Roadmap and Concluding Remarks

Manuel E. Acacio

Agenda

Attention: Tutorial is being recorded

Time (CET)	Time (ET)	Topic	Presenter
14:00 – 14:40	8:00 – 8:40	**Flexible Accelerators**	Tushar Krishna
14:40 – 15:10	8:40 – 9:10	**Cycle accurate simulation and Overview of STONNE**	José Luis Abellán
15:10 – 16:10	9:10 – 10:10	**(Hands-on) STONNE Deep-Dive**	Francisco Muñoz-Martínez
16:10 – 16:40	10:10 – 10:40	**Coffee Break**	
16:40 – 17:10	10:40 – 11:10	**(Hands-on) STONNE Deep-Dive**	Francisco Muñoz-Martínez
17:10 – 17:40	11:10 – 11:40	**Dataflow exploration for Graph Neural Networks**	Raveesh Garg
17:50 – 18:00	11:50 – 12:00	**Roadmap for Future Development**	Manuel Acacio

Tutorial Website https://stonne-simulator.github.io/ASPLOSTUT.html

includes agenda and STONNE/OMEGA installation instructions

STONNE Roadmap

- Current version of STONNE (v1) publicly available at:

  `https://github.com/stonne-simulator/stonne`

- Ongoing and future development of the STONNE framework envisages two main directions:

  1. Extension of the functionality of STONNE

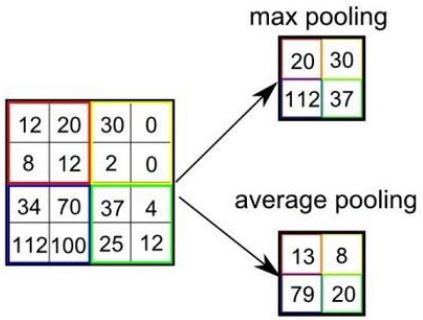
  2. Integration of STONNE with other tools

- Several of these new features are expected to be available in a subsequent release (v2) scheduled by end of the July 2022

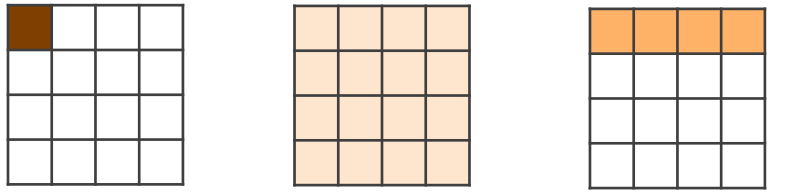
STONNE Roadmap

1. Extension of the functionality of STONNE:

- Simulation of non computation-intensive layers (e.g. Max/Avg Pooling)

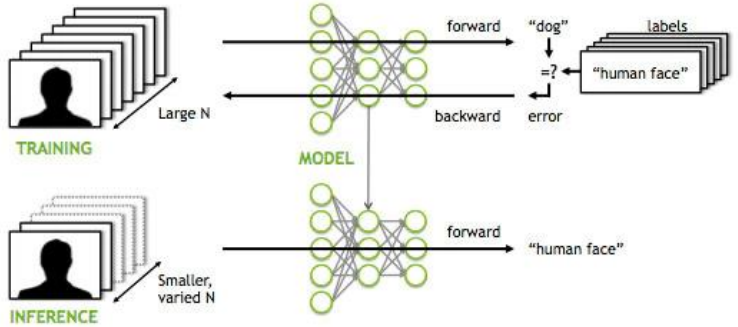


- Additional dataflows for GEMM computation:



- Link with other DL frameworks (e.g. TensorFlow Lite)

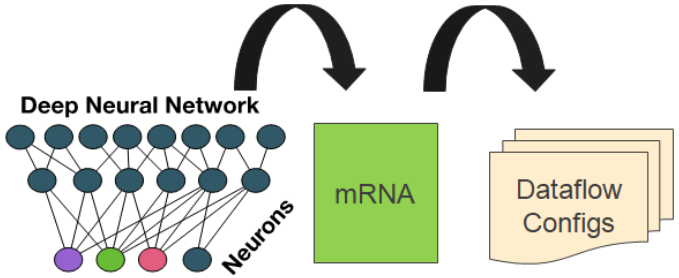
- Simulation of training processes



STONNE Roadmap

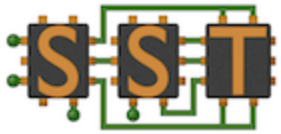
2. Integration of STONNE with other tools:

- Inclusion of **mRNA** tool[1] for transparent configuration of flexible accelerator architectures



- Inclusion of STONNE in **SST**[2] (Structural Simulation Toolkit)

`https://github.com/stonne-simulator/sst-elements-with-stonne`



- Support for a detailed memory hierarchy (e.g. *Buffets*[3] support), including DRAMsim3[4] connection

- Integration of STONNE in gem5[5] to allow for detailed study of CPU-Accelerator interaction

[1]Z. Zhao *et al.*, "*mRNA: Enabling Efficient Mapping Space Exploration for a Reconfigurable Neural Accelerator*," Proc. of ISPASS 2019.
[2]http://sst-simulator.org/
[3]M. Pellauer *et al.*, "*Buffets: An efficient and composable storage idiom for explicit decoupled data orchestration*," Proc. of ASPLOS 2019.
[4]https://github.com/umd-memsys/DRAMsim3
[5]https://www.gem5.org/

OMEGA Roadmap

- Current version of OMEGA (v1) publicly available at:

  `https://github.com/stonne-simulator/omega`

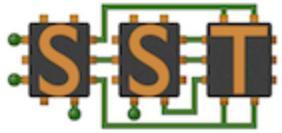
- Ongoing and future development of the OMEGA framework also envisages two main directions:

  1. Extension of the functionality of OMEGA

  2. Integration of OMEGA with other tools

OMEGA Roadmap

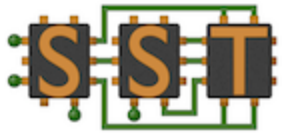
1. Extension of the functionality of OMEGA:
    - Generalize the analysis for dataflows for multiphase computation like DL recommendation models or HPC kernels like Conjungate gradient.
    - Build a mapping optimizer on top of this work:
        - The **taxonomy** of the proposed dataflows will help formalize the design-space for mapping search.
        - The **OMEGA framework** can be used as a cost model that the mapping search tool can employ.



OMEGA Roadmap

2. Integration of OMEGA with other tools:

- It is possible to use the OMEGA frameworks with other simulation tools or analytical models in addition to STONNE.
- The parameters and timestamps from these tools can be fed into the inter-phase cost model.
- With STONNE's integration with SST, it would also be possible to instantiate STONNE two times and write a configuration file for each inter-phase dataflow strategy to execute GNNs and study the execution of GNNs with traditional memory hierarchy.



Concluding remarks

- Some takeaways:
  - Use of current analytical models can lead to significant inaccuracies in performance and energy consumption estimations when it comes to more complex DDN accelerator microarchitectures
  - Having to build the RTL model of a DNN accelerator does not allow for rapid quantification of the efficacy of architectural enhancements during the early stages of a design

    ⇨ Need for cycle-level DNN architectural simulation

  - STONNE is able to model rigid and flexible accelerators, and data-dependent optimizations, all performing the actual computation of complete DNN models
    - STONNE can be easily extended to model new accelerator architectures
  - Also, STONNE can be a building block for modeling other DL accelerators:
    - The OMEGA framework builds on top of STONNE to enable modeling of the cost of the pipelined GNN dataflows

Concluding remarks

- Willing to contribute?
  - For bug notices please create a new issue on the corresponding Git repository (main simulator, SST plugin or OMEGA)

    `https://github.com/stonne-simulator`

  - For additional extensions to be included in the public release or suggestions for enhancements, please contact:

    `stonnesimulator@gmail.com`

Thank you for your interest in the STONNE ecosystem!

Tushar Krishna

José L. Abellán

Franciso Muñoz-Martínez

Raveesh Garg

Manuel E. Acacio

STONNE: A Simulation Tool for Neural Network Engines

Francisco Muñoz-Martínez, Raveesh Garg,
José L. Abellán, Manuel E. Acacio, Tushar Krishna