Increasing Dynamism in Plasticine

Yaqi Zhang
yaqiz@stanford.edu

Alexander Rucker
acrucker@stanford.edu

Matthew Vilim
mvilim@stanford.edu

Stanford University

STANFORD DAWN

Background

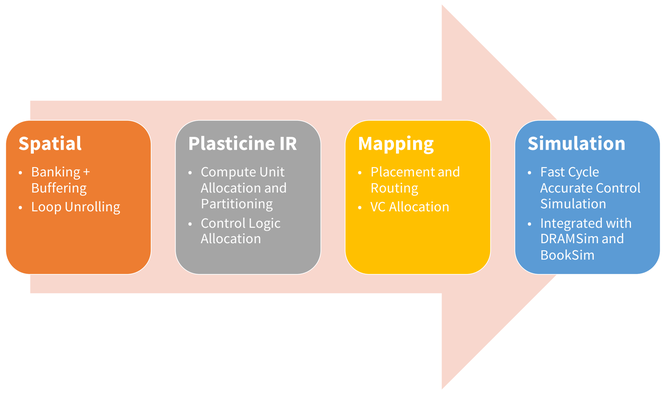
Plasticine is a vector Coarse-Grained Reconfigurable Array:

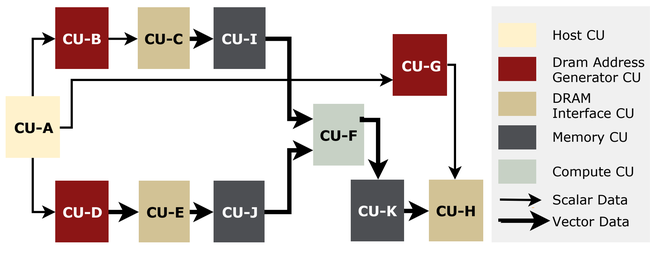
- 6-stage, 16-lane 32-bit floating point SIMD pipelines
- Distributed 256-kByte memories
- Memory controllers support dense and sparse DRAM access

Plasticine demonstrated up to 95x speedup and 77x performance per watt vs. a Stratix V FPGA.

How can we retain Plasticine's performance and efficiency while enabling new applications?

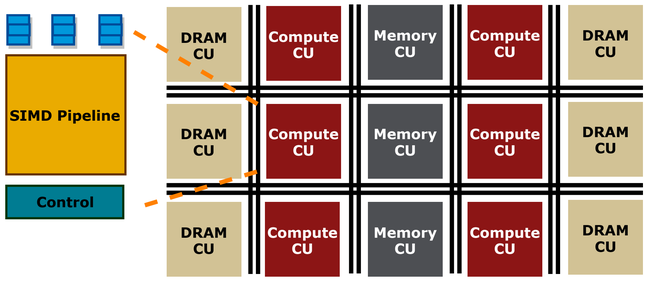
Compiler & Mapping Flow





Virtual Compute Unit Data-Flow Graph

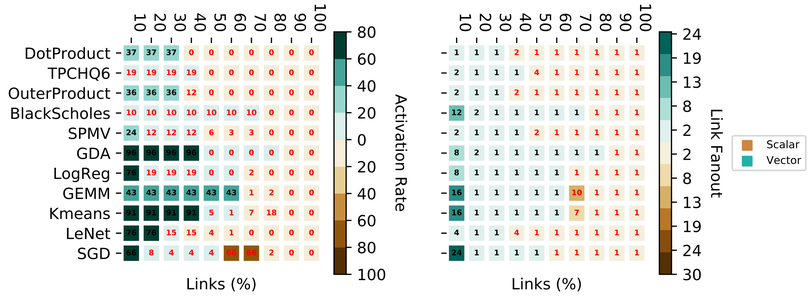
Mapping



Physical Compute Unit

Hybrid Networks

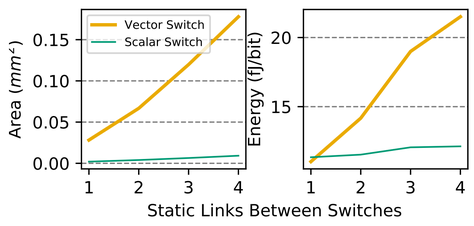
Different applications have different link activation rates and fanouts:



How can we improve link utilization?

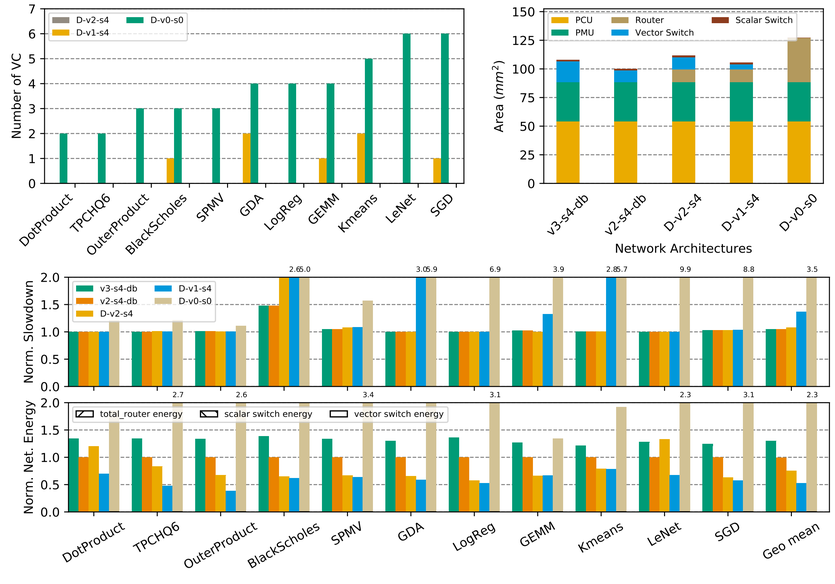
- Use static network for high-bandwidth and broadcast links
- Use dynamic network to encourage link sharing on low-activation links
- Specialize networks at different granularities

Area and energy scaling for switch and router on 28nm technology:



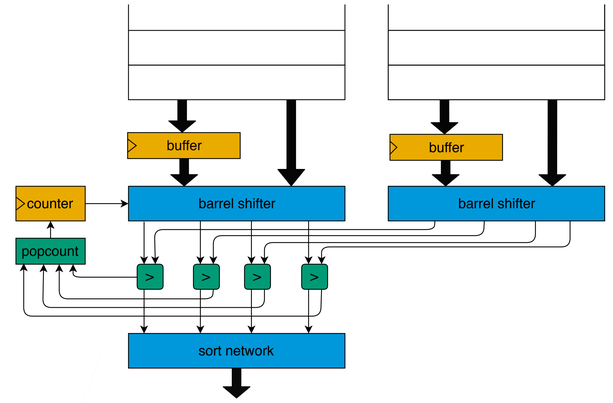
#VC	Buffer Slots	Area (μm^2)	Energy(fJ/bit)
4	2	72426.39	31.19
2	4	70390.01	28.44
4	4	127695.32	36.11
8	4	241833.85	52.47

Network evaluation for different benchmarks:

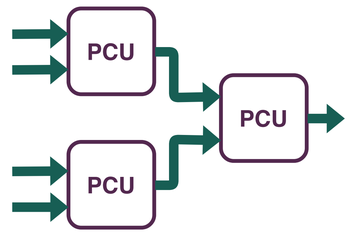


Future Work

Minor hardware additions to the PCU to allow fast sort:



Multi-way merge reduces memory traffic:



What's the next class of applications to target?

- Transactional/online applications?
- Streaming data analytics and networking?
- Graph analytics?

What advances will be necessary to target these applications?

Improve achieved density for generic control constructs:

- Data-dependent conditionals
- Finite state machine-based control
- Parsing support
- Support for more complicated data structures