Accelergy: An Architecture-Level Energy and Area Estimator for Accelerator Designs

Yannan Nellie Wu\textsuperscript{1}, Joel S. Emer\textsuperscript{1,2}, Vivienne Sze\textsuperscript{1}

\textsuperscript{1} MIT \hspace{1cm} \textsuperscript{2} NVIDIA
Domain-Specific Accelerators Improve Energy Efficiency

Data and computation-intensive applications are power hungry

- Object Detection
  - Deep Neural Network Accelerator
- Database Processing
  - Database Accelerator

We must quickly evaluate energy efficiency of arbitrary potential designs in the large design space.
From Architecture Blueprints to Physical Systems

- How many levels in the memory hierarchy?
- How large are the memories at each level?
- How many PEs are there?
- What are the X and Y dimensions of the PE array?
- ...

*processing element

Global Buffer (GLB)

Architecture Stage

Processing Element (PE) 0

PE 2

PE 3
From Architecture Blueprints to Physical Systems

Global Buffer (GLB)

*processing element

PE0

PE2

PE3

Architecture Stage

RTL Model

Physical Layout

Fabricated System

[ISSCC 2016]
Physical-Level Energy Estimation and Design Exploration

- How many levels in the memory hierarchy?
- How large are the memories at each level?
- How many PEs are there?
- What are the X and Y dimensions of the PE array?
- ...

Slow design space exploration
- Long simulations on gate-level components
- Long turn-around time for each potential design
Building systems with emerging technologies can be even more time-consuming, limiting the amount of design space.

Physical-Level Energy Estimation and Design Exploration

[Architecture Stage] → [Building System] → [Physical System]

Energy

Physical-Level Energy Estimator

Optical Computation [Nature Photonics 2017]

Non-volatile Memory Computation [NANOARCH 2017]
Architecture-Level Energy Estimation and Design Exploration

Fast design space exploration
- Short simulations on architecture-level components
- Short turn-around time for each potential design
Existing Architecture-Level Energy Estimators

- Architecture-level energy modeling for general purpose processors
  - Wattch[Brooks, ISCA2000], McPAT[Li, MICRO2009], GPUWattch[Leng, ISCA2013], PowerTrain[Lee, ISLPED2015]

**CPU/GPU-Centric Architecture Model**

Use a fixed set of **compound components** to represent the architecture

Components that can be decomposed into lower level components
Existing Architecture-Level Energy Estimators

- Architecture-level energy modeling for general purpose processors
  - Wattch [Brooks, ISCA2000], McPAT [Li, MICRO2009], GPUWattch [Leng, ISCA2013], PowerTrain [Lee, ISLPED2015]

CPU/GPU-Centric Architecture Model

The fixed set of compound components is not sufficient to describe various optimizations in the diverse accelerator design space.
Outline

• **Accelergy Infrastructure**
  – Performs architecture-level estimations to enable rapid design space exploration
  – Supports modeling of diverse architecture designs
  – Supports estimations with various technologies
  – Open-source tool that is available at [http://accelergy.mit.edu/](http://accelergy.mit.edu/)

• **Validation on a conventional digital accelerator design**
  – 95% accurate on total energy consumption comparing to post layout simulation

• **Modeling of processing in memory (PIM) based DNN accelerator designs with a systematic evaluation system**
Accelergy High-Level Infrastructure

Architecture Description

Global Buffer (GLB)

PE0 ➔ PE2 ➔ PE3

Available at http://accelergy.mit.edu/
We use **primitive components**, i.e., key building blocks, to define a compound component.
Accelergy High-Level Infrastructure

**Architecture Description**
- Global Buffer (GLB)
- PE0
- PE2
- PE3

**Compound Component Description**
- GLB
  - SRAM
    - tech: 45nm
    - width: 64
    - depth: 1024
  - control
    - tech: 45nm
- PE
  - MAC
    - tech: 45nm
    - width: 16
  - latch
    - tech: 45nm
    - width: 16

**Energy Estimation Plug-ins**

Available at [http://accelergy.mit.edu/](http://accelergy.mit.edu/)
Accelergy High-Level Infrastructure

Architecture Description

Global Buffer (GLB)

PE0

PE2

PE3

Compound Component Description

GLB

SRAM

tech: 45nm
width: 64
depth: 1024

countrol

tech: 45nm

PE

MAC

tech: 45nm
width: 16

latch

tech: 45nm
width: 16

Simple Example Estimation Plug-in

<table>
<thead>
<tr>
<th>name</th>
<th>tech.</th>
<th>width</th>
<th>action</th>
<th>energy (pJ)</th>
<th>area (μm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC</td>
<td>45nm</td>
<td>16b</td>
<td>mac</td>
<td>1.5</td>
<td>800</td>
</tr>
<tr>
<td>latch</td>
<td>45nm</td>
<td>16b</td>
<td>access</td>
<td>0.2</td>
<td>20</td>
</tr>
</tbody>
</table>

Available at [http://accelergy.mit.edu/](http://accelergy.mit.edu/)
Accelergy High-Level Infrastructure

**Architecture Description**
- Global Buffer (GLB)
- PE:
  - PE0
  - PE2
  - PE3
- GLB: SRAM
  - tech: 45nm
  - width: 64
  - depth: 1024
- PE: MAC
  - tech: 45nm
  - width: 16
- PE: latch
  - tech: 45nm
  - width: 16

**Compound Component Description**
- Control
  - tech: 45nm

**Action Counts**
- PE0: compute: 1000
- PE1: ...

**Simple Example Estimation Plug-in**

<table>
<thead>
<tr>
<th>name</th>
<th>tech</th>
<th>width</th>
<th>action</th>
<th>energy (pJ)</th>
<th>area ((\mu m^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC</td>
<td>45nm</td>
<td>16b</td>
<td>mac</td>
<td>1.5</td>
<td>800</td>
</tr>
<tr>
<td>latch</td>
<td>45nm</td>
<td>16b</td>
<td>access</td>
<td>0.2</td>
<td>20</td>
</tr>
</tbody>
</table>

Comes from a performance model (e.g., cycle accurate simulator)

Available at [http://accelergy.mit.edu/](http://accelergy.mit.edu/)
Accelergy High-Level Infrastructure

Comes from a performance model (e.g., cycle accurate simulator)

<table>
<thead>
<tr>
<th>Name</th>
<th>Action</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE0</td>
<td>compute</td>
<td>1000</td>
</tr>
<tr>
<td>PE1</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Energy/Area Estimation

<table>
<thead>
<tr>
<th>Name</th>
<th>Energy (pJ)</th>
<th>Area ((\text{\mu m}^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE0</td>
<td>1700</td>
<td>820</td>
</tr>
<tr>
<td>PE1</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Simple Example Estimation Plug-in

<table>
<thead>
<tr>
<th>Name</th>
<th>Tech.</th>
<th>Width</th>
<th>Action</th>
<th>Energy (pJ)</th>
<th>Area ((\text{\mu m}^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC</td>
<td>45nm</td>
<td>16b</td>
<td>mac</td>
<td>1.5</td>
<td>800</td>
</tr>
<tr>
<td>latch</td>
<td>45nm</td>
<td>16b</td>
<td>access</td>
<td>0.2</td>
<td>20</td>
</tr>
</tbody>
</table>

Available at [http://accelergy.mit.edu/](http://accelergy.mit.edu/)
Plug-ins for Fine-Grain Action Energy Estimation

- External energy/area models that accurately reflect the properties of a macro
  - e.g., multiplier with zero-gating

Energy characterizations of the zero-gated multiplier
(normalized to idle)

<table>
<thead>
<tr>
<th>name</th>
<th>tech.</th>
<th>width</th>
<th>action</th>
<th>energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>multiplier</td>
<td>65nm</td>
<td>16b</td>
<td>random multiply</td>
<td>23.0</td>
</tr>
<tr>
<td>multiplier</td>
<td>65nm</td>
<td>16b</td>
<td>reused multiply</td>
<td>16.8</td>
</tr>
<tr>
<td>multiplier</td>
<td>65nm</td>
<td>16b</td>
<td>gated multiply</td>
<td>1.3</td>
</tr>
</tbody>
</table>
Plug-ins for Fine-Grain Action Energy Estimation

- External energy/area models that accurately reflect the properties of a macro
  - e.g., multiplier with zero-gating

Energy characterizations of the zero-gated multiplier (normalized to idle)

With the characterization provided in the plug-in, we can see significant energy savings for sparse workloads.
Plug-ins for Fine-Grain Action Energy Estimation

- External energy/area models that accurately reflect the properties of a macro
  - e.g., multiplier with zero-gating

Energy characterizations of the zero-gated multiplier (normalized to idle)

With the characterization provided in the plug-in, we can see significant energy savings for sparse workloads.
Outline

• Accelergy Infrastructure
  – Performs architecture-level estimations to enable rapid design space exploration
  – Supports modeling of diverse architecture designs
  – Supports estimations with various technologies
  – Open-source tool that is available at http://accelergy.mit.edu/

• Validation on a conventional digital accelerator design
  – 95% accurate on total energy consumption comparing to post layout simulation

• Modeling of processing in memory (PIM) based DNN accelerator designs with a systematic evaluation system
Energy Validation on Eyeriss [Chen, ISSCC 2016]

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

Eyeriss Architecture

Ifmap = input feature map
Psum = partial sum
PE = processing element
*_spad = * _scratchpad
Energy Validation on Eyeriss [Chen, ISSCC 2016]

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

Zero-gating optimization
If there is a 0 ifmap data
• Gate on reading the weights data => gated-read
• Gate on computing the MAC => gated-MAC
Energy Validation on Eyeriss [Chen, ISSCC 2016]

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

*Total energy might not add up to exact 100.0% due to rounding*
Comparisons with existing work: Aladdin[Shao, ISCA2014]
Systematic DNN Accelerator Evaluation System

- Require systematic way to
  - Evaluate and compare wide range of DNN processor designs
    - With different high-level architecture designs
    - With different underlying technologies
  - Rapidly explore design space

Flexible and fast DNN accelerator performance model

**Flexibility and performance**

- Architecture description
- Compound component description

**Output**

- Action Counts
- Energy & Area Estimations
- Energy estimation plug-in 0
- Energy estimation plug-in 1
- ...
Systematic DNN Accelerator Evaluation System

• Require systematic way to
  – Evaluate and compare wide range of DNN processor designs
    • With different high-level architecture designs
    • With different underlying technologies
  – Rapidly explore design space
• Timeloop [Parashar, ISPASS2019]
  – Optimizing compiler for DNN processing
  – Performance Model
Outline

• **Accelergy Infrastructure**
  – Performs architecture-level estimations to enable rapid design space exploration
  – Supports modeling of diverse architecture designs
  – Supports estimations with various technologies
  – Open-source tool that is available at [http://accelergy.mit.edu/](http://accelergy.mit.edu/)

• **Validation on a conventional digital accelerator design**
  – 95% accurate on total energy consumption comparing to post layout simulation

• **Modeling of processing in memory (PIM) based DNN accelerator designs with a systematic evaluation system**
PIM DNN Architectures

- Example PIM DNN architectures

Activation is input voltage \( V_i \)

Weight is resistor conductance \( G_i \) \[= 1/\text{resistance}\]

Partial sum is output current \( I_i \)

\[ I = I_1 + I_2 = V_1 \times G_1 + V_2 \times G_2 \]
Estimation for PIM Accelerators

Architecture Description
- Global Buffer (GLB)
  - SRAM: tech: 45nm, width: 64, depth: 1024
  - Control: tech: 45nm
- PE
  - MAC: tech: memristor, width: 16

Example Estimation Plug-in

Action Counts
<table>
<thead>
<tr>
<th>name</th>
<th>action</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEO</td>
<td>compute</td>
<td>1000</td>
</tr>
<tr>
<td>PE1</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

Energy/Area Estimation
* scaled from ISAAC [MICRO 2016]

<table>
<thead>
<tr>
<th>name</th>
<th>energy (pJ)</th>
<th>area (μm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEO</td>
<td>1700</td>
<td>14.6</td>
</tr>
<tr>
<td>PE1</td>
<td></td>
<td>820-1.21E-2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>name</th>
<th>tech.</th>
<th>width</th>
<th>action</th>
<th>energy (pJ)</th>
<th>area (μm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC</td>
<td>45nm</td>
<td>16b</td>
<td>mac</td>
<td>1.5</td>
<td>800</td>
</tr>
<tr>
<td>MAC</td>
<td>memristor</td>
<td>16b</td>
<td>mac</td>
<td>1.46E-2*</td>
<td>1.21E-2*</td>
</tr>
<tr>
<td>ADC</td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>DAC</td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
Accelry Modeling of PIM Architectures

• Parameterizable templates that allow easy representation of PIM architectures
Accelergy Modeling of PIM Architectures

• Parameterizable templates
  – Architecture Template allows architecture parameter sweeping, e.g.,
    • number of PE rows
    • number of PE columns
    • size of global buffer, etc.
  – Component design template allows implementation optimization, e.g.,
    • optimize DAC-based D2A conversion system
    • optimize the design of the flash ADC in the A2D conversion system, etc.
Energy Modeling Validation on PIM Design

• Validation on the ADC-based design proposed in CASCADE [Chou, MICRO2019]
• Design Specs
  – 80 64x64 1-bit Memristor Arrays
  – 1-bit DACs
  – 6-bit ADCs
  – 16-bit data representations
• Workload: VGG Net convolutional layers
• Energy estimation tables: extracted numbers from the paper/cited sources
Energy Modeling Validation on PIM Design

Total Energy Estimation and Breakdown Validation

The architecture is correctly modeled:
- 95% accurate total energy estimation
- tracks the breakdown across different components

Total energy might not add up to exact 100.0% due to rounding.
Energy Modeling on PIM Design

Energy Breakdown Across VGG Convolutional Layers

Captures the energy breakdown of each convolutional layer

Published at [Wu, ISPASS 2020]
Summary

• Accelergy is an architecture-level energy estimator that
  – Accelerates accelerator design space exploration
  – Provides flexibility to
    • Describe and evaluated a wide range of accelerator designs
    • Support different technologies with user defined plug-ins, e.g., CMOS, RRAM, etc.
  – Achieves high accuracy energy estimations
    • 95% accurate for the Eyeriss accelerator and Cascade PIM accelerator

• The Timeloop-Accelergy system allows fast explorations on
  – High-level architecture properties, e.g., PE array size
  – Lower-level implementation optimizations on the components in the design, e.g., ADC design optimizations

Acknowledgement: DARPA, Facebook, MIT Presidential Fellowship
Resources

• **Papers:**

• **Accelergy Website:** [http://accelergy.mit.edu/](http://accelergy.mit.edu/)
  – Open-sourced tools and infrastructure docker
  – Slide decks
  – MICRO2019 tutorial

• **Upcoming Tutorial at ISCA2020**
  – Free and open to public
  – Will happen between May 30th – June 3rd
  – Email us at accelergy@mit.edu for any questions!