Cycle Accurate Simulation of AI Applications using STONNE, SST-STONNE and OMEGA

Francisco Muñoz-Martínez*[1], ***Raveesh Garg*[2]***, José L. Abellán[1], Manuel E. Acacio[1], Clay Hughes[3], Siva Rajamanickam[3], and Tushar Krishna[2]

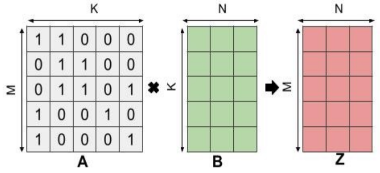
*Joint contribution

[1]Universidad de Murcia, [2]Georgia Institute of Technology, [3]Sandia National Laboratories

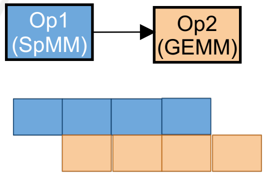
Georgia Tech College of Computing
Center for Research into
Novel Computing Hierarchies

Complexity in AI execution

Complexity in applications

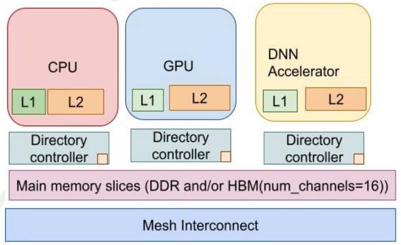


Sparsity in one or more tensor



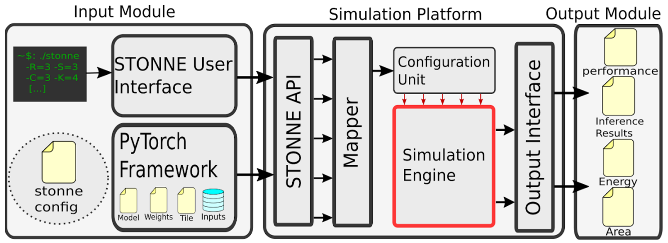
Pipelining dependent operations
Eg. Graph Neural Networks (GNNs)

Complexity in Hardware

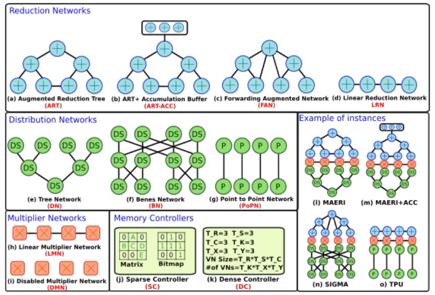


Heterogeneous Hardware with shared memory

STONNE: A Simulation TOol for Neural Networks Engines

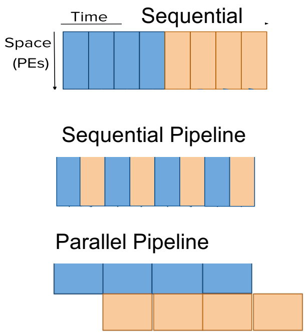


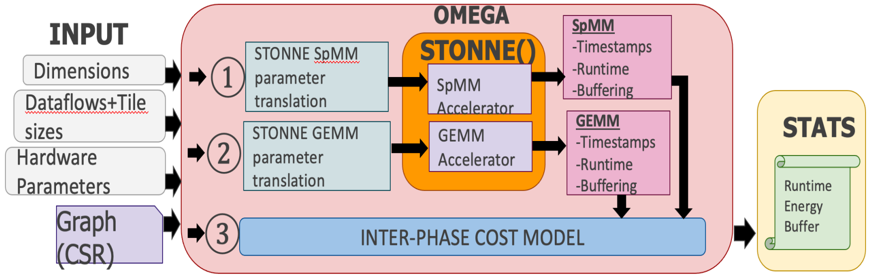
```
~$: ./stonne
  -R=3 -S=3
  -C=3 -K=4
  [...]
```



- A cycle accurate simulator running DNN models on flexible accelerators.
- Can model any topology for distribution, reduction and multiplier network, thus modular and flexible.
- Written in C++. Reports performance for DenseGEMM, DenseCONV, SpMM and SpGEMM kernels.
- **Source:** https://github.com/stonne-simulator/stonne

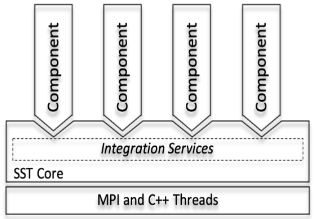
OMEGA: Observing Mapping Efficiency over GNN Accelerators

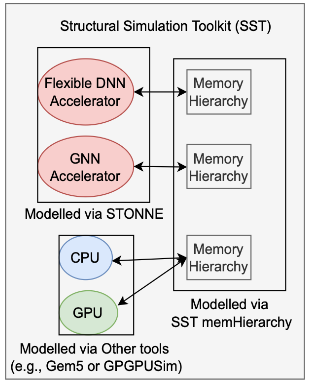


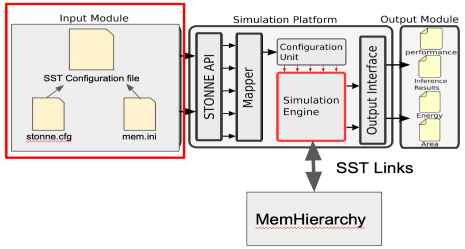


- Builds over STONNE to add support for inter-operation pipelining.
- Models inter-operation dataflow choices for Graph Neural Networks
- **Source:** https://github.com/stonne-simulator/omega

SST-STONNE Integration to Model Complex Backends







- **S**tructural **S**imulation **T**oolkit (SST) enables full-system simulation between multiple components.
- Integrates STONNE simulator instances with a memory hierarchy.
- **Source** - https://github.com/stonne-simulator/sst-elements-with-stonne