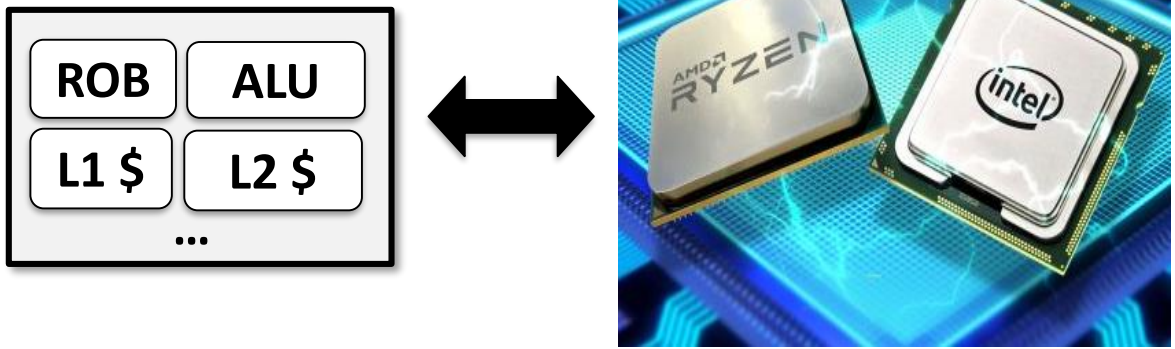
Name	Action	Counts
PE[0]. AG[0]	count	50
PE[0]. AG[1]	count	50
…		

Action Counts

Existing Work - Modeling Complicated Designs

- **Existing work that aims to succinctly model complicated architectures**
  - **Wattch**[ISCA2000], **McPAT**[MICRO2009]

CPU-Centric Architecture Model



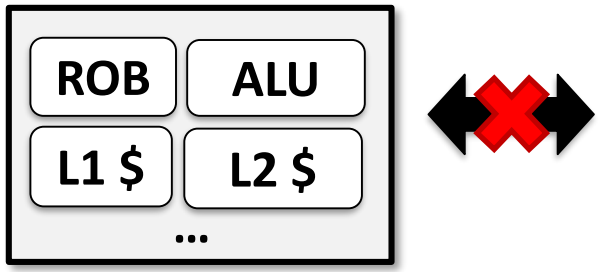
Use a fixed set of compound components to represent the architecture

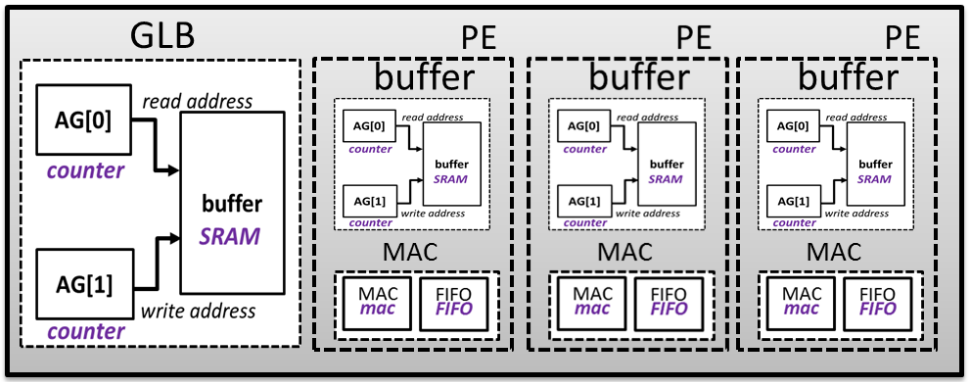
Components that can be decomposed into lower level components

Existing Work - Modeling Complicated Designs

- **Existing work that aims to succinctly model complicated architectures**
  - **Wattch**[ISCA2000], **McPAT**[MICRO2009]

CPU-Centric Architecture Model



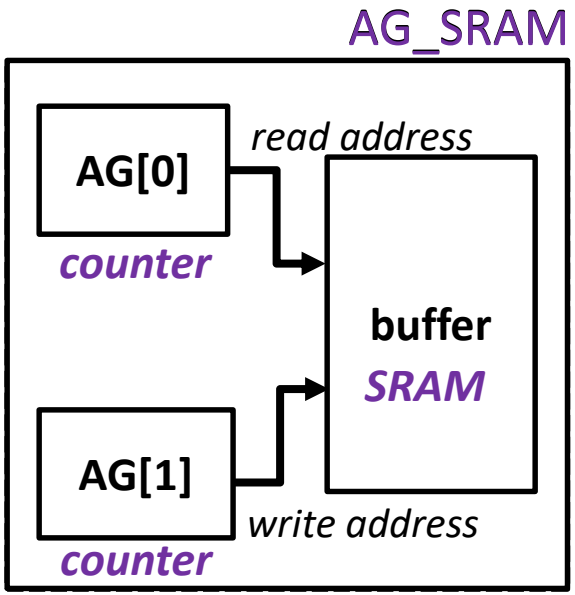


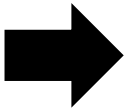
The fixed set of compound components is not sufficient to describe arbitrary accelerator designs

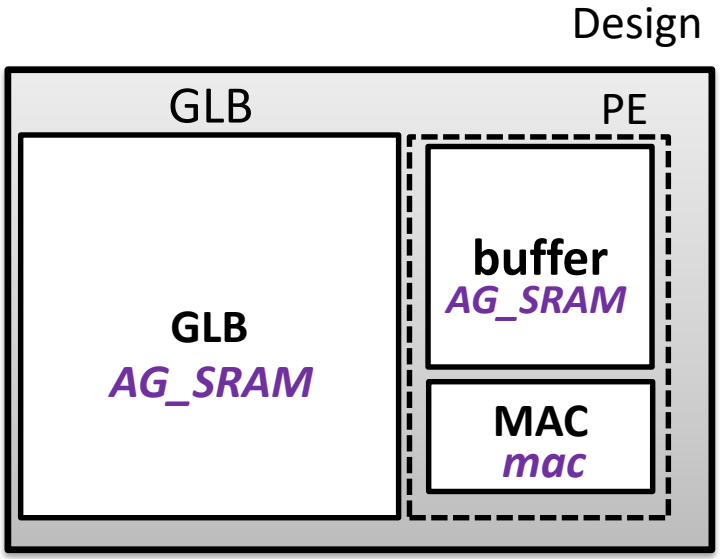
Accelergy Overview

- An architecture-level energy estimator
- Flexibly characterizes various primitive components of different technologies
- **Succinctly models diverse and complicated designs**
- Improves estimation accuracy via fine-grained classification of operations
- Achieves 95% accuracy in evaluating a deep neural network (DNN) accelerator – Eyeriss [ISSCC 2016]

Accelergy: Succinctly Model Arbitrary Architecture





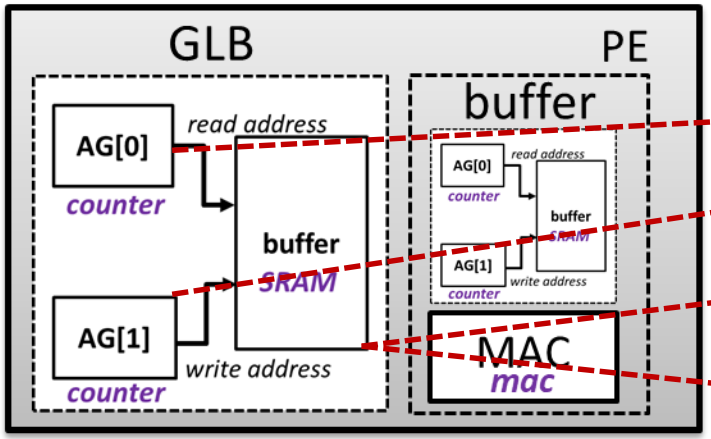


AG_SRAM is an user-defined compound component

Architecture described with compound components and primitive components

Accelergy: Succinctly Model Arbitrary Action Counts

Design

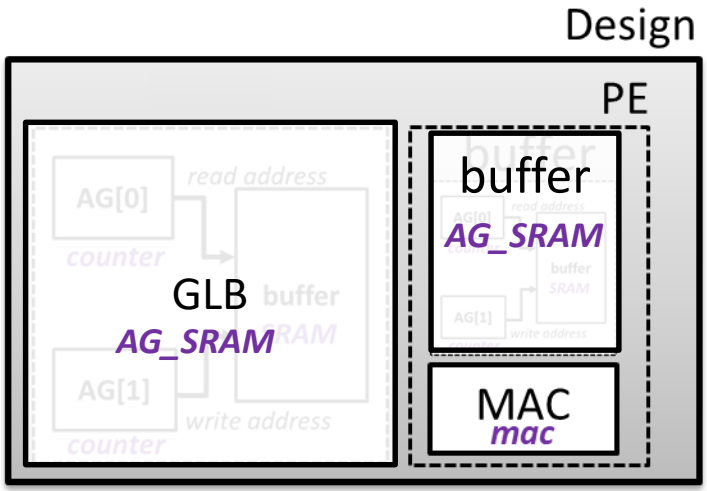


Action Counts

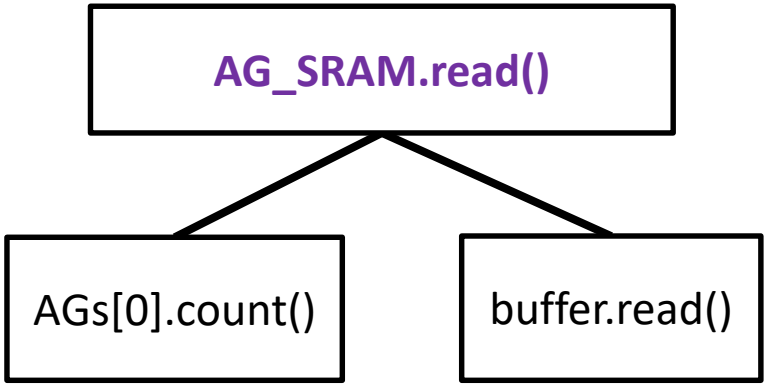
Name	Action	Counts
GLB.AG[0]	count()	50
GLB.AG[1]	count()	20
GLB.buffer	read()	50
GLB.buffer	write()	20
...		

Tedious action counts in terms of primitive component actions

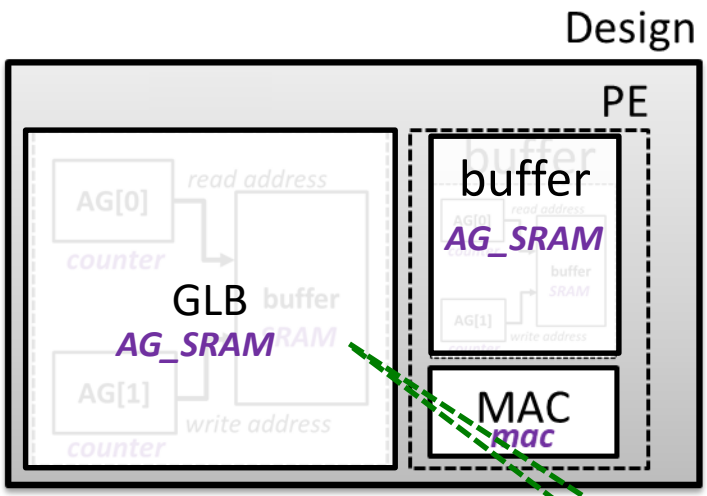
Accelergy: Succinctly Model Arbitrary Action Counts



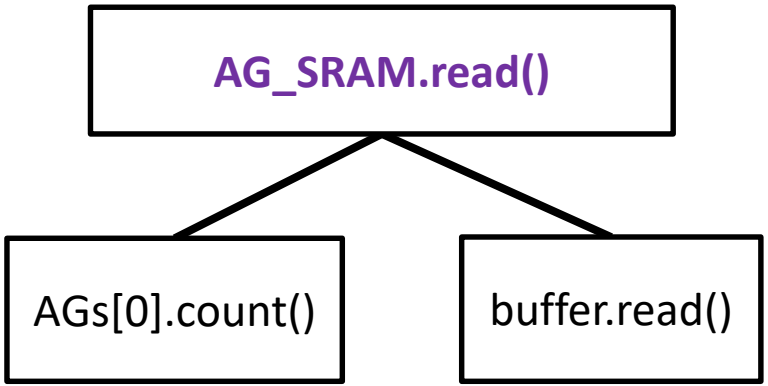
User-defined compound actions



Accelergy: Succinctly Model Arbitrary Action Counts



User-defined compound actions

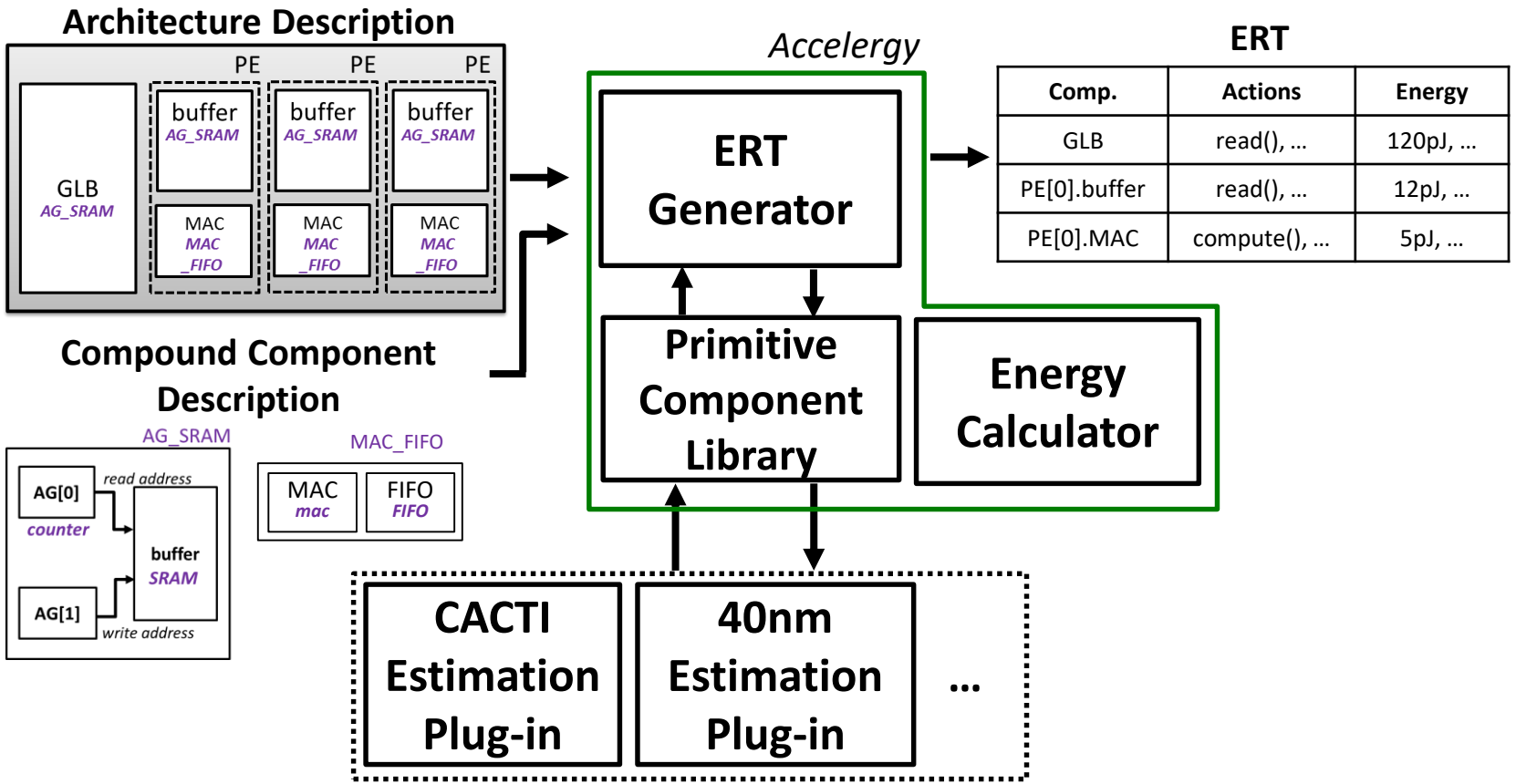


Action Counts

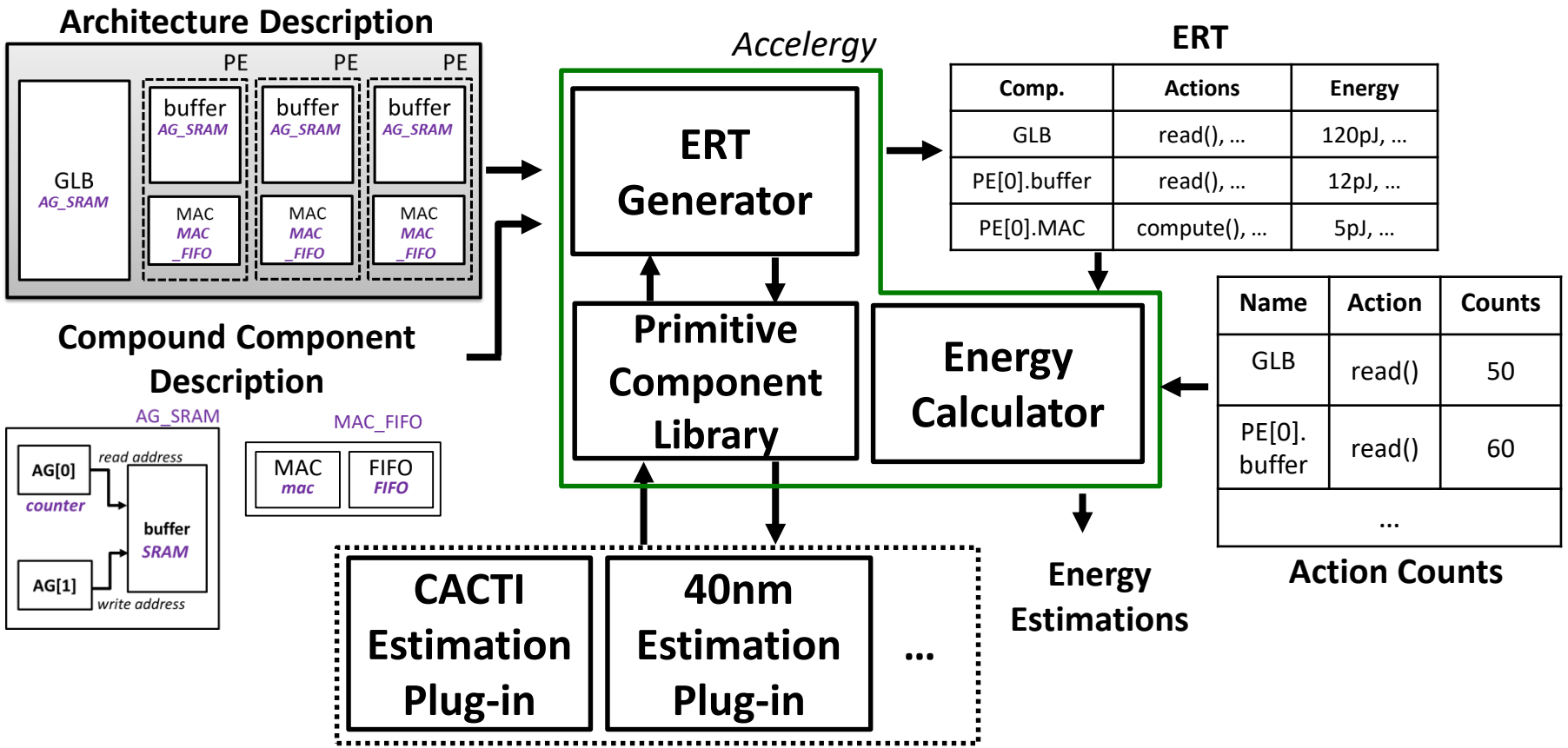
Name	Action	Counts
GLB	read()	50
GLB	write()	20

Succinct action counts with compound component actions

Accelergy: Succinctly Model Complex Designs



Accelergy: Succinctly Model Complex Designs



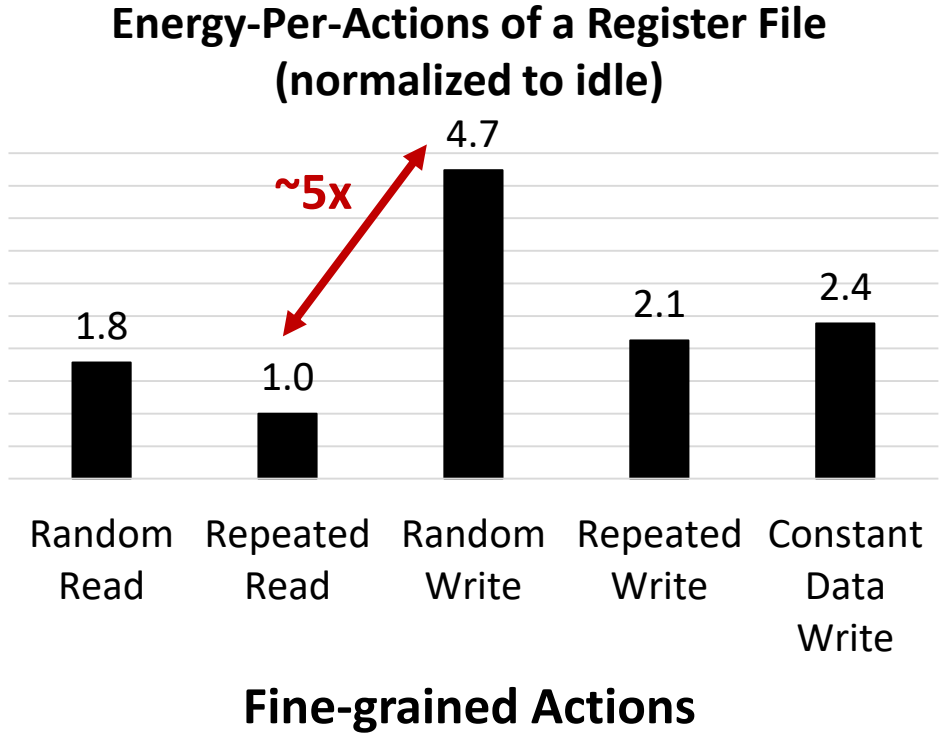
Additional Challenge: Inaccurate Modeling of Energy/Action

- Existing architecture-level energy estimators only model coarse action types

Component	Action	Energy
GLB	access()	100pJ
Buffer	access()	10pJ
ALU	compute()	5pJ

Coarse-grained Actions

Coarse-grained estimations reduce estimation accuracies



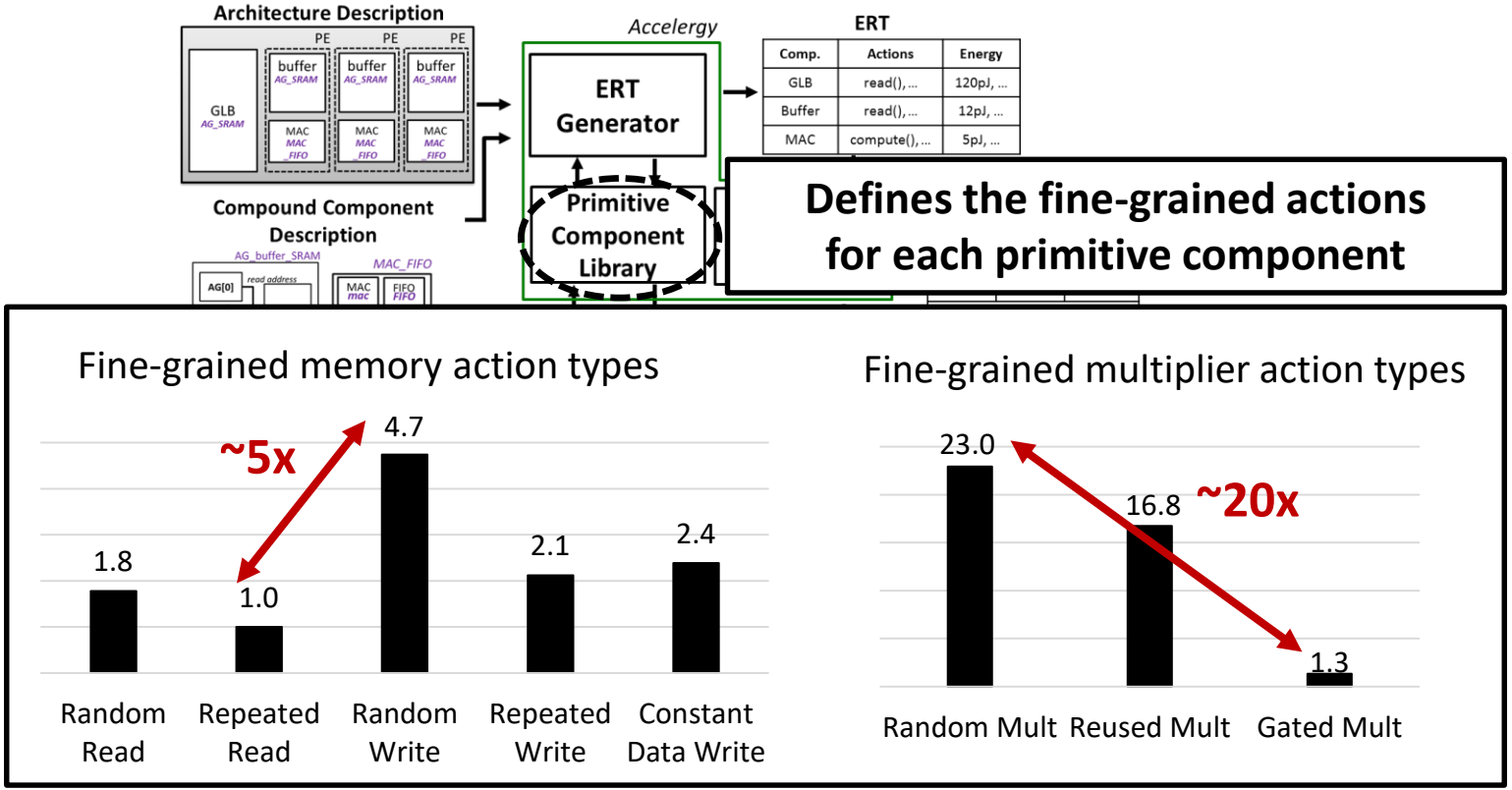
Fine-grained Actions

Accelergy Overview

- An architecture-level energy estimator
- Flexibly characterizes various primitive components of different technologies
- Succinctly models diverse and complicated designs
- **Improves estimation accuracy via fine-grained actions**
- Achieves 95% accuracy in evaluating a deep neural network (DNN) accelerator – Eyeriss [ISSCC 2016]

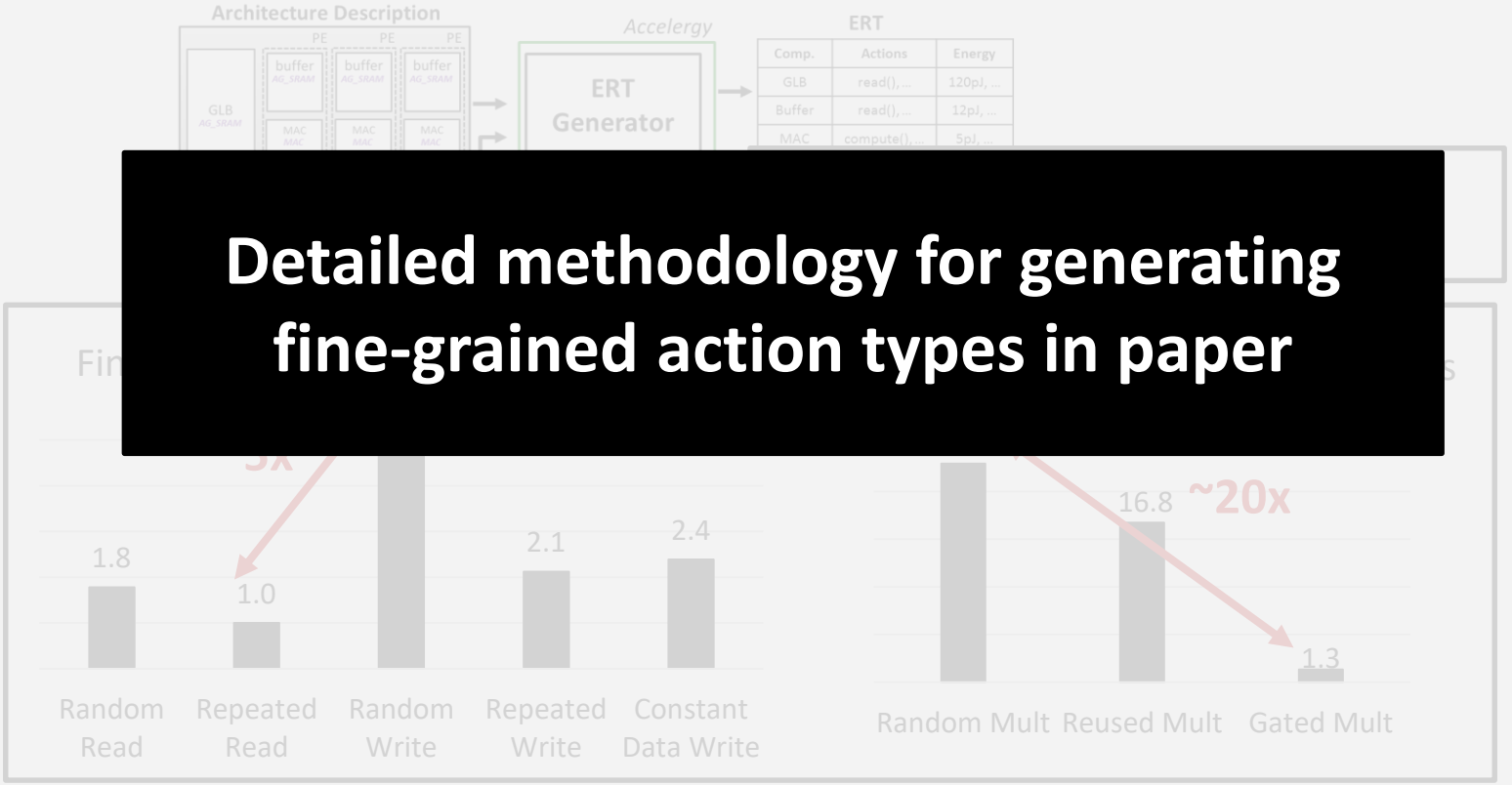
Accelergy: Fine-grained Action Classification

- **Accurate estimation with a primitive component library**



Accelergy: Fine-grained Action Classification

- Accurate estimation with a primitive component library*



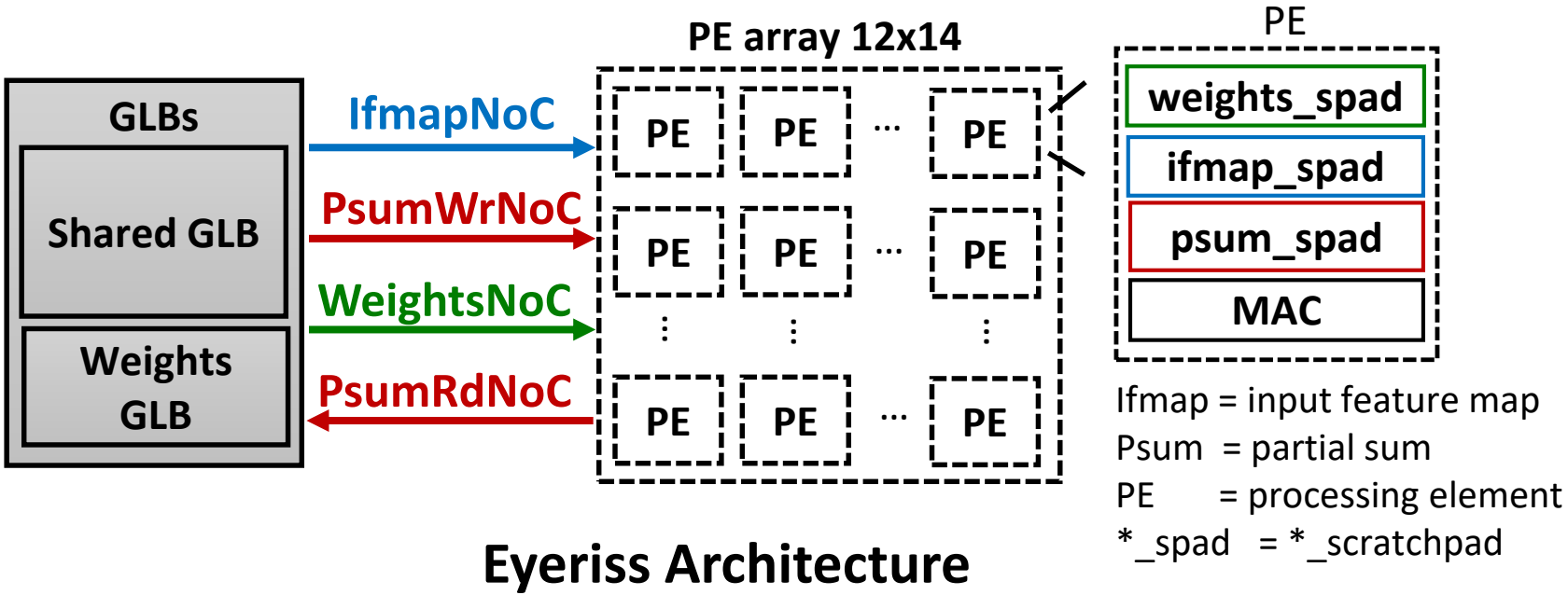
Detailed methodology for generating fine-grained action types in paper

Accelergy Overview

- An architecture-level energy estimator
- Flexibly characterizes various primitive components of different technologies
- Succinctly models diverse and complicated designs
- Improves estimation accuracy via fine-grained actions
- **Achieves 95% accuracy in evaluating a deep neural network (DNN) accelerator – Eyeriss [ISSCC 2016]**

Energy Evaluations on Eyeriss

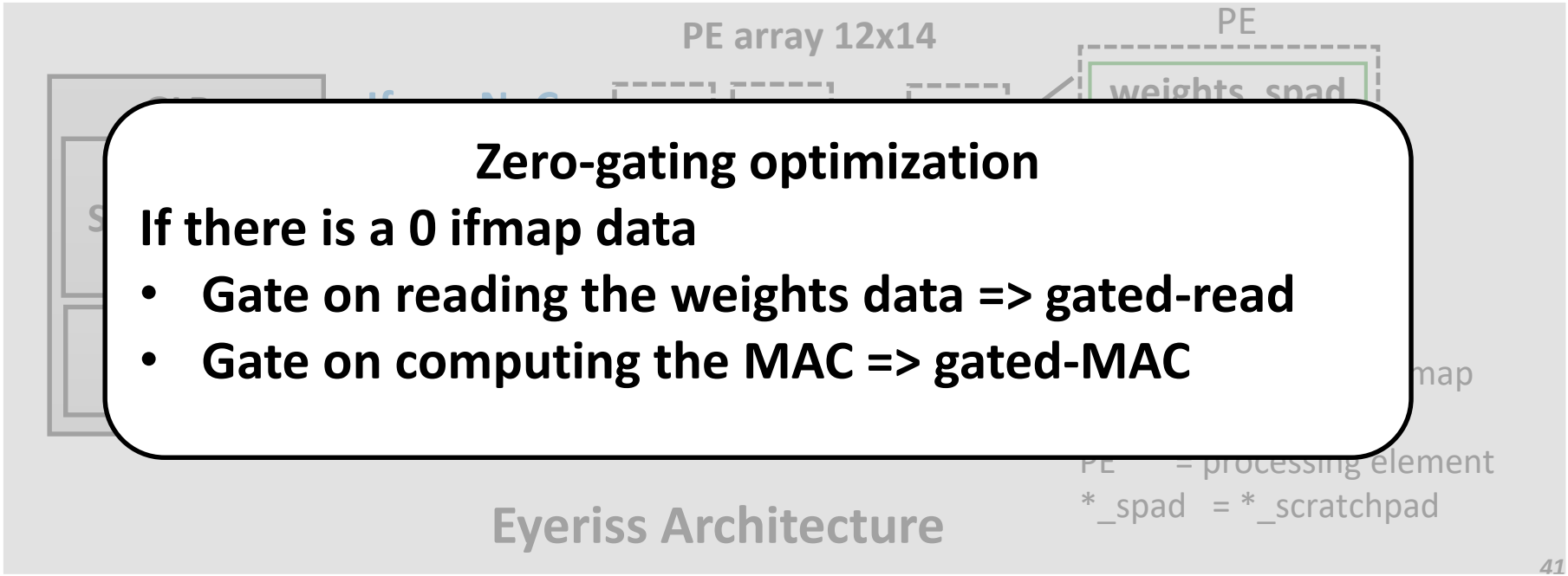
- **Experimental Setup:**
  - **Workload: Alexnet weights & ImageNet input feature maps**
  - **Ground Truth: Energy obtained from post-layout simulations**



Eyeriss Architecture

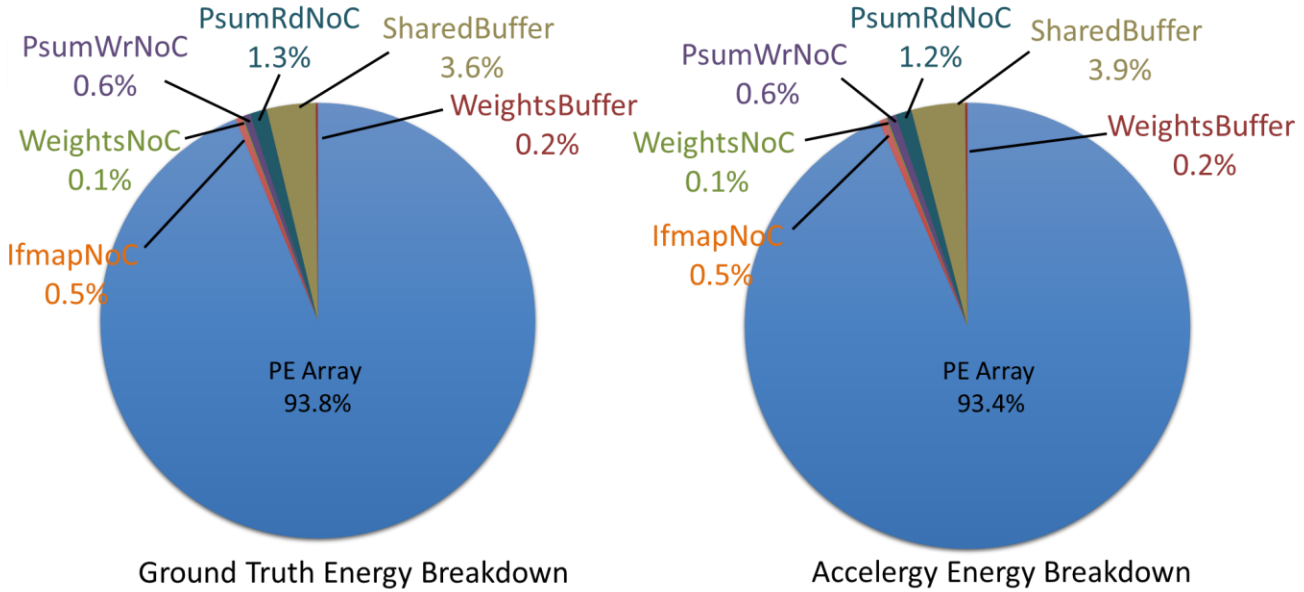
Energy Evaluations on Eyeriss

- **Experimental Setup:**
  - **Workload: Alexnet weights & ImageNet input feature maps**
  - **Ground Truth: Energy obtained from post-layout simulations**



Total Energy and Coarse Energy Breakdown

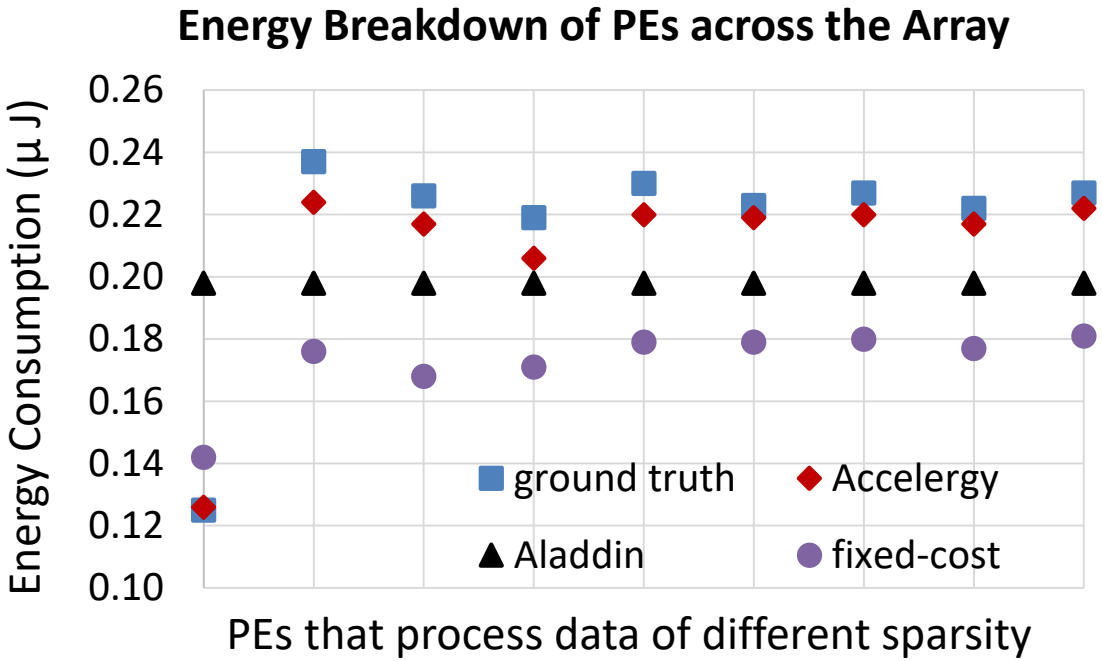
- **Total energy estimation is 95% accurate of the post-layout energy.**
- **Estimated relative breakdown of the important units in the design is <u>within 8%</u> of the post-layout energy.**



**Total energy might not add up to exact 100.0% due to rounding*

PE Array Energy Breakdown

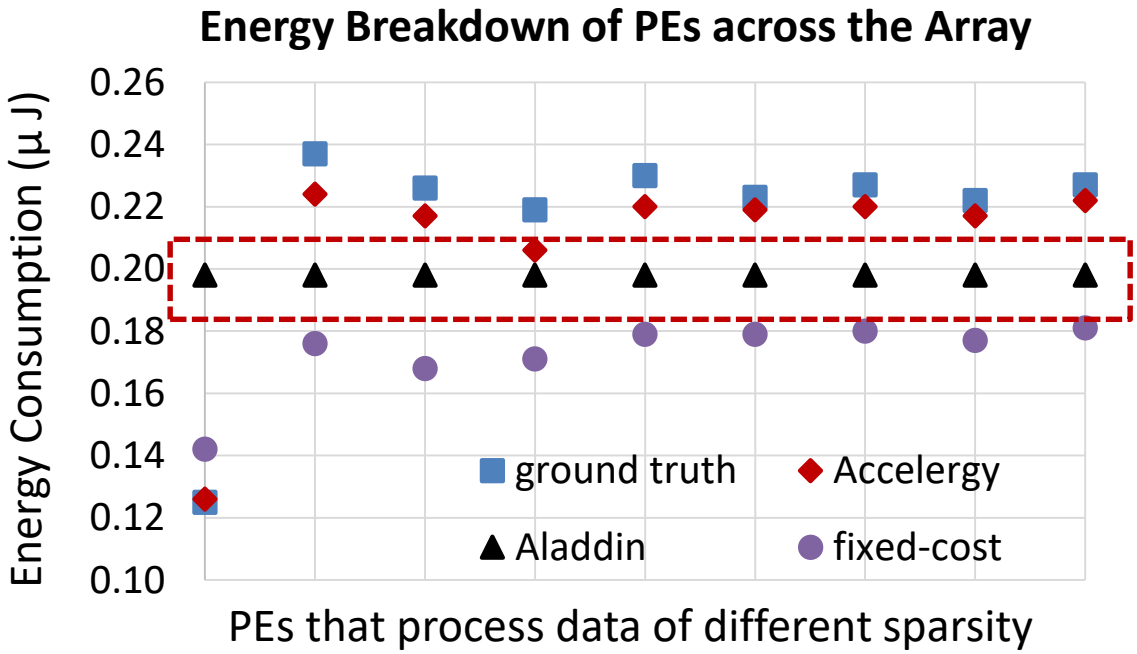
- **Comparisons with existing work: Aladdin and fixed-cost**



PE Array Energy Breakdown

- **Comparisons with existing work: Aladdin and fixed-cost**

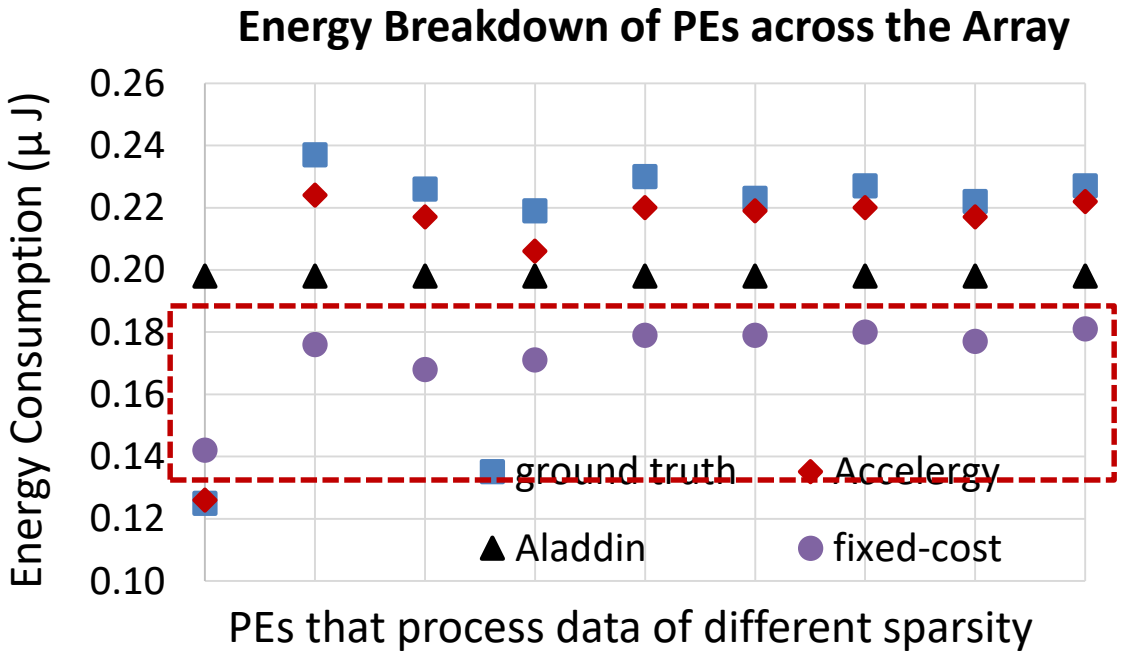
Not aware of the fine-grained actions related to zero-gating optimization



PE Array Energy Breakdown

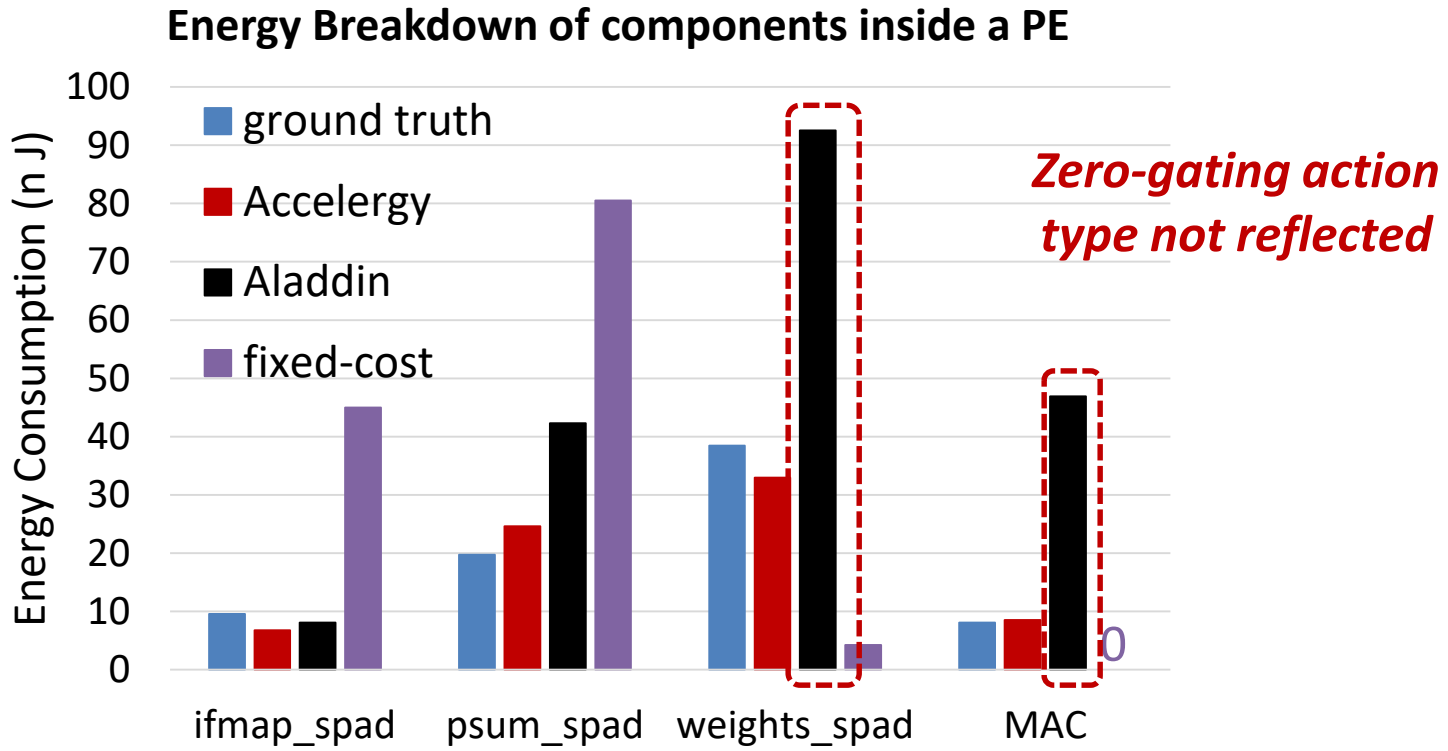
- **Comparisons with existing work: Aladdin and fixed-cost**

Inaccurate energy characterization of components



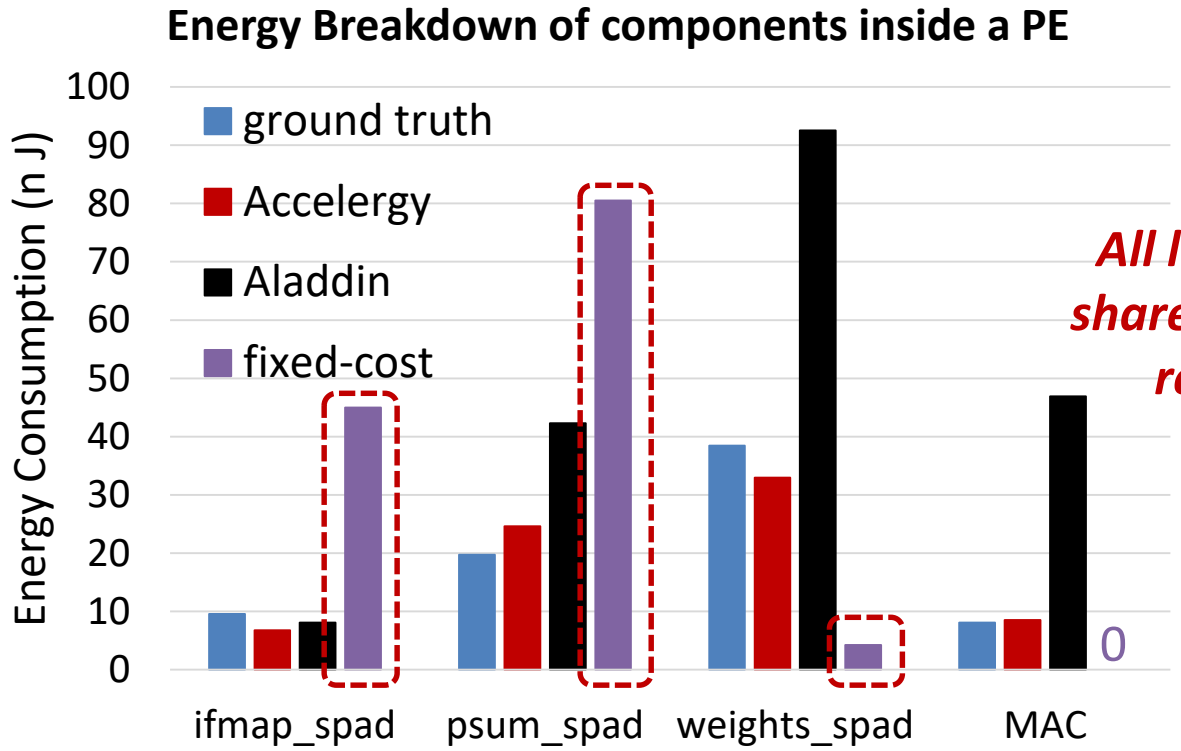
PE Energy Breakdown

- **Comparisons with existing work: Aladdin and fixed-cost**



PE Energy Breakdown

- **Comparisons with existing work: Aladdin and fixed-cost**



All local scratchpads share the same energy reference table

Conclusion

- Accelergy is an architecture-level energy estimator that
  - Accelerates accelerator design space exploration
  - Provides flexibility to
    - Describe a diverse range of accelerator designs
    - Support estimation of different technologies, e.g., CMOS, RRAM, optical
  - Achieves high accuracy energy estimations
    - 95% accurate for the Eyeriss accelerator
- Open-source code available at: http://accelergy.mit.edu

Acknowledgement: DARPA, Facebook, MIT Presidential Fellowship