The DECADES Project

Margaret Martonosi[1], David Wentzlaff[1], Luca Carloni[2]
[1]Princeton University, [2]Columbia University

Specialized Functions: Software Defined Hardware (SDH)

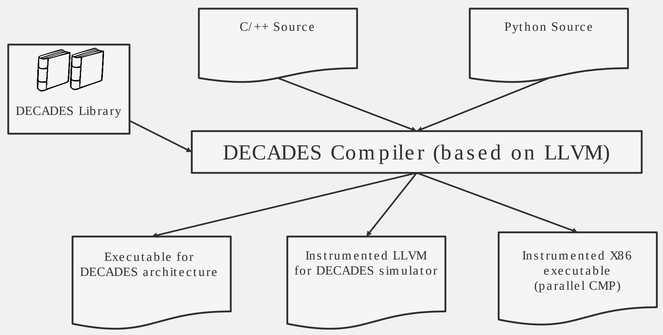
BACKGROUND AND OVERVIEW: THE PROBLEM

- Due to end of Moore/Dennard Era, we face massive challenges in power/performance scaling.
- Particularly vexing for important emerging workloads like Machine Learning and Graph Analytics.
- Compute Accelerators are a partial solution, but:
  1. Need dynamic customization and flexibility.
  2. Need solutions to memory bottlenecks.

SOLUTION: THE DECADES APPROACH

- APIs and language constructs for decoupled data supply, spatial placement, and multi-granular memory access.
- Static and dynamic compiler adaptations and automated parallelism.
- Configurable tile-based hardware: Cores, Accelerators, Intelligent Storage.
- Scalable full-system simulator and FPGA-based emulator, as well as x86 targets.
- Full-chip fabrication with HW/SW full-system measurements.
- **Technology Transfer** : Team has extensive experience in Open-Source SW and HW ecosystem, and licensing of configurable HW tiles.

DECADES COMPILER



Overview

Two front ends: C/++ through Clang and Python through Numba.
LLVM is linked against DECADES runtime and library.
DECADES transformations: DOALL and Decoupled Parallelism.
Compiled down to: RISC-V executables, DECADES Simulator, X86 executable.

Decoupled Supply/Compute

Automatic program slicing into a memory access slice and a compute slice.
Terminal loads are identified and annotated for additional analysis and optimization.

Intelligent Storage

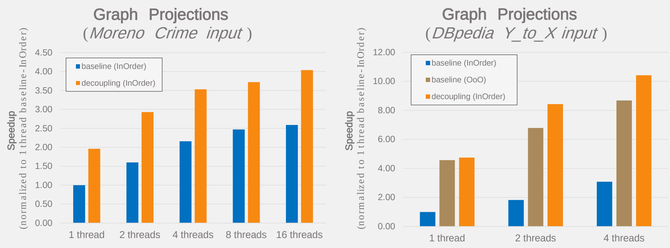
Semi-manual slicing of program into compute core slice and intelligent storage (IS) slice.
Used for situations when automatic slicing is not sufficient (e.g. sparse to dense transformations).

SPMD Parallelism

Kernels are supplied with a native tile id which can be used to write SPMD programs.
Can be used in tandem with either of the above (i.e. multiple compute tiles and supply/IS tiles).

DECADES HARDWARE

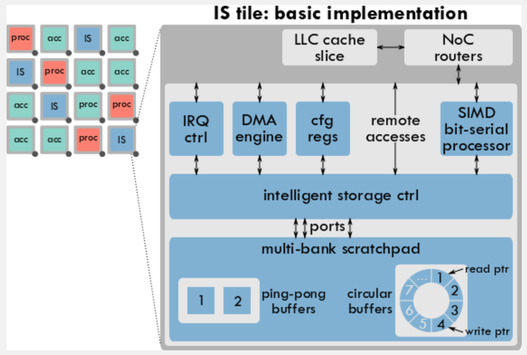
Decoupled Memory Supply



Decoupled Memory Features :

- Separate memory fetch unit for lookahead and prefetch.
- Orthogonal to DOALL Parallelism.
- 4X-10X speedup from memory latency tolerance.
- Outperforms OoO designs for memory-bound applications.

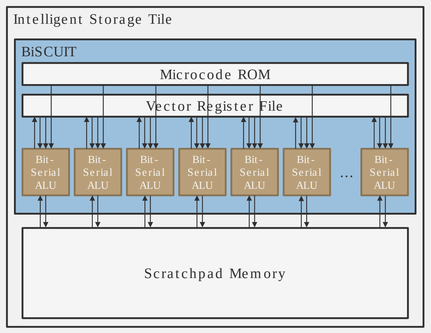
Intelligent Storage Tile



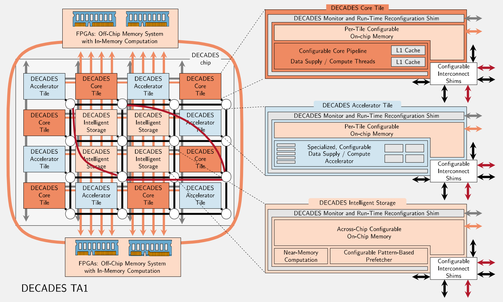
Intelligent Storage Tile Features :

- Shared scratchpad for processors and accelerators.
- Explicit data movement to control data reuse and memory bandwidth usage, while leveraging the DECADES network-on-chip.
- Runtime configurability to match the needs of a given application.
- Near-scratchpad computation.

BiSCUIT Bit-Serial Compute

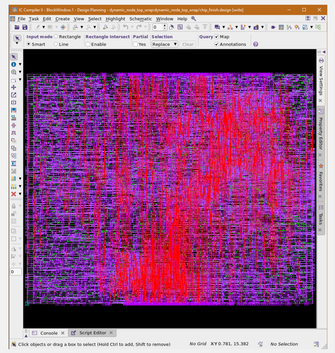


DECADES PLATFORM ARCHITECTURE



DECADES TA1

PLACED AND ROUTED TEST CHIP



RESULTS

* Optimized = DECADES algorithmic improvements + compiler innovations

Graph Projections

Innovations

- explicit edge list.
- intra-node parallelism.
- Decoupled supply/compute identifies long latency loads and issues them in parallel.
- Handles large real-world graphs e.g. YouTube.

Results

- Sequential baseline is **8x** faster than SDH baseline.
- Scaling up to 16 threads yields **37x** speedup.

Vertex Nomination

Innovations

- Single global path computation.
- Used subset of Python compilable to a native executable with Numba.
- Python to DECADES runtime through Numba.
- Parallelism available in global graph traversal
- Compiler-aware modular data-structures (e.g. graph).

Results

- Up to 10,000x speedup over SDH baseline.

Funding and Acknowledgements
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.

mrm@Princeton.edu **wentzlaf@Princeton.edu** **luca@cs.Columbia.edu**

THE ELECTRONICS RESURGENCE INITIATIVE