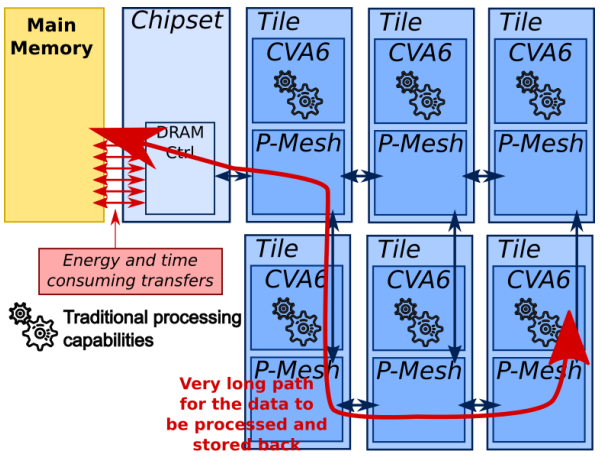
RIDIM:Reconfigurable stream dataflow computing near memory

Kevin Martin, Philippe Coussy, Univ. Bretagne-Sud, Lab-STICC
Jean-François Nezan, Maxime Pelcat, INSA Rennes, IETR
Steven Derrien, Univ. RennesI, Irisa/INRIA
Shuvra S. Bhattacharyya, Univ. of Maryland, College Park, USA

Summary

Holistic software/hardware model by integrating processing capabilities all along the path from the main memory to the processor
Model of computation "Passive-Active Flow Graph" (PAFG)

Typical NoC-based many-core



Some key figures

One addition:
- 3 memory accesses
- 1 operation

One adder: 1 cycle, few picoJoules
One memory access: 400 cycles, 300 nanoJoules

The total energy spent for moving data has reached excessive proportions

- 62% of energy in mobile systems
- 80% area of a chip dedicated to memory and data movement

Related work

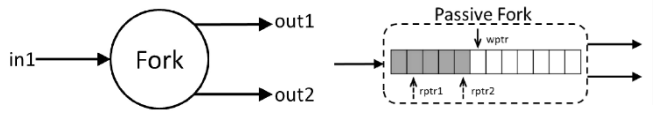
- SnackNoC
  - Need to rewrite the application to follow a producer-consumer data model
- *Data-stream processing*
  - KPN model of computation, no energy study
- Near-Memory Computing (NMC)
  - Software stack
- In-Memory Computing (IMC)
  - Tool-chain

Karthik Sangaiah, Michael Lui, Ragh Kuttappa, Baris Taskin, and Mark Hempstead. *Snacknoc: Processing in the communication layer.* HPCA, 2020.
Jens Rettkowski and Diana Göhringer. *Data stream processing in networks-on-chip.* ISVLSI, 2017.
Gagandeep Singh, Lorenzo Chelini, Stefano Corda, Ahsan Javed Awan, Sander Stuijk, Roel Jordans, Henk Corporaal, and Albert-Jan Boonstra. *Near-memory computing: Past, present, and future*. Microprocessors and Microsystems, 2019.

PAFG for non-experts

Generalization of the concept of dataflow edges into multi-input, multi-output components that are called "passive blocks".

- Designer-specified memory management optimization
- Flexible embedding of computations into passive blocks




```
Firing of fork actor:
1. Read data from input FIFO in1 and store it
   in a temporary variable d1
2. Send data value in d1 to output FIFO out1
3. Send data value in d1 to output FIFO out2
```

Passivization of the fork actor:
- ✓ No explicit data movement
- ✓ No firing

Proposed Approach

Compute power inside the routers of the network-on-chip, or inside the DRAM controller.
CGRA: Coarse Grained Reconfigurable Architecture
- ✓ Reconfigurable component at word-level

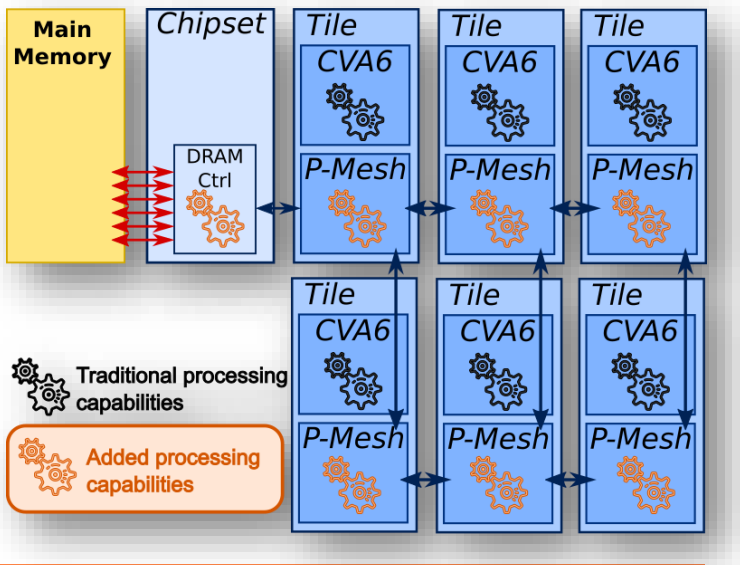
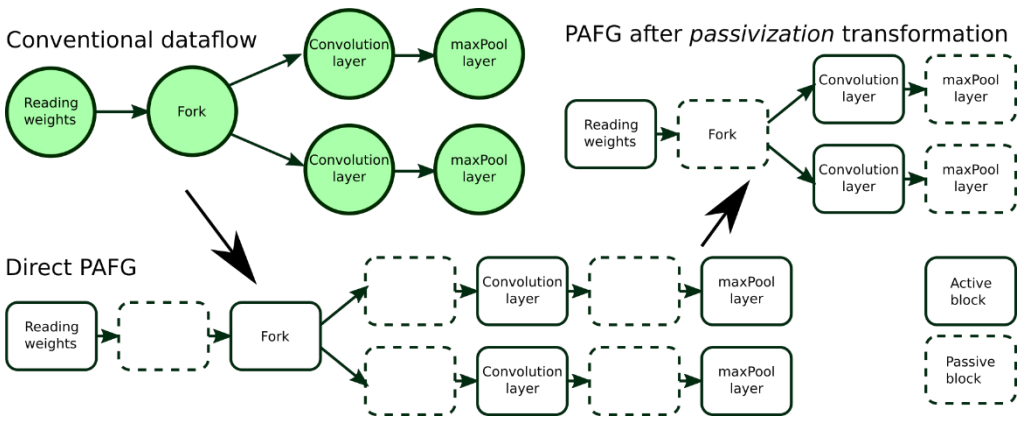


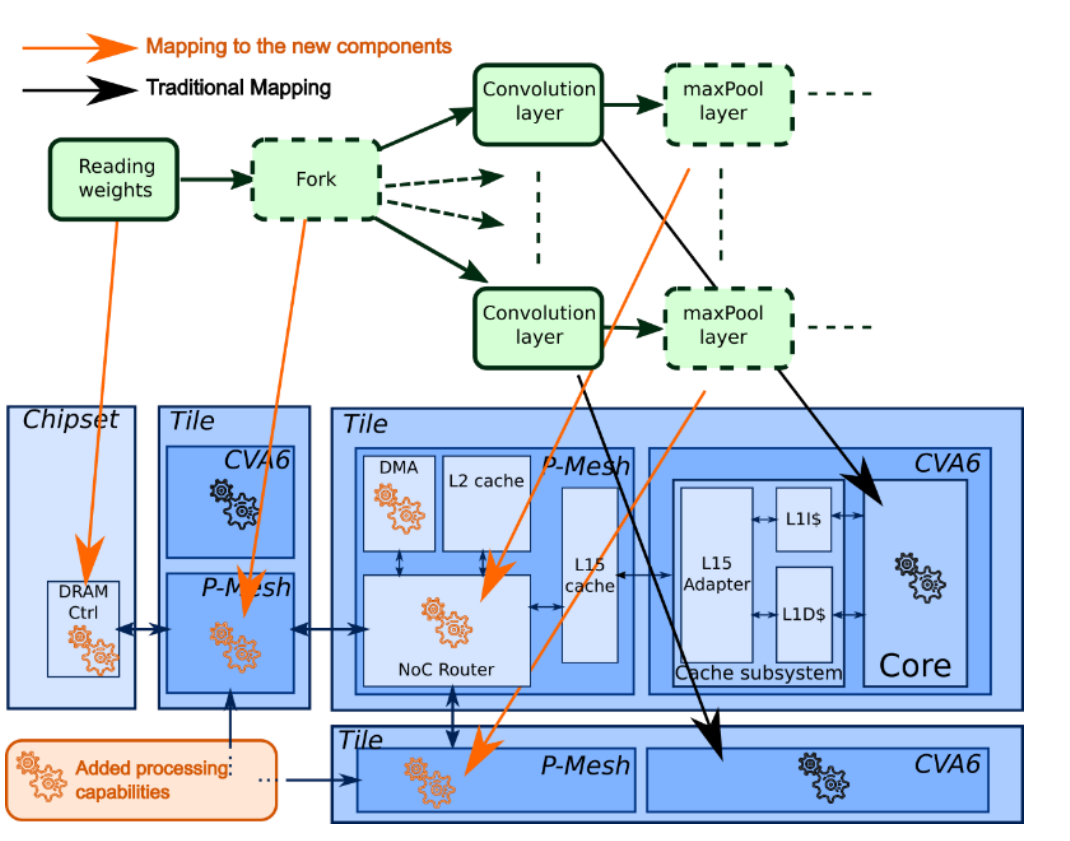
Illustration of the approach

Squeezenet:

Deep neural network for computer vision, designed for small networks and lower number of parameters, while achieving the same level of accuracy than bigger networks



MaxPool layer:
- Several megabytes as input
- Few hundreds of kilobytes as output
- ✓ Good candidate for passivization



Mapping of passive actors in the network
Mapping of active actors on the processors