Increasing Dynamism in Plasticine

Yaqi Zhang
yaqiz@stanford.edu

Alexander Rucker
acrucker@stanford.edu

Matthew Vilim
mvilim@stanford.edu

Stanford University

STANFORD DAWN

Background

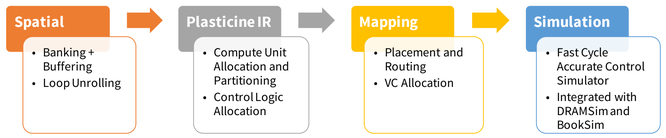
Plasticine a vector Coarse-Grained Reconfigurable Array:

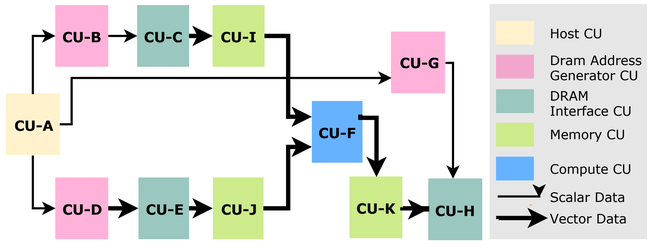
- 6-stage, 16-lane 32-bit floating point SIMD pipelines
- Distributed 256-kByte memories
- DRAM controllers with tile load and scatter-gather support

Plasticine demonstrated up to 95x speedup vs. an FPGA, and 77 times performance per Watt.

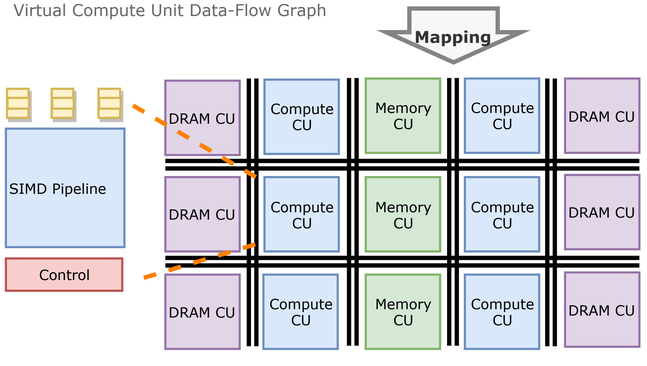
How can we retain Plasticine's performance and efficiency while enabling new applications?

Compiler & Mapping Flow





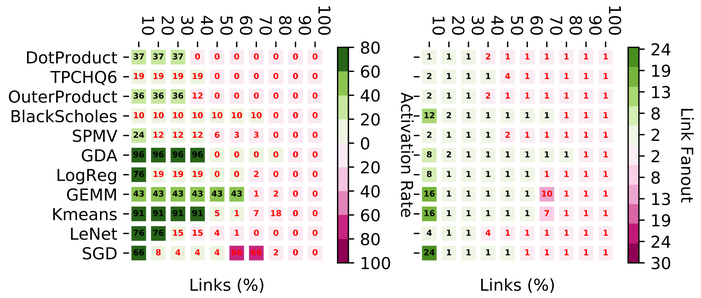
Virtual Compute Unit Data-Flow Graph



Physical Compute Unit

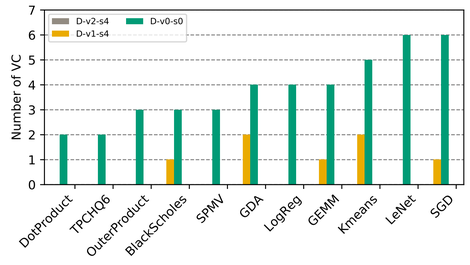
Hybrid Networks

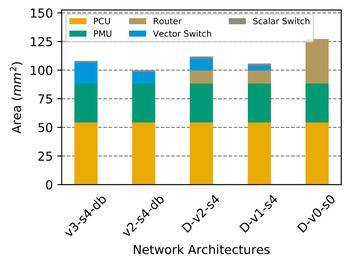
Different applications have different link activation rates and fanouts:

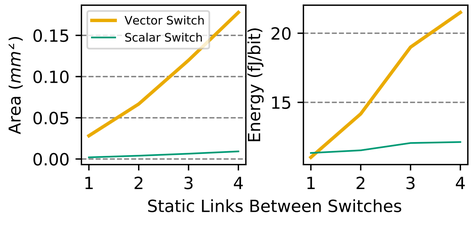


How can we improve link utilization?

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

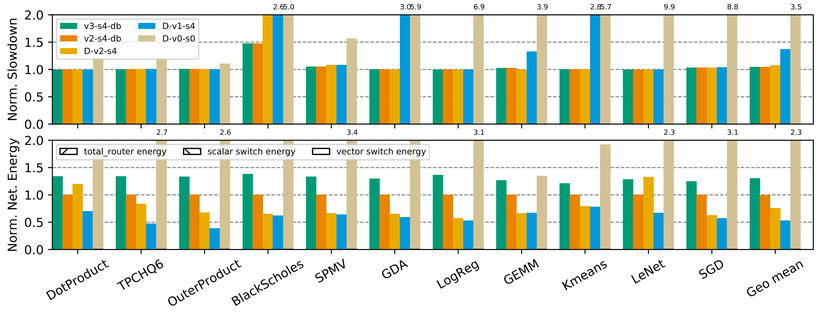






#VC	Buffer Slots	Area (m^2)	Energy(fJ/bit)
4	2	72426.39	31.19
2	4	70390.01	28.44
4	4	127695.32	36.11
8	4	241833.85	52.47

Router Area and Energy Scaling



Future Work

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?

- Enhanced support for data-dependent conditionals
- Support for finite state machine-based control
- Tightly integrated parsing support
- Support for more complicated data structures