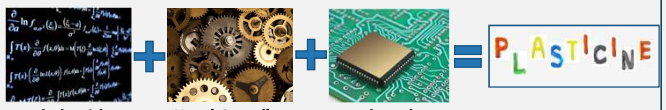
Plasticine: A Universal Data Analytics Accelerator

Oyekunle Olukotun(PI), Chris Re, Christos Kozyrakis (Stanford University)
Ram Sivaramakrishnan (PI), Mark Luttrell, Jun-Uk Luke Shin (SambaNova Systems, Inc.)

Architectures Thrust: Software Defined Hardware (SDH)

Overview



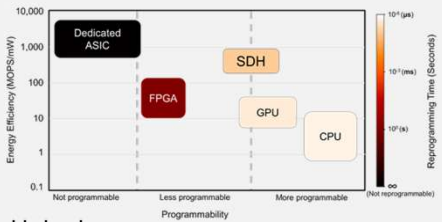
Novel algorithms from the database and machine learning domains designed with Plasticine in mind.

Novel compiler that takes high level applications, performs algorithm and data optimizations, and translates them to efficient Plasticine configurations

Novel hardware that enables reconfiguration without the cost.

Combined via a **unified software layer** targeting multiple domains.

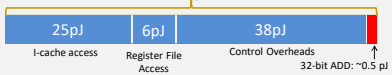
Background



Programmable hardware: Instruction-based processors (CPUs, GPGPUs)

Reconfigurable hardware: Reconfigurable data paths (FPGAs, SDH)

Instructions add overheads:



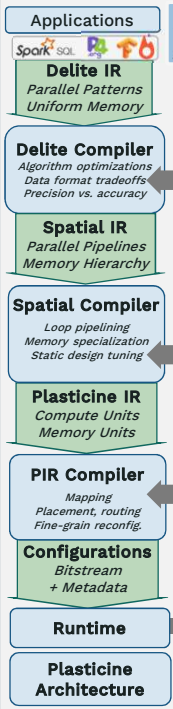
Reconfigurable hardware avoids instruction overheads

Previous Work	Exploit Nested Parallelism	Reconfigurable, distributed scratchpad memories	Hierarchical interconnection	Programmability	Fast Compilation
ADRES	×	×	×	✓	✓
DySER	×	×	×	✓	✓
Garp	×	×	×	✓	✓
PipeRench	×	×	×	✓	✓
Tartain	×	✓	✓	✓	✓
RaPiD	×	✓	×	✓	?
HRL	×	×	×	✓	×
Triggered Instructions	✓	✓	×	✓	×
Mosaic	✓	×	×	✓	?
FPGA	✓	✓	×	✓	×

Previous work within limited application domains:

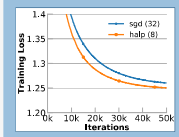
- Deep learning only: GraphCore, TPU
- Database only: Oracle, SAP

Proposed Work



Parallel patterns provide a backbone for the Delite intermediate representation (IR), a hierarchical dataflow graph preserving information across a wide range of domains

Delite algorithm optimizations enable big gains



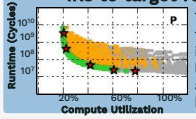
High-accuracy low-precision (on LSTM training)

High-accuracy low-precision training enables benefits of low-precision without a loss in accuracy.

YellowFin is a training tuner that can mitigate the asynchrony introduced from distributed training.

Worst-case optimal joins can outperform classic database joins by orders of magnitude.

Spatial lowers and optimizes high level program IRs to target reconfigurable hardware

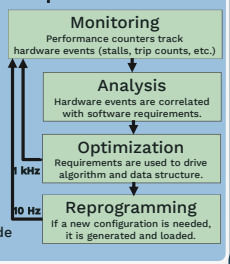


The Spatial IR is designed to be targeted by parallel patterns, but includes low level target information like memory hierarchy to drive automatic tuning to best utilize the hardware's memory and compute.

Rapid reconfiguration enables adaption to changing application requirements

The optimal algorithm and data layout depend on the characteristics of the data. The compiler will automatically rewrite programs without user intervention:

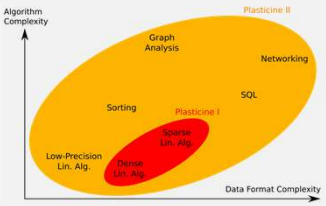
1. The programmer or Delite compiler produces a list of **tunable program features**
2. **Hardware counters** are generated for these features
3. The features are used in a machine learning model to guide **optimization decisions**



Dynamic applications pose unique hardware challenges

Applications have data ranging from 4 bits (low-precision math) to 12,000 bits (networking). This data may be dense or sparse, and control ranges from predictable to complex and data dependent.

Plasticine allows specialization for each of these characteristics with configurable hardware support like dense and sparse memory controllers.



Performance and Energy Efficiency Comparisons (Single Tile)

	Performance			Performance/Watt		
Class	vs. CPU	vs. GPU	vs. ASIC	vs. CPU	vs. GPU	vs. ASIC
Dense (AlexNet)	30.0x	1.7x	78.6x	52x	9.2x	0.5x
Sparse (BFS)	8.2x	0.8x	3.7x	55x	9.7x	1.3x

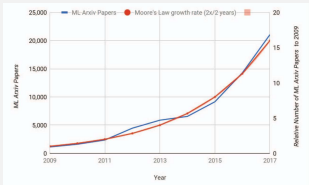
CPU: 2-socket, 18-core Xeon E5-2699 v3 CPU · **GPU:** NVIDIA Titan X
Dense ASIC: Eyeriss · **Sparse ASIC**: Graphicionado

Projections show a **single tile** is within **2x** energy efficiency of a small, efficient CNN ASIC on dense workloads and comparable to a custom graph processing ASIC on sparse workloads. The Plasticine chip will have multiple such tiles and more DRAM channels.

Impact

Enable Next Generation Data Analytics

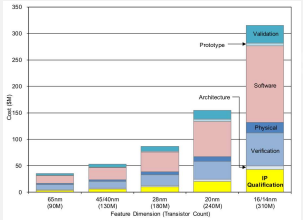
Modern ML and data analytics workloads are constantly changing, have increasing compute and memory requirements, data dependent computation, and variable precision.



Flexibility at runtime and interfacing to large DRAM allows sustaining compute throughput, without sacrificing compute density and energy efficiency.

Reduce Accelerator Development Costs

Software accounts for increasingly larger fractions of chip development costs with technology scaling.



High-level programming models like SQL, TensorFlow, and PyTorch increase designer productivity and allow Plasticine to be used by domain experts with no experience with reconfigurable hardware.

The intermediate **Spatial** compiler enables the use of these DSLs by automatically performing target-specific lowering and optimizations. Spatial's IR is abstract enough to be targeted by parallel patterns, but low enough to allow hardware modeling.

Elastic Cloud Deployment

Plasticine provides a secure mechanism to perform fine-grained resource sharing which allows tailoring the compute and memory requirements for a given application to maximize Performance / $, while minimizing interference and increasing utilization.

Democratize Hardware Acceleration

While developers in different domains use different programming frameworks, the underlying characteristics amenable to hardware acceleration largely remains constant. By supporting frameworks such as Tensorflow, PyTorch, and P4, our approach has the potential to transparently accelerate applications from various application domains.

Contact

Technical Point of Contact:
Professor Kunle Olukotun (kunle@stanford.edu)
Departments of Electrical Engineering and Computer Science, Stanford University

BAA: HR001117S0055

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA).
The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

THE ELECTRONICS RESURGENCE INITIATIVE