Plasticine: A Reconfigurable Architecture For Parallel Patterns

By: **Snehil Verma**

Design Gap

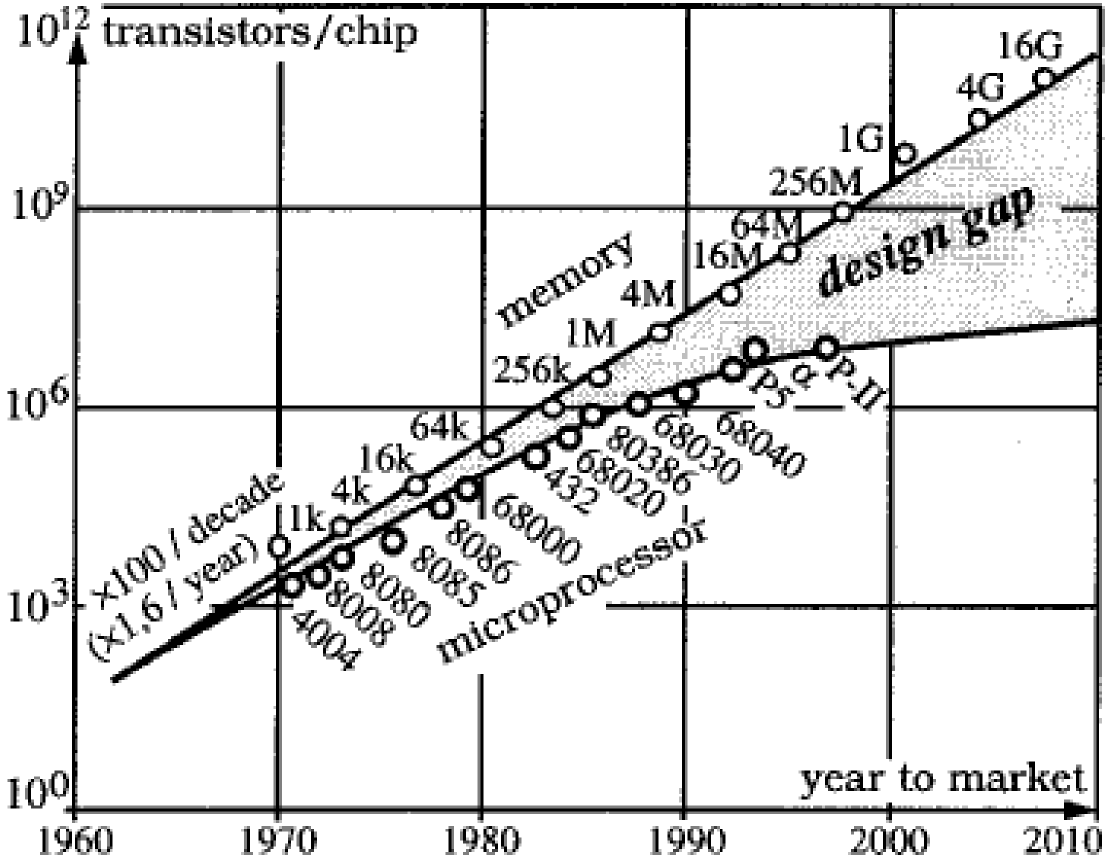
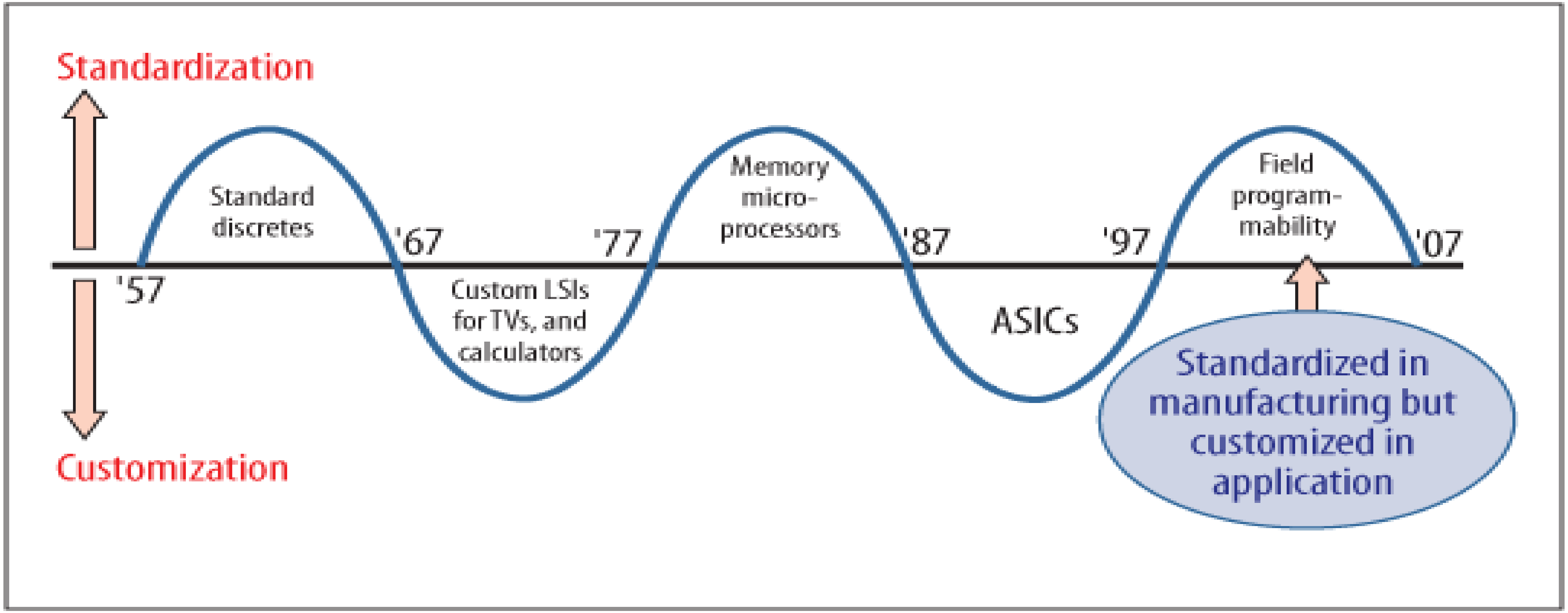


Fig. 2: The Gordon Moore curve w. design gap [11].

Makimoto wave



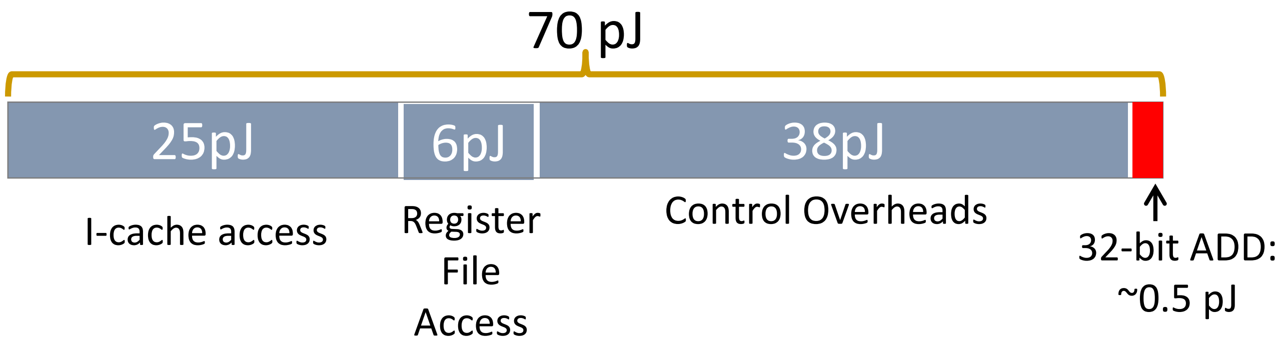
Flexibility: Programmable vs. Reconfigurable

- Programmable hardware: Instruction-based processors (CPUs, GPGPUs)
- Reconfigurable hardware: Statically reconfigurable datapaths (FPGAs, CGRAs)

Flexibility: Programmable vs. Reconfigurable

- Programmable hardware: Instruction-based processors (CPUs, GPGPUs)
- Reconfigurable hardware: Statically reconfigurable datapaths (FPGAs, CGRAs)

- Instructions **add overheads**:
  - Instruction fetch, decode, register file
  - 40% of datapath energy on CPU[1]
  - 30% of dynamic power on GPU [2]



- Reconfigurable hardware: **No instruction overheads**

[1] Hameed et al, Understanding Sources of Inefficiency in General-purpose Chips, ISCA 2010

[2] Leng et al, GPUWattch: Enabling Energy Optimizations in GPGPUs, ISCA 2013

FPGA: The Good And Bad

- **Bit-level reconfigurable logic elements + static interconnect**

Baidu 百度

intel FPGA

Microsoft

amazon web services

FPGA: The Good And Bad

- **Bit-level reconfigurable logic elements + static interconnect**

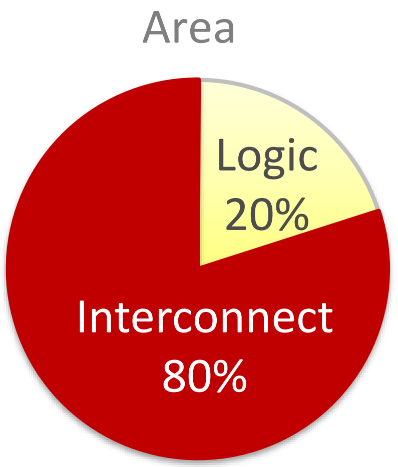
- **Flexibility**
- **Performance / Watt**
- **Commercially successful, mature toolchain support**

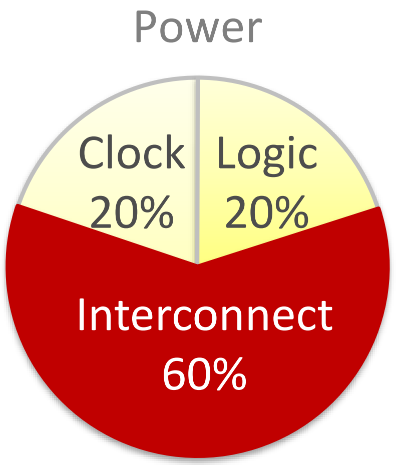
FPGA: The Good And Bad

- **Bit-level reconfigurable logic elements + static interconnect**

- **Flexibility**
- **Performance / Watt**
- **Commercially successful, mature toolchain support**

- **Fine-grained reconfigurability overheads:**
  - 60% area, power spent in the interconnect
- **Long compile times (hours)**
- **Low-level programming models**



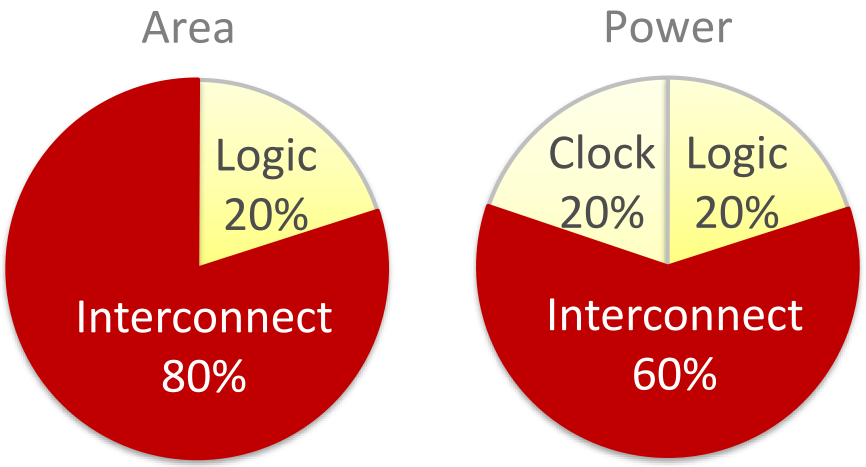


FPGA: The Good And Bad

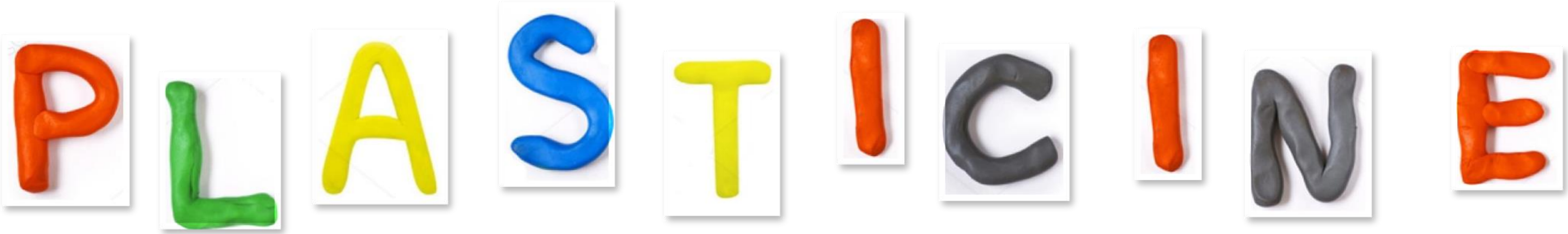
- **Bit-level reconfigurable logic elements + static interconnect**

- **Flexibility**
- **Performance / Watt**
- **Commercially successful, mature toolchain support**

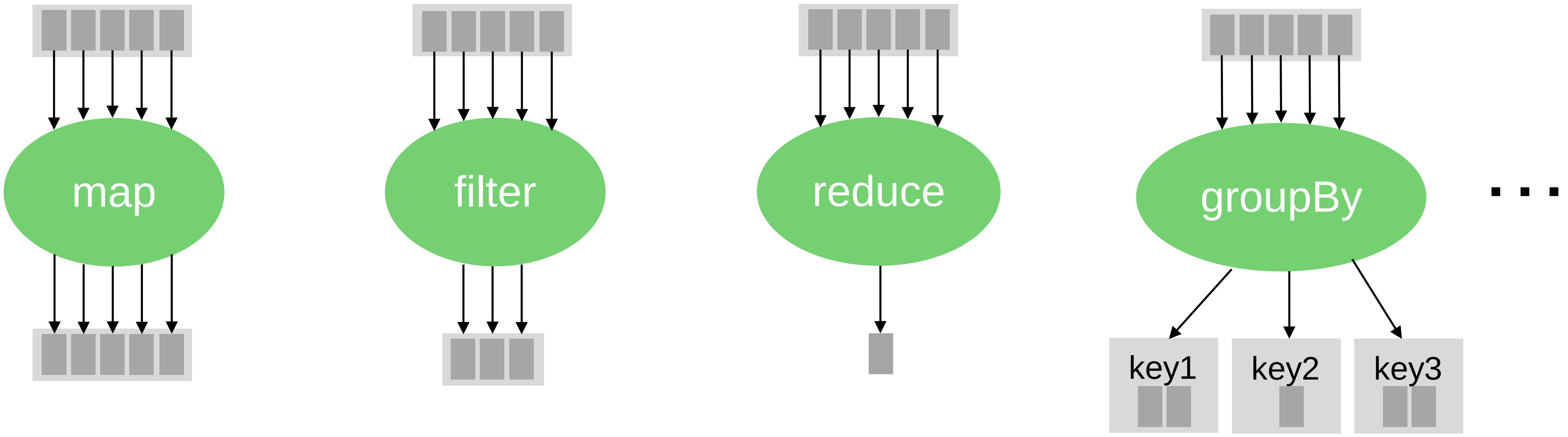
- **Fine-grained reconfigurability overheads:**
  - 60% area, power spent in the interconnect
- **Long compile times (hours)**
- **Low-level programming models**



Design reconfigurable hardware with the right abstractions

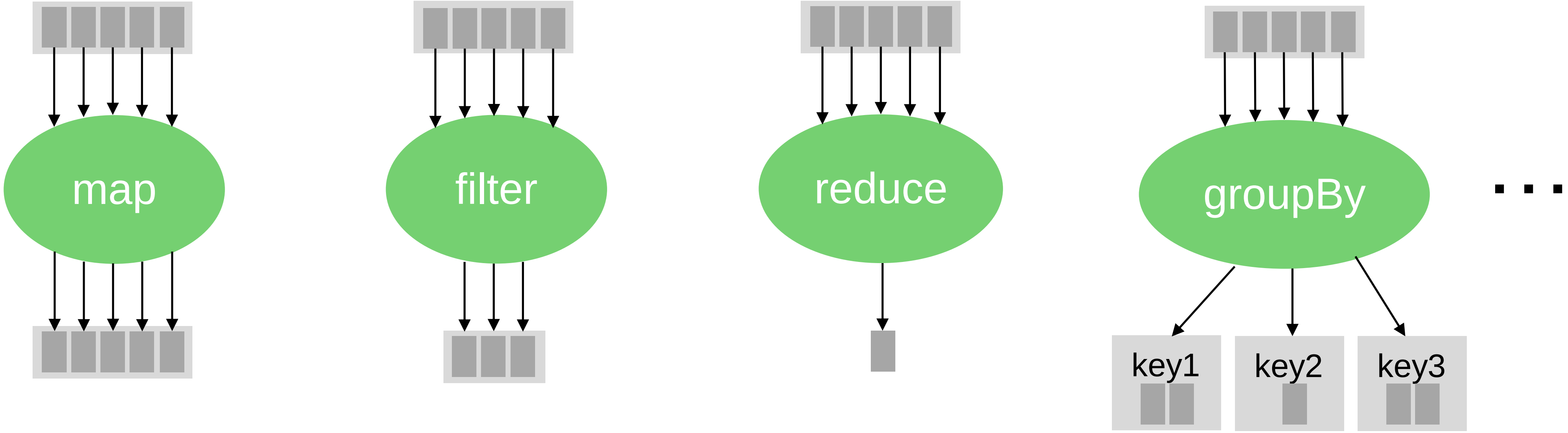

PLASTICINE

Parallel Patterns



- Captures parallelism, locality
- High-level, expressive

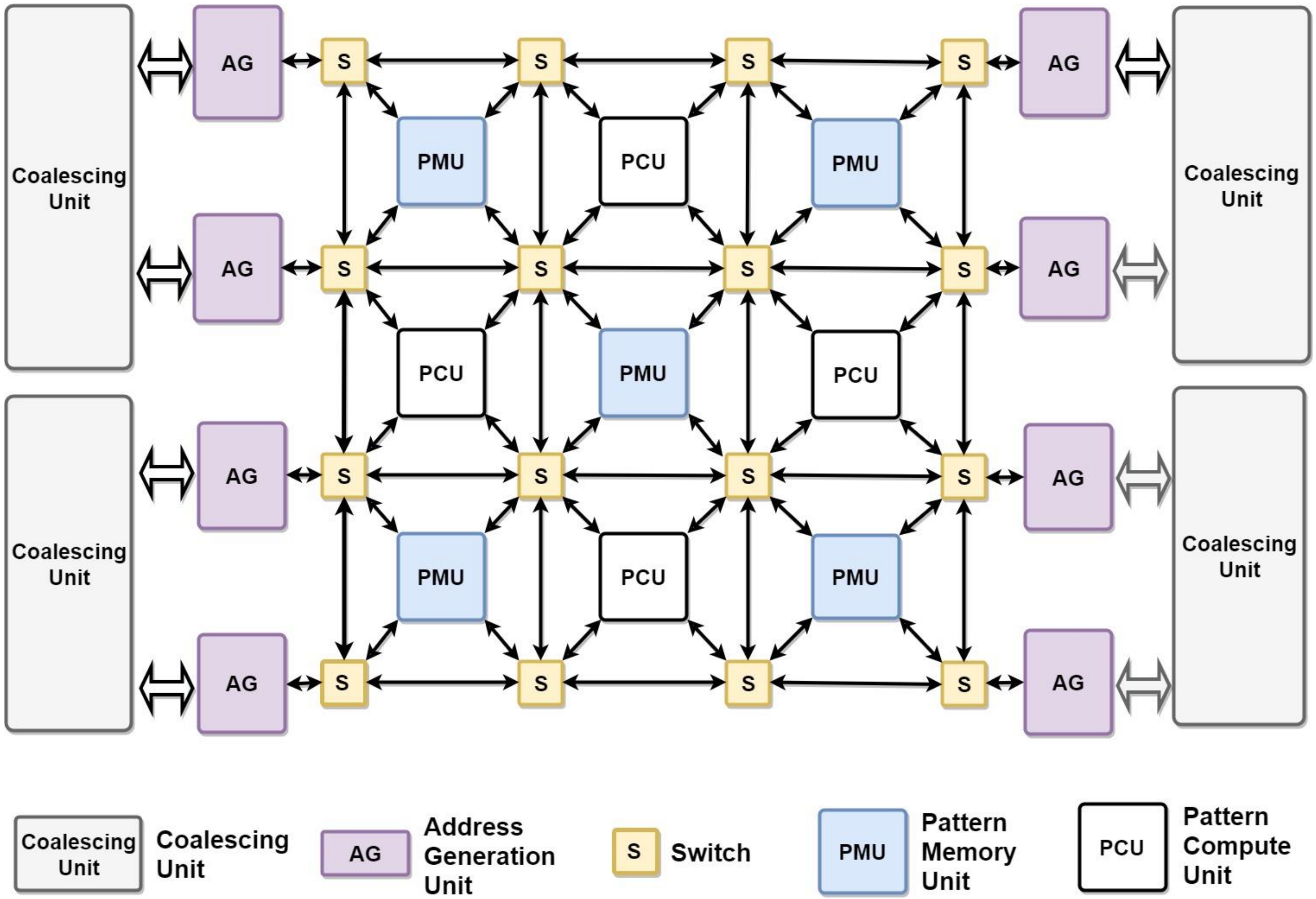
Parallel Patterns



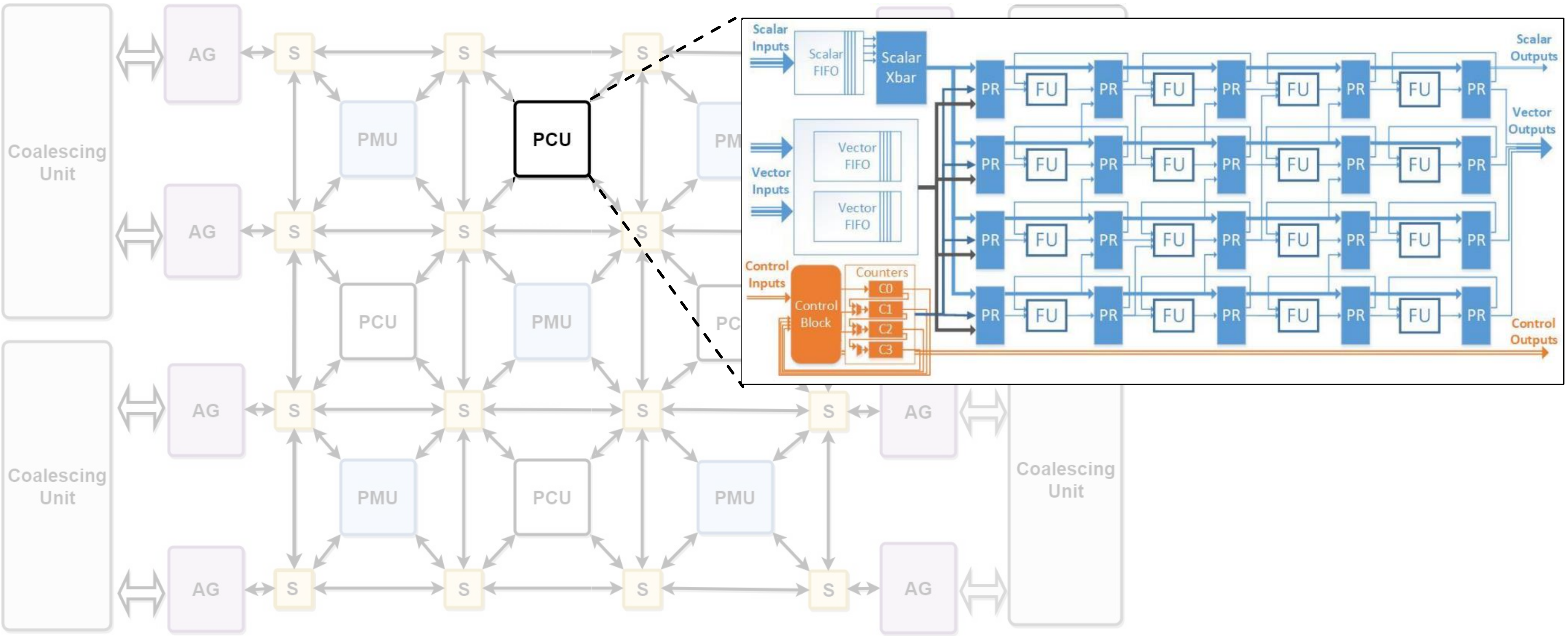
- Captures parallelism, locality
- High-level, expressive

Design reconfigurable primitives to accelerate parallel patterns

Plasticine: Top-Level

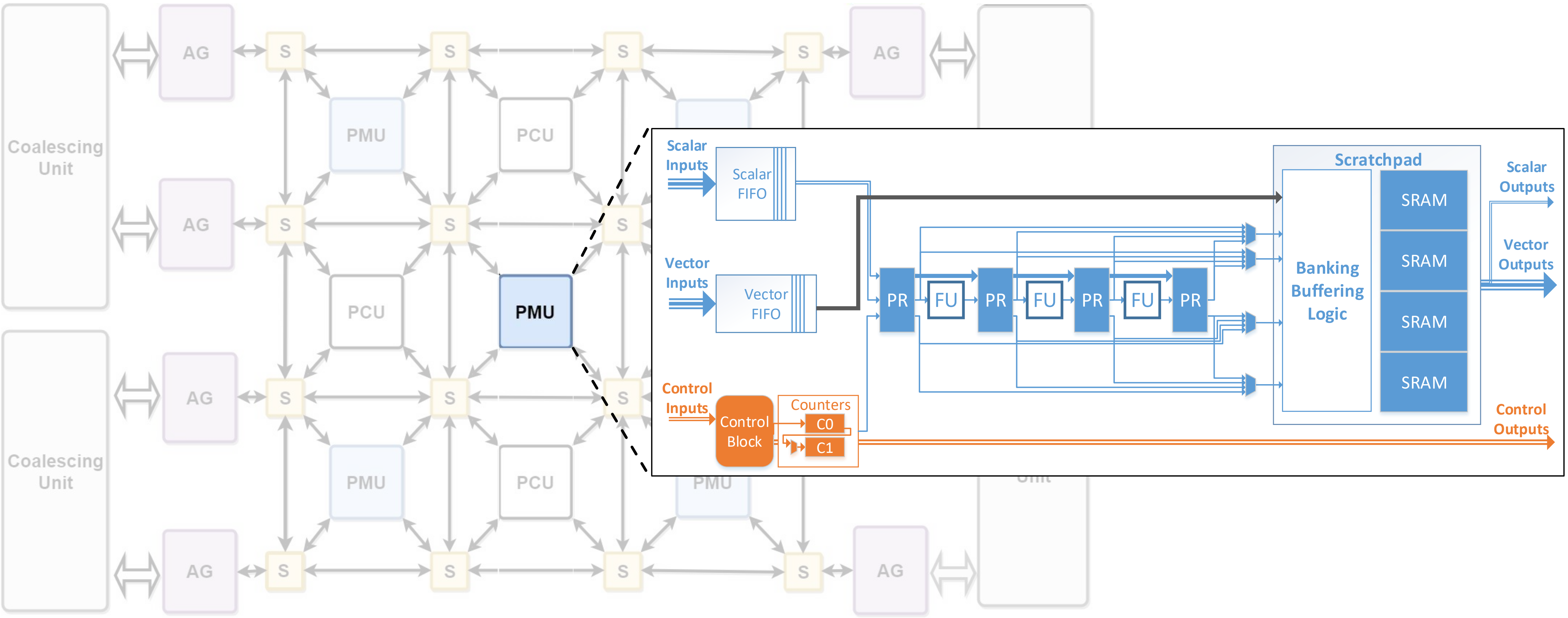


Plasticine: PCU



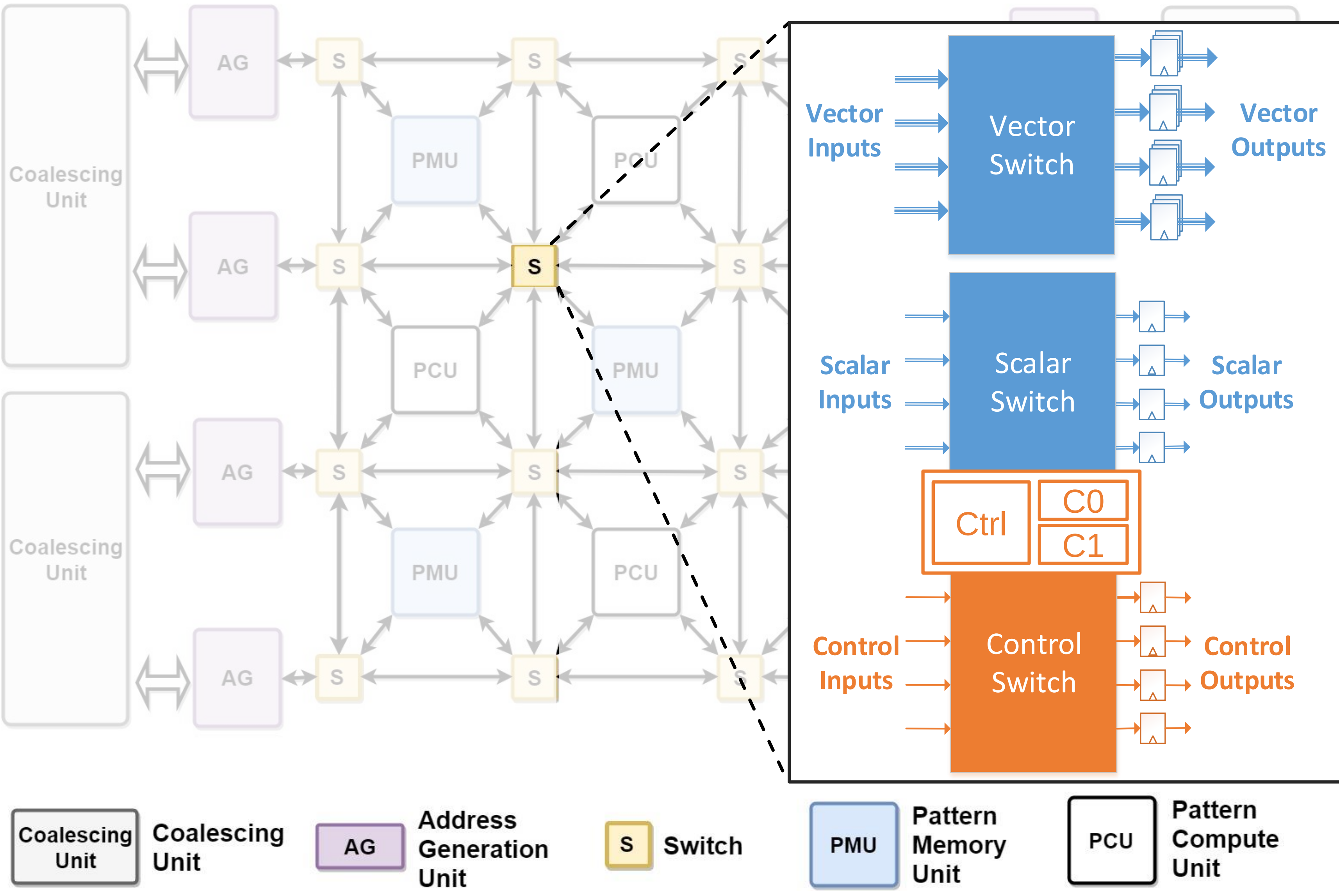
Coalescing Unit	Coalescing Unit
AG	Address Generation Unit
S	Switch
PMU	Pattern Memory Unit
PCU	Pattern Compute Unit

Plasticine: PMU



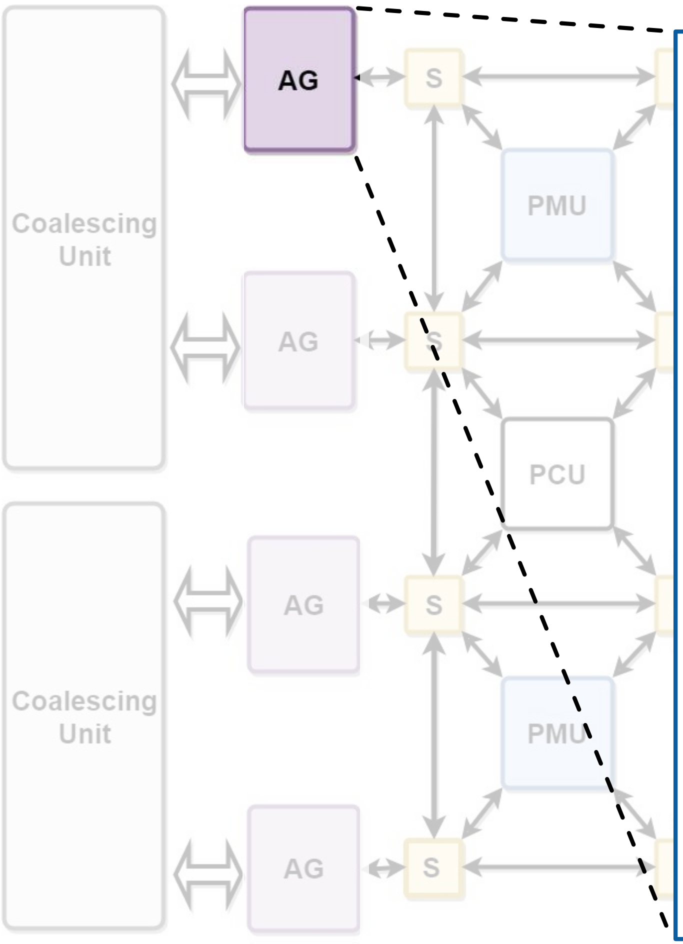
Coalescing Unit	Coalescing Unit	AG	Address Generation Unit	S	Switch	PMU	Pattern Memory Unit	PCU	Pattern Compute Unit

Plasticine: Interconnect

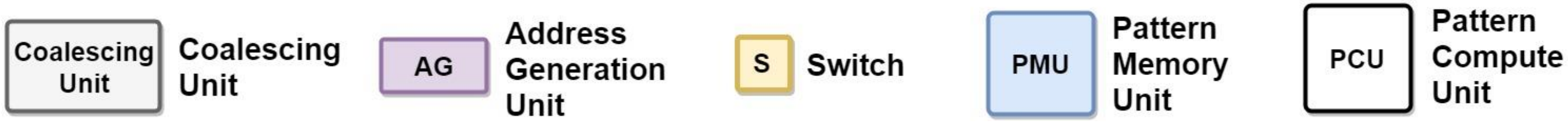


- Three granularities
- Counters, Control within switches

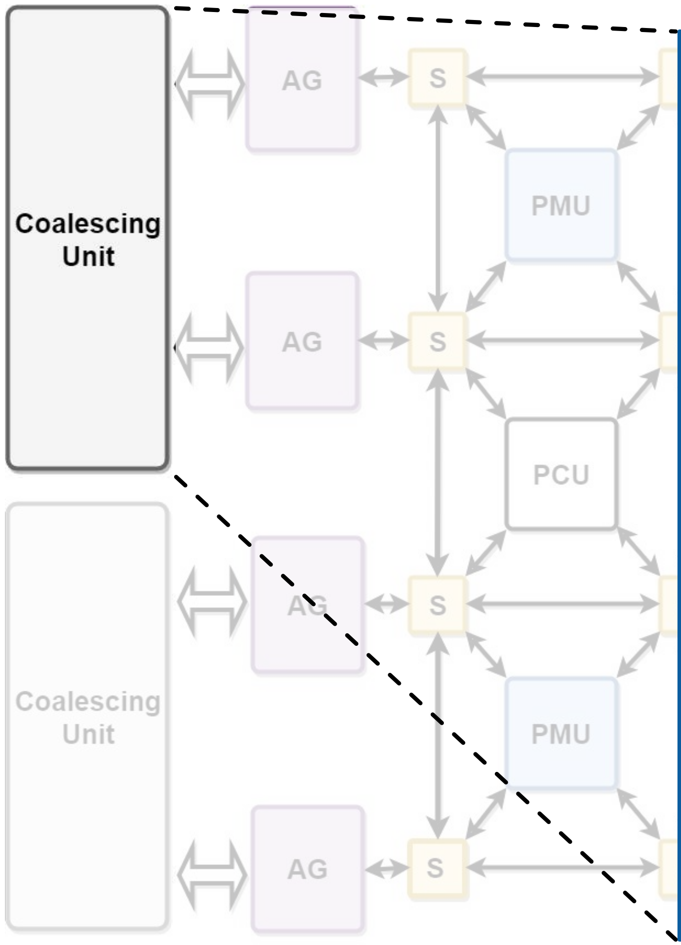
Plasticine: Address Generators



- Integer FUs to generate DRAM requests
- Design similar to PMU.



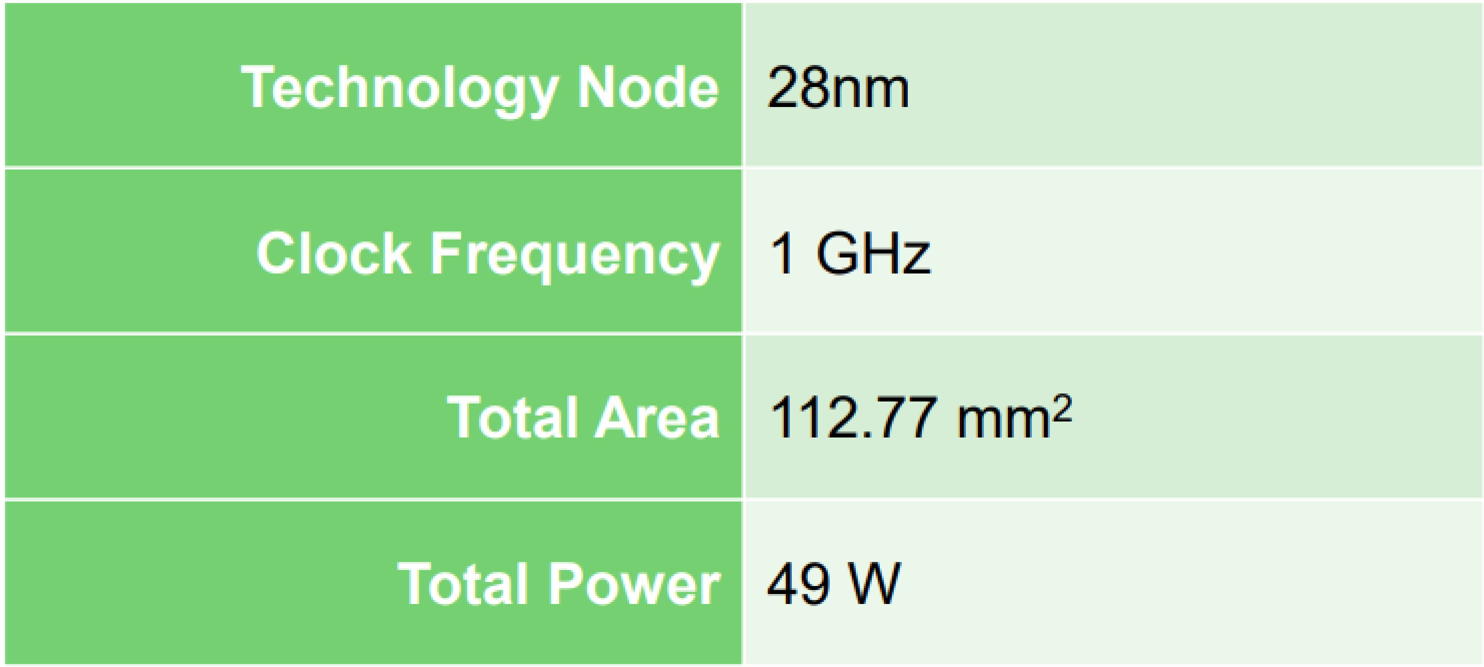
Plasticine: Coalescing Units



- **Consists of buffer**
  - Allows large number of outstanding requests
- **Arbitration between multiple address streams**
  - Shares physical DRAM channel between multiple AGs

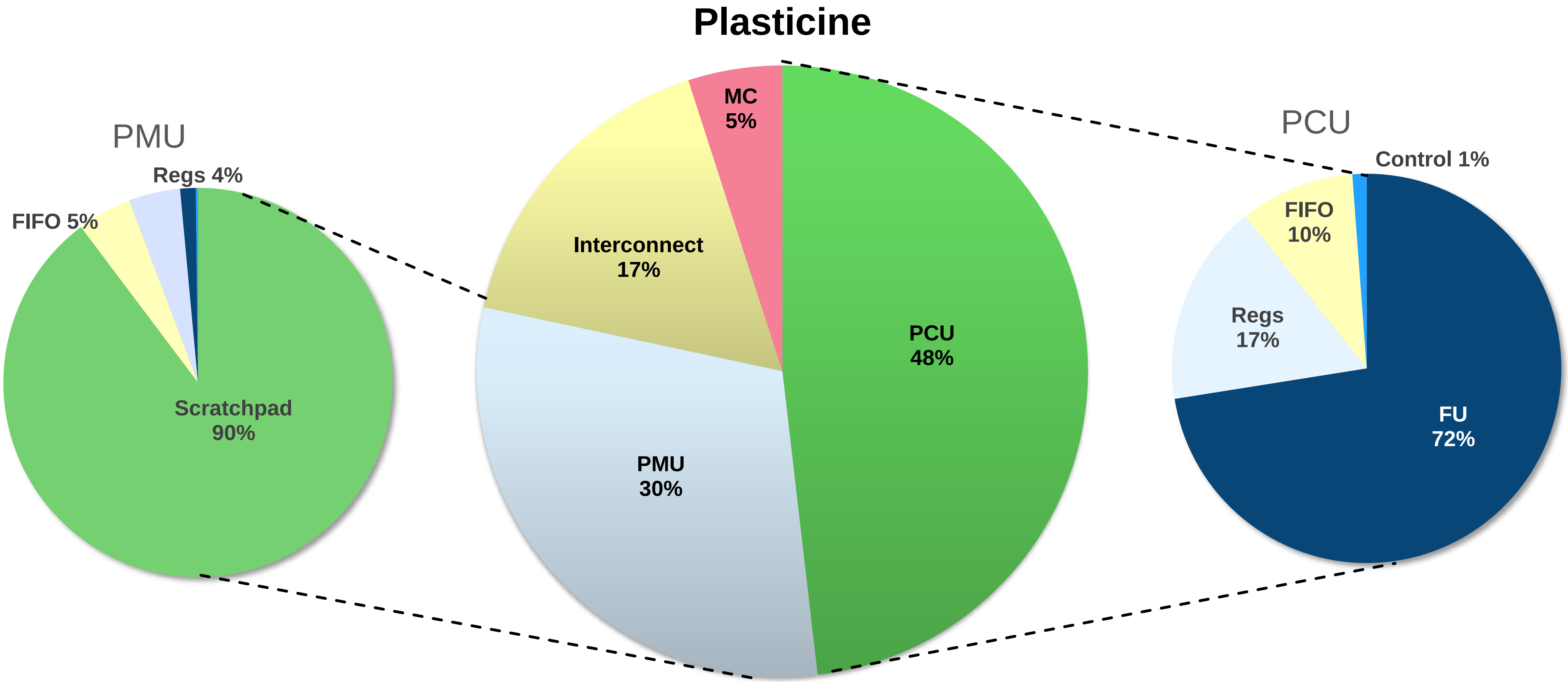
Coalescing Unit
Coalescing Unit
AG
Address Generation Unit
S
Switch
PMU
Pattern Memory Unit
PCU
Pattern Compute Unit

Plasticine Clock, Area, and Power



Technology Node	28nm
Clock Frequency	1 GHz
Total Area	112.77 mm^2
Total Power	49 W

Area Breakdown

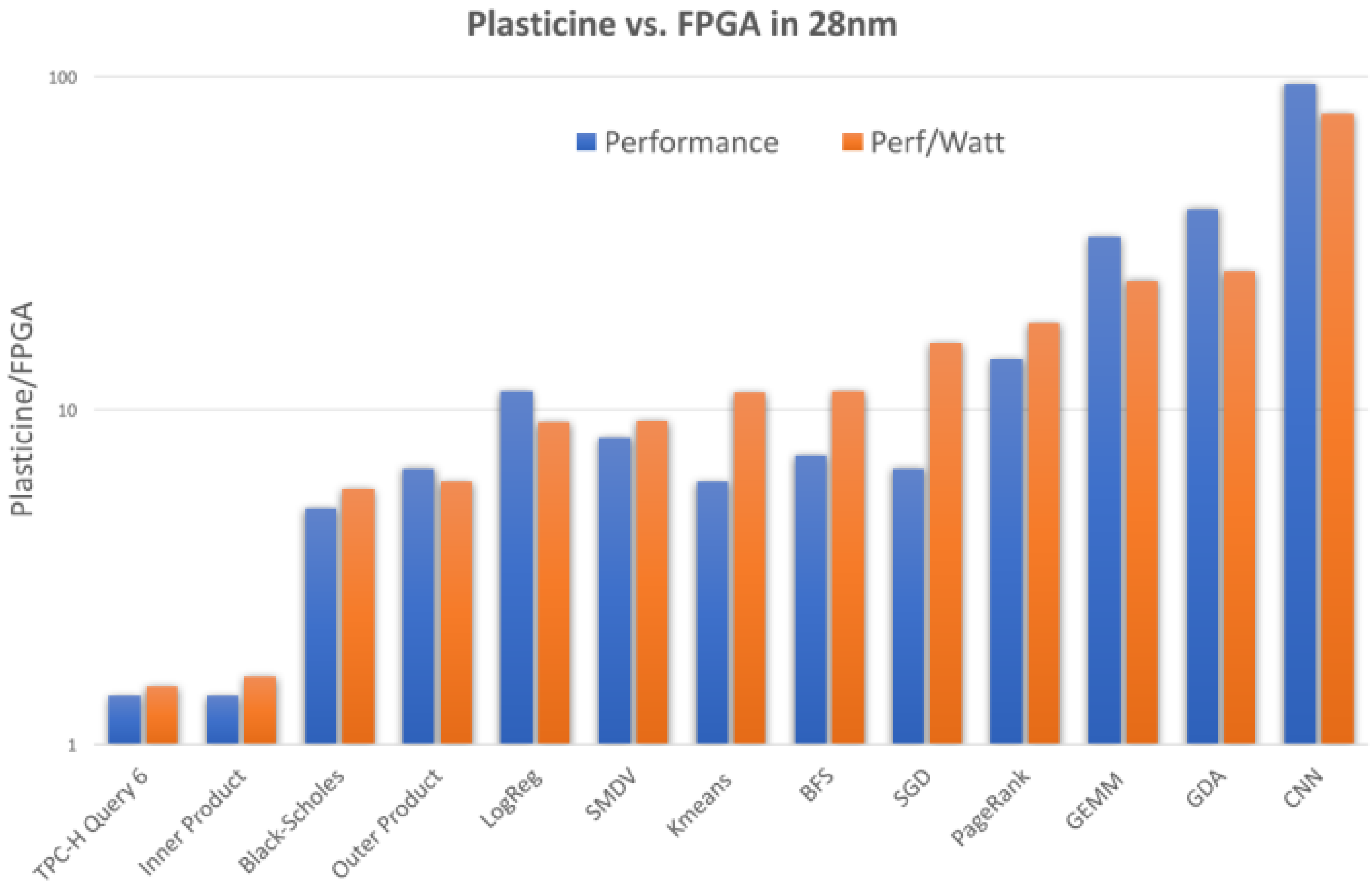


Experimental Setup

- **Plasticine:**
  - Implemented in Chisel, RTL synthesized with 28nm library
  - 4 DDR3-1600 DRAM channels, peak memory bandwidth = 51.2 GB/s
  - 1 GHz clock

- **FPGA:**
  - Altera Stratix V, 28 nm technology
  - 6 DDR3-800 DRAM channels, peak memory bandwidth = 37.5 GB/s
  - 150 MHz clock

Plasticine v/s FPGA



Conclusion

- Co-designing reconfigurable architecture and programming models based on parallel patterns leads to efficient, programmable systems

- Up to **95x** improvement in Performance, **77x** improvement in Perf/W over FPGA in similar process technology, with an area of 113 sq mm.

Acknowledgement

- Some of the slides are adapted from **Raghu Prabhakar**’s talk on “Plasticine: A Reconfigurable Architecture For Parallel Patterns.”

Thank you

Questions?