6.5930/1

Hardware Architectures for Deep Learning

Final Project

April 1, 2024

Joel Emer and Vivienne Sze

Massachusetts Institute of Technology
Electrical Engineering & Computer Science

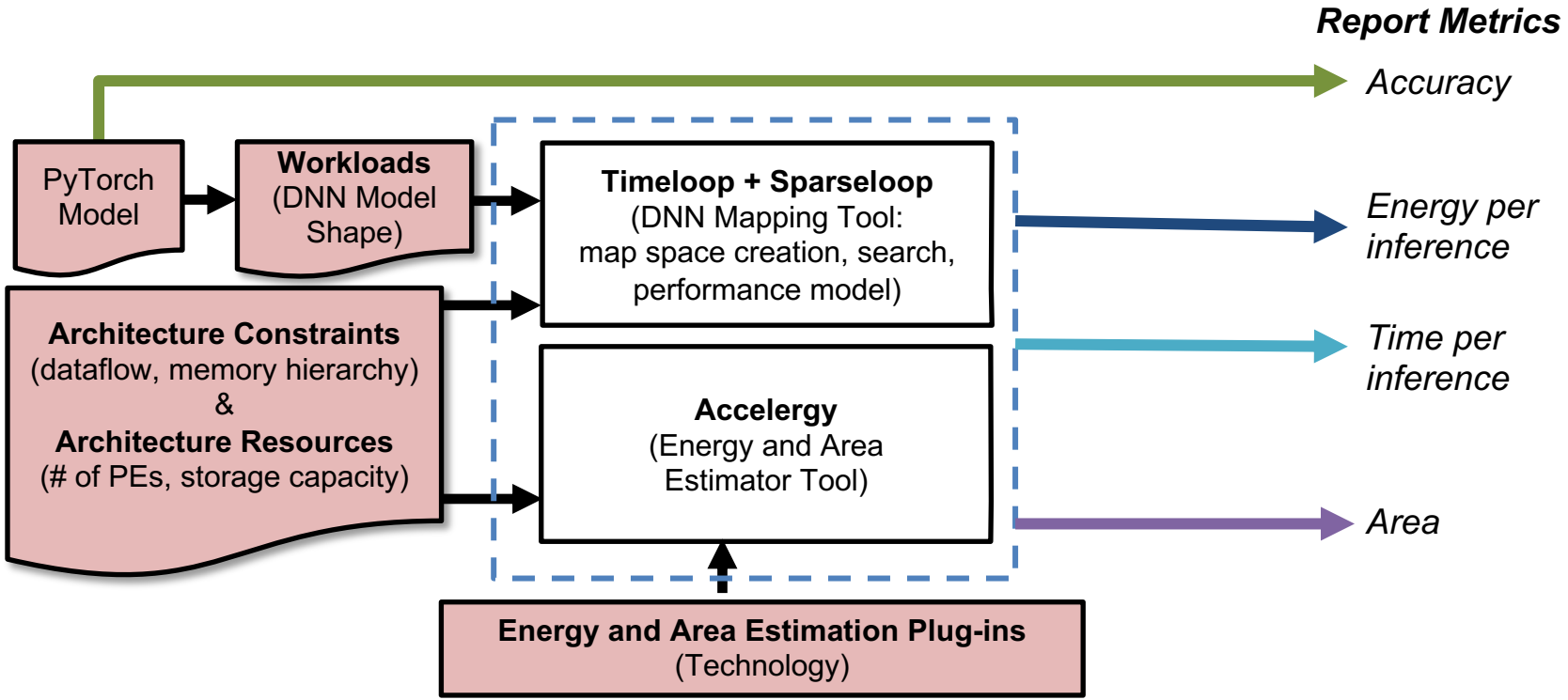
Final Project (6.5930 students only)

- Project
  - Choose from a list of suggested projects
  - May propose own project (requires formal proposal and approval by course staff)
- Teams
  - Two or three people
  - Each will have a TA mentor
- Schedule
  - April 1 – List of projects released
  - **April 8** – Submit project selection and proposal
  - May 6 – Project report due
  - May 8/13 – Poster presentation

Info on Final Project

- Based on tools used in labs (especially Labs 3 & 4)
  - Timeloop & Sparseloop (Mapping and Action Counts)
  - Accelergy (Energy and Area Estimation)
- You will be given baseline designs
- Example projects: ***Only need to change one aspect of design***
  - Workload (DNN Model)
  - Plug-in (Technology)
  - Architecture Constraints (Dataflow & Memory Hierarchy)
  - Architecture Resources (# of PEs & Storage Capacity)
- Stretch goal: Change more than one aspect of design
  - e.g., Architecture Constraints (Dataflow & Memory Hierarchy) + Plug-ins
- Report metrics
  - Area, energy consumption per inference, time per inference
  - If change DNN model, need to also report accuracy

More Info on Final Project



Inputs to tools highlighted in ***red.***
Project can involve changing one or more of these inputs

Example Projects

1. New technology (RRAM, SRAM, Optical) **[Plug-ins]**
   - Develop new plug-in and evaluate with weight-stationary architecture
2. Evaluate behavior of existing DNN Models **[Workloads]**
   - Select at least three CNN image classification models from survey paper: https://ieeexplore.ieee.org/document/8506339, or explore other application domains, e.g., video, or other types of DNNs, e.g., transformers
   - Evaluate on two architectures and analyze impact of architecture characteristics
3. Explore workloads, e.g., multi-branch, RNNs, transformers **[Workloads]**
   - Evaluate on single architecture and match existing results
   - Extend converter from PyTorch DNN models (or ONNX) to new workloads

Example Projects

4. Iso area design space exploration **[Architecture Resources]**
   - Evaluate single architecture for fixed area, change # of PEs and buffer size
   - Create search algorithm and tool
5. Model and analyze previously proposed architecture **[Architecture]**
   - Compare to a baseline model across diverse workloads and analyze benefits
     - For example, build model for one of the papers explored during paper review (e.g., Tetris, Wire-aware, TPU-1, Heterogenous Dataflow Accelerator, Short Cut Mining, Fused-Layer, MAERI, etc.)
   - Explore impact of sweeping design parameters, e.g., buffer partitioning
6. Propose and evaluate new architecture **[Architecture]**
   - Multiple arrays (weight stationary)
   - Evaluate pipelining of multiple layers in a single accelerator
   - Explore impact of sparsity

Project Proposal (Due Mon, April 8)

- **Brief overview of project**
  - Indicate which example project your project is based on, or
  - If different from example projects, you will need to provide more details
- **Motivation (Justify selection of problem)**
  - What is the problem? What is/are the key question(s) your project is trying to answer? Why is it important? If successful, what difference will your results make?
- **Technical Contributions**
  - What are the current solutions (cite previous work)? How is your approach different? Why do you think it will work? What are the key challenges that you might face?
- **Evaluation (How will you evaluate your solution)**
  - What metrics does it affect (energy, speed, cost, accuracy)? What will you compare with (baseline)? What will you measure to quantify your results? Are there any downsides (i.e., what are the tradeoffs)? What are the experiments/simulations that you will need to perform to validate your approach?
- **Timeline (Plan for the next four weeks)**
  - What are the key milestones and how do you plan to divide up the work in the team?

Project Deliverables

- May 6 – Project Report
  - 4 pages, not including references
  - References can take up additional page(s) [unlimited]
  - Use template
    - https://www.iscaconf.org/isca2020/submit/isca2020-latex-template.zip
    - Make sure submission is NOT anonymous (turn off anonymous flags)
  - Technical summary sheet (does not count towards page limit)

- May 8 & 13 – Poster Presentations

Report/Poster Requirements

- Clearly introduce the problem and why it is important
- Describe and cite previous work (2-3 references) – how is it different from what you did in the project
- Describe the three most interesting insights / innovative ideas pursued in your project
  - Make sure you indicate what parts worked and what parts did not work
- Provide any suggestions for future work
- Make the figures clear – i.e., keep the font large enough for readability, do not rely on color for your figures, do not try to include too much detail

Technical Summary Sheet (1-page)

- Project Title
- Authors
- List the three key insights/innovations of project
- List the contributions of each partner
- Provide name of project repository submission

Project Grading (20 points)

- **Project Proposal + Problem Definition and Relevance (2 Points)**
  - What are you going to study?
  - Why is it important?
  - What baseline are you going to use? What are you going to vary?
- **Technical Contributions & Evaluation (10 points)**
  - Summary of new insights. Do we learn anything new from this work?
  - Identify key technical challenges. What is the difficulty of the problem?
  - Description of experiment setup. Are the experiments unbiased?
  - Quantified results across relevant metrics (e.g., accuracy, energy, speed, area).
  - Description and evaluation of tradeoffs (benefit versus cost)
  - Comprehensive analysis and interpretation of results. Detailed insights.
- **Report and Poster (4 Points)**
  - Description of project idea and results
  - Judged on clarity, organization and intuition behind project
    - Is intuition given on the technical approach? Is the paper and presentation well organized? Are figures/graph readable and clearly labeled? Are they explained in the text and talk?
  - Contributions of team members
- **Artifacts – (2 Points)**
  - Reusable documented infrastructure for simulator, configuration files, workloads, etc.
  - Usability by others (e.g., to be used by students in future classes or made open-sourced for the research community)