THE ELECTRONICS RESURGENCE INITIATIVE

DECADES:

DEEPLY-CUSTOMIZED ACCELERATOR-ORIENTED DATA SUPPLY SYSTEMS SYNTHESIS

LUCA CARLONI

DEPARTMENT OF COMPUTER SCIENCE
COLUMBIA UNIVERSITY

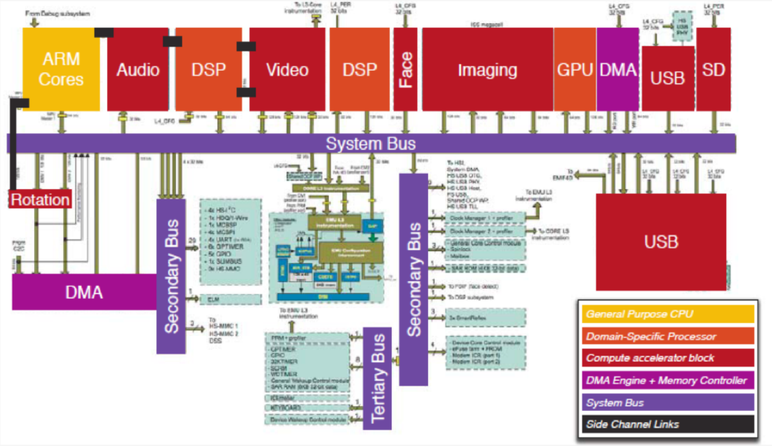
THE DATA SUPPLY CHALLENGE

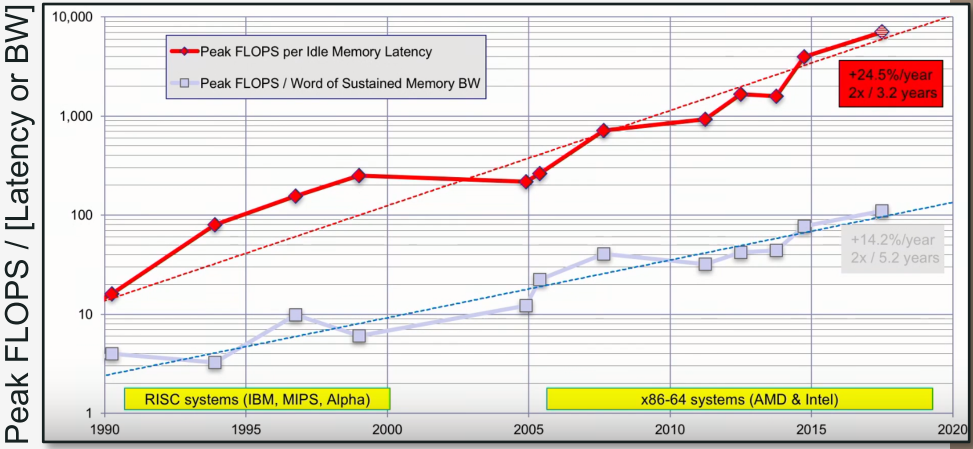
Modern computer systems are increasingly heterogeneous

- Accelerator-oriented parallelism to meet aggressive performance and power targets

As accelerators have sped up compute portions, the main challenge is data supply

- Key bottlenecks lie in memory and communication overheads associated with supplying specialized accelerators with data
- Different apps have distinct data supply needs





John McCalpin, SC’16 Keynote

DECADES: A VERTICALLY-INTEGRATED APPROACH

Language and Compiler Support

- Enhance data locality
- Optimize spatial mapping of threads
- Enable in-memory computing

Very Coarse-Grained Reconfigurable Tile-Based Architecture

- Coarser than CGRA → VCGRTA
- 3 classes of reconfigurable tiles
- Reconfigurable interconnection network
- Reconfigurable in-memory computing

Multi-Tiered Demonstration Strategy

- Scalable full-system simulation
- Multi-FPGA emulation infrastructure
- 225-tile DECADES chip prototype

DECADES: A VERTICALLY-INTEGRATED APPROACH

Language and Compiler Support

(M. Martonosi)

- Enhance data locality
- Optimize spatial mapping of threads
- Enable in-memory computing

Very Coarse-Grained Reconfigurable Tile-Based Architecture

(L. Carloni)

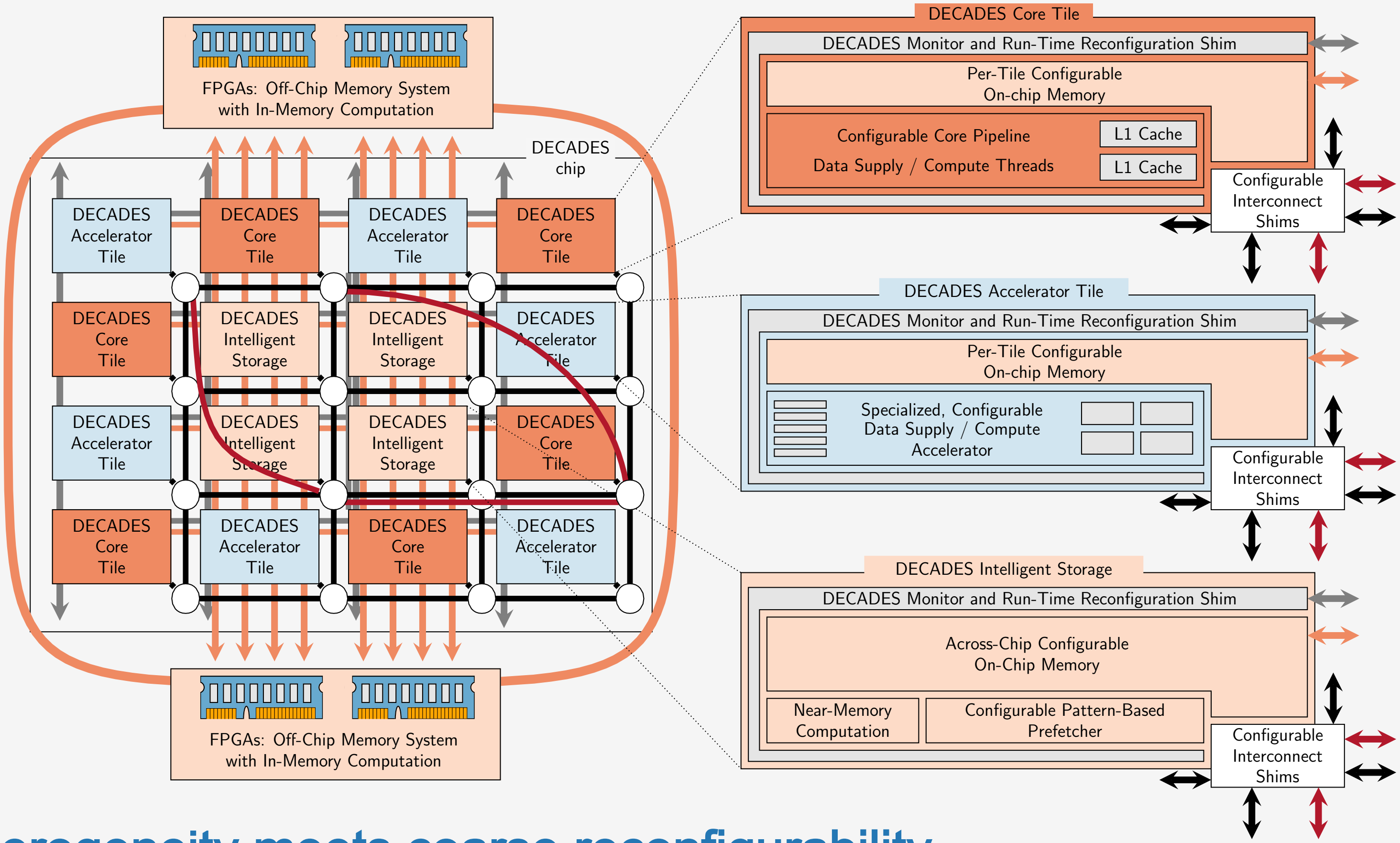
- Coarser than CGRA → VCGRTA
- 3 classes of reconfigurable tiles
- Reconfigurable interconnection network
- Reconfigurable in-memory computing

Multi-Tiered Demonstration Strategy

(D. Wentzlaff)

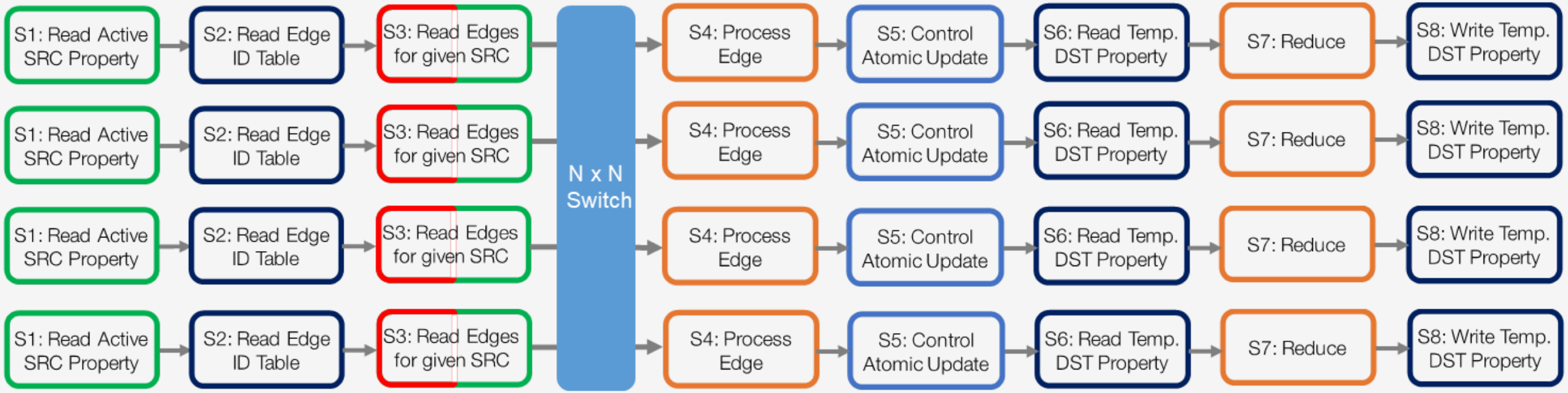
- Scalable full-system simulation
- Multi-FPGA emulation infrastructure
- 225-tile DECADES chip prototype

DECADES PLATFORM ARCHITECTURE


FPGAs: Off-Chip Memory System
with In-Memory Computation
DECADES
chip
DECADES Accelerator Tile
DECADES Core Tile
DECADES Accelerator Tile
DECADES Core Tile
DECADES Core Tile
DECADES Intelligent Storage
DECADES Intelligent Storage
DECADES Accelerator Tile
DECADES Accelerator Tile
DECADES Intelligent Storage
DECADES Intelligent Storage
DECADES Core Tile
DECADES Core Tile
DECADES Accelerator Tile
DECADES Core Tile
DECADES Accelerator Tile
FPGAs: Off-Chip Memory System
with In-Memory Computation
DECADES Core Tile
DECADES Monitor and Run-Time Reconfiguration Shim
Per-Tile Configurable
On-chip Memory
Configurable Core Pipeline
Data Supply / Compute Threads
L1 Cache
L1 Cache
Configurable Interconnect Shims
DECADES Accelerator Tile
DECADES Monitor and Run-Time Reconfiguration Shim
Per-Tile Configurable
On-chip Memory
Specialized, Configurable
Data Supply / Compute
Accelerator
Configurable Interconnect Shims
DECADES Intelligent Storage
DECADES Monitor and Run-Time Reconfiguration Shim
Across-Chip Configurable
On-Chip Memory
Near-Memory
Computation
Configurable Pattern-Based
Prefetcher
Configurable Interconnect Shims

Heterogeneity meets coarse reconfigurability

PRIOR WORK: GRAPHICIONADO



S1: Read Active SRC Property → S2: Read Edge ID Table → S3: Read Edges for given SRC → N x N Switch → S4: Process Edge → S5: Control Atomic Update → S6: Read Temp. DST Property → S7: Reduce → S8: Write Temp. DST Property

S1: Read Active SRC Property → S2: Read Edge ID Table → S3: Read Edges for given SRC → N x N Switch → S4: Process Edge → S5: Control Atomic Update → S6: Read Temp. DST Property → S7: Reduce → S8: Write Temp. DST Property

S1: Read Active SRC Property → S2: Read Edge ID Table → S3: Read Edges for given SRC → N x N Switch → S4: Process Edge → S5: Control Atomic Update → S6: Read Temp. DST Property → S7: Reduce → S8: Write Temp. DST Property

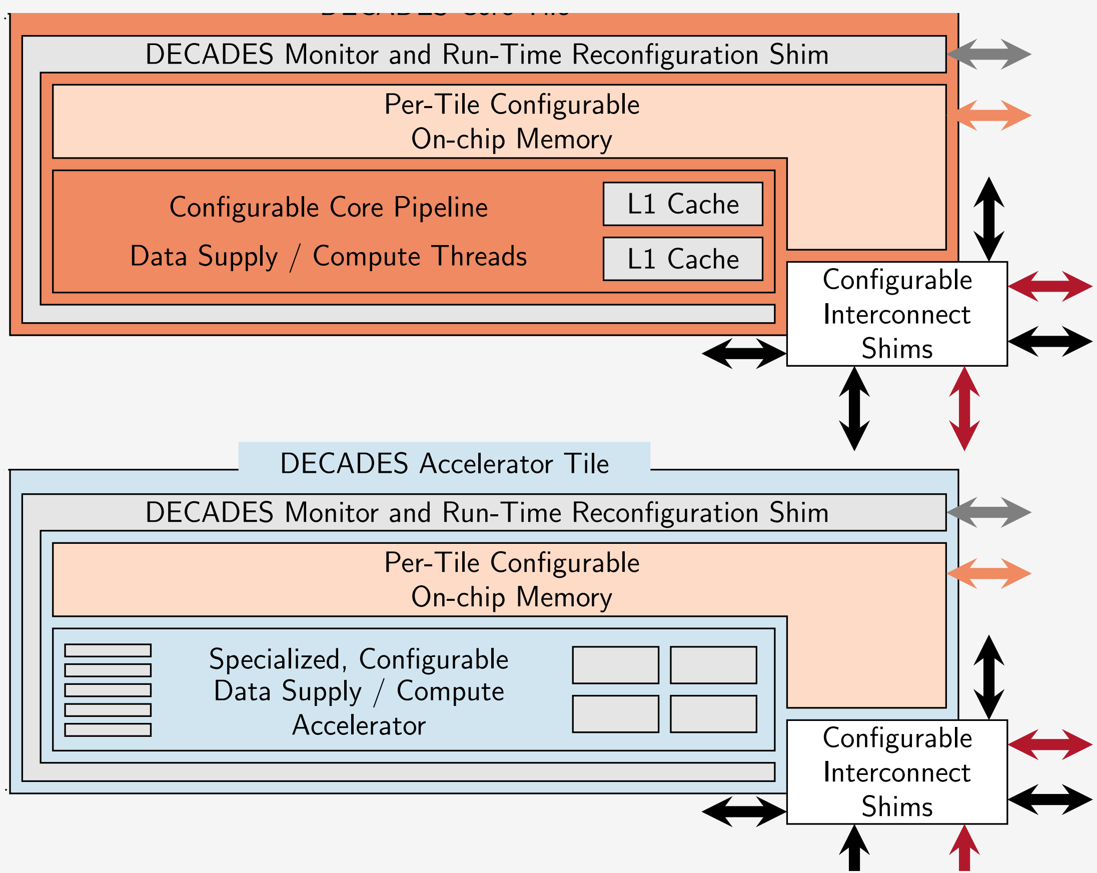
S1: Read Active SRC Property → S2: Read Edge ID Table → S3: Read Edges for given SRC → N x N Switch → S4: Process Edge → S5: Control Atomic Update → S6: Read Temp. DST Property → S7: Reduce → S8: Write Temp. DST Property

- Application-Specific Memory Hierarchy for Bandwidth-Bound Graph Analytics
  - Customized memory hierarchy to minimize off-chip memory access traffic [3x reduction]
  - Ease the design/use of accelerators
  - Dataflow pipeline based on high-level abstraction eases the programming and enables hardware reuse for different graph applications
  - Specialized HW accelerator for graph analytics successfully achieves ~3x speedup and 50x+ energy saving compared to state-of-the-art software framework on 32-core CPU

[Ham et al., MICRO-49, 2016. IEEE Micro Top Picks Honorable Mention]

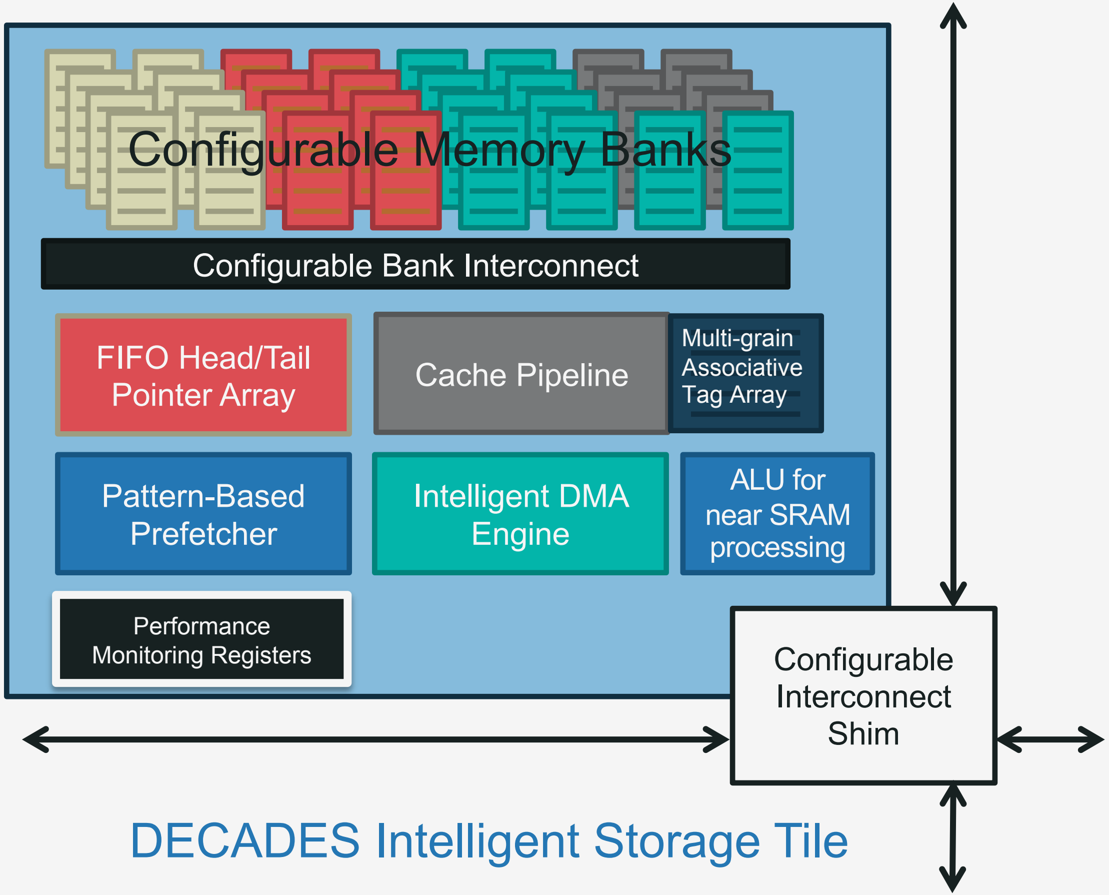
DECADES CORE AND ACCELERATOR TILES

- Computations mapped onto core tiles or available accelerator tiles
  - Each tile is wrapped in monitor/reconfiguration shim
- Dynamic reconfiguration of Supply-Compute decoupling, power-performance tradeoffs, and interconnect


DECADES Core Tile
DECADES Monitor and Run-Time Reconfiguration Shim
Per-Tile Configurable On-chip Memory
Configurable Core Pipeline
Data Supply / Compute Threads
L1 Cache
L1 Cache
Configurable Interconnect Shims
DECADES Accelerator Tile
DECADES Monitor and Run-Time Reconfiguration Shim
Per-Tile Configurable On-chip Memory
Specialized, Configurable Data Supply / Compute Accelerator
Configurable Interconnect Shims

INTELLIGENT DATA MANAGEMENT

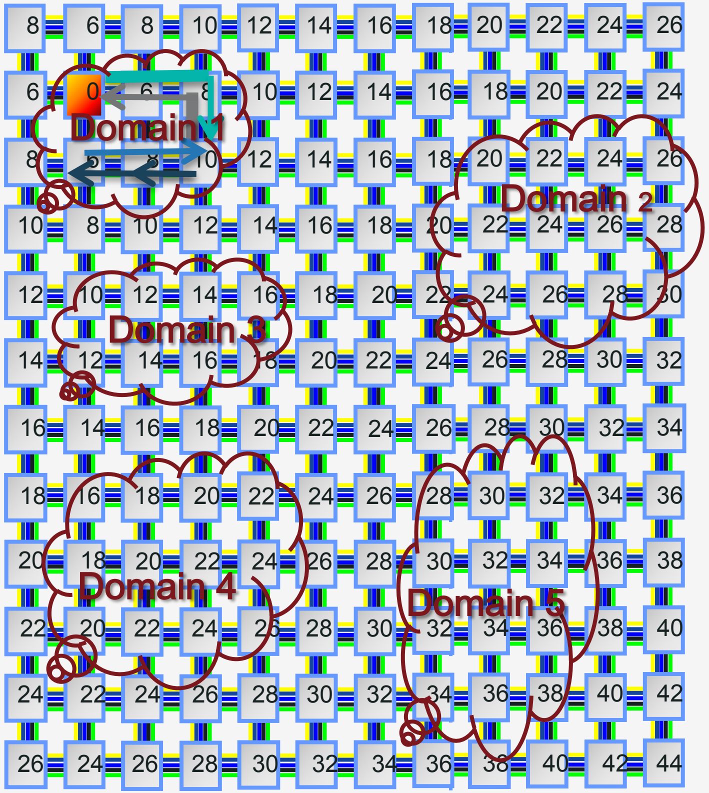
- Specialization #1: Map apps onto mix of compute tiles and intelligent storage (IS) tiles
- Specialization #2: Select and configure appropriate storage features within IS
  - Configurable memory banks + address and prefetching features
  - Simple near-SRAM ALU


Configurable Memory Banks
Configurable Bank Interconnect
FIFO Head/Tail Pointer Array
Cache Pipeline
Multi-grain Associative Tag Array
Pattern-Based Prefetcher
Intelligent DMA Engine
ALU for near SRAM processing
Performance Monitoring Registers
Configurable Interconnect Shim

DECADES Intelligent Storage Tile

PRIOR WORK: COHERENCE DOMAIN RESTRICTION FLEXIBLE MEMORY

- Flexible memory system on top of cache coherent system
  - Enables the exact minimal communication needed
  - Build incoherent coherent domains
- Restriction on application- or page-level
  - Improves performance
    - Shorter network on-chip distances
    - Less interfering memory coherence traffic
  - Reduces energy
    - Fewer on-chip network links need to be transited
    - Less area dedicated to tracking cache line sharers
  - Reduces area
    - Track fewer sharers on large configurations

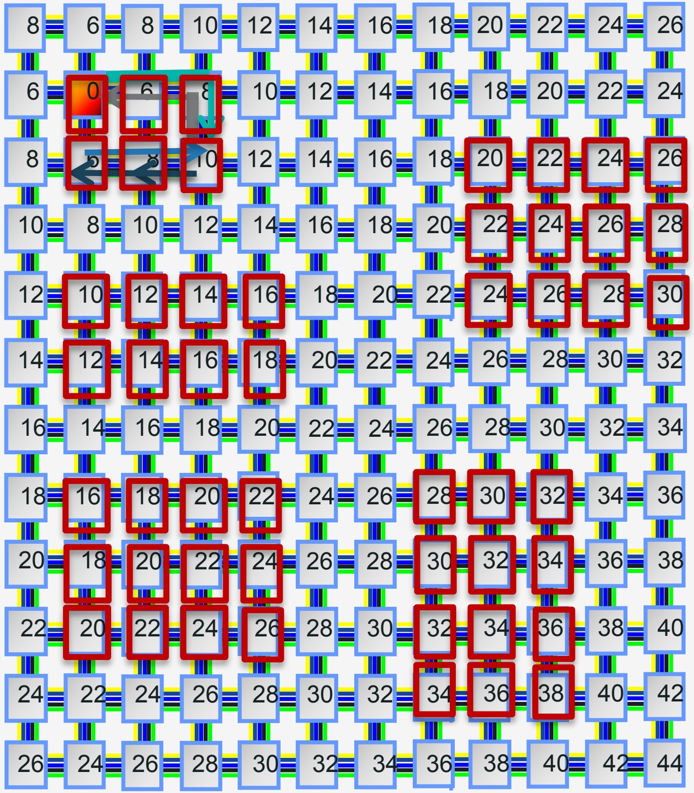


8	6	8	10	12	14	16	18	20	22	24	26
6	0	6	8	10	12	14	16	18	20	22	24
8	6	8	10	12	14	16	18	20	22	24	26
10	8	10	12	14	16	18	20	22	24	26	28
12	10	12	14	16	18	20	22	24	26	28	30
14	12	14	16	18	20	22	24	26	28	30	32
16	14	16	18	20	22	24	26	28	30	32	34
18	16	18	20	22	24	26	28	30	32	34	36
20	18	20	22	24	26	28	30	32	34	36	38
22	20	22	24	26	28	30	32	34	36	38	40
24	22	24	26	28	30	32	34	36	38	40	42
26	24	26	28	30	32	34	36	38	40	42	44

[Fu et al, MICRO 2015]

PRIOR WORK: COHERENCE DOMAIN RESTRICTION FLEXIBLE MEMORY

- Flexible memory system on top of cache coherent system
  - Enables the exact minimal communication needed
  - Build incoherent coherent domains
- Restriction on application- or page-level
  - Improves performance
    - Shorter network on-chip distances
    - Less interfering memory coherence traffic
  - Reduces energy
    - Fewer on-chip network links need to be transited
    - Less area dedicated to tracking cache line sharers
  - Reduces area
    - Track fewer sharers on large configurations

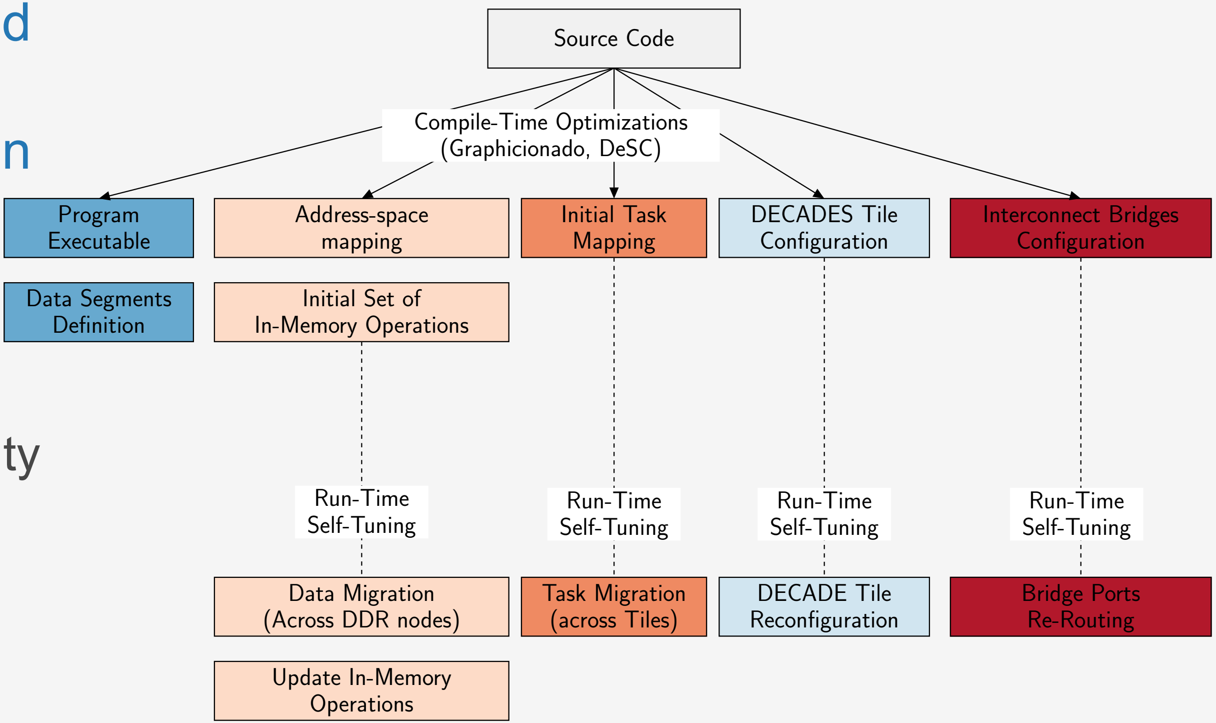


8	6	8	10	12	14	16	18	20	22	24	26
6	0	6	8	10	12	14	16	18	20	22	24
8	6	8	10	12	14	16	18	20	22	24	26
10	8	10	12	14	16	18	20	22	24	26	28
12	10	12	14	16	18	20	22	24	26	28	30
14	12	14	16	18	20	22	24	26	28	30	32
16	14	16	18	20	22	24	26	28	30	32	34
18	16	18	20	22	24	26	28	30	32	34	36
20	18	20	22	24	26	28	30	32	34	36	38
22	20	22	24	26	28	30	32	34	36	38	40
24	22	24	26	28	30	32	34	36	38	40	42
26	24	26	28	30	32	34	36	38	40	42	44

[Fu et al, MICRO 2015]

LANGUAGE, COMPILER & RUNTIME SYSTEM

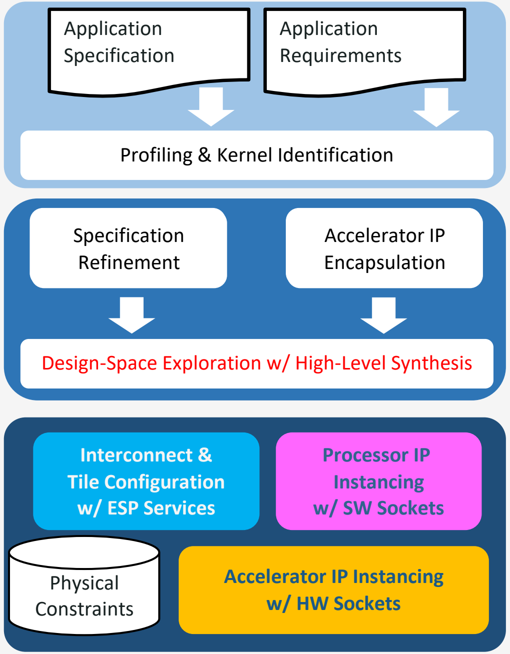
- Compiler Analysis and Support for memory hierarch specialization
  - Bandwidth optimizations through cache optimizations and locality/granularity tailoring
  - Latency tolerance through decoupling
- Build on DeSC LLVM compiler infrastructure


Source Code
Compile-Time Optimizations (Graphicionado, DeSC)
Program Executable
Data Segments Definition
Address-space mapping
Initial Set of In-Memory Operations
Initial Task Mapping
DECADES Tile Configuration
Interconnect Bridges Configuration
Run-Time Self-Tuning
Run-Time Self-Tuning
Run-Time Self-Tuning
Run-Time Self-Tuning
Data Migration (Across DDR nodes)
Update In-Memory Operations
Task Migration (across Tiles)
DECADE Tile Reconfiguration
Bridge Ports Re-Routing

[Ham/Aragon/Martonosi, MICRO-48, 2015]

PRIOR WORK: EMBEDDED SCALABLE PLATFORMS

- Flexible Tile-Based Architecture
- System-Level Design Methodology



Application Specification

Application Requirements

Profiling & Kernel Identification

Specification Refinement

Accelerator IP Encapsulation

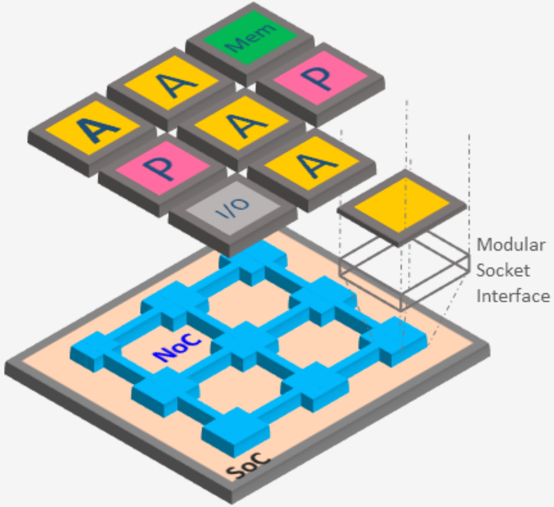
Design-Space Exploration w/ High-Level Synthesis

Interconnect & Tile Configuration w/ ESP Services

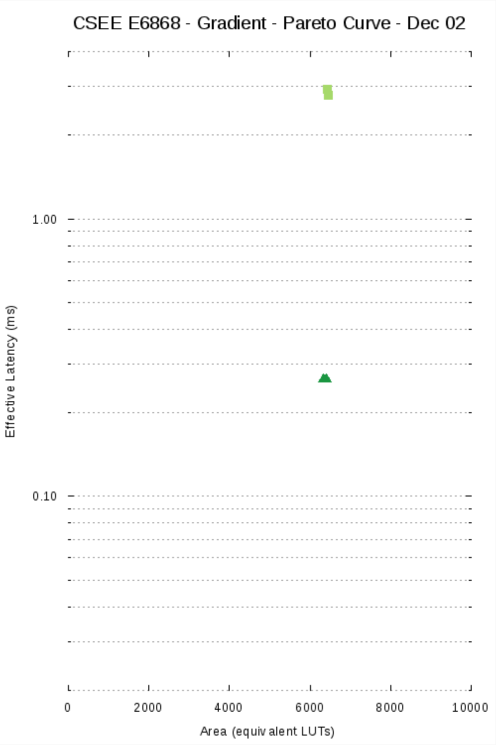
Processor IP Instancing w/ SW Sockets

Physical Constraints

Accelerator IP Instancing w/ HW Sockets



- SoC Design Productivity



CSEE E6868 - Gradient - Pareto Curve - Dec 02

In the span of 1 month:

21: student teams

661: improved designs

32: avg. number of improved designs per team

1.5: avg. number of new designs per team/day

99: Pareto curve changes

11: final number of Pareto-optimal designs

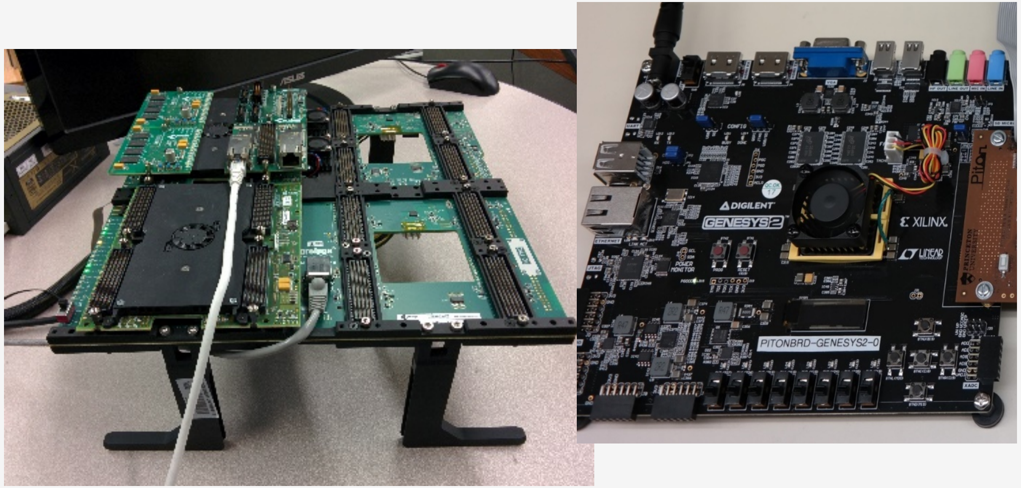
26x: Performance range

10x: Area range

[Carloni, DAC 2016]

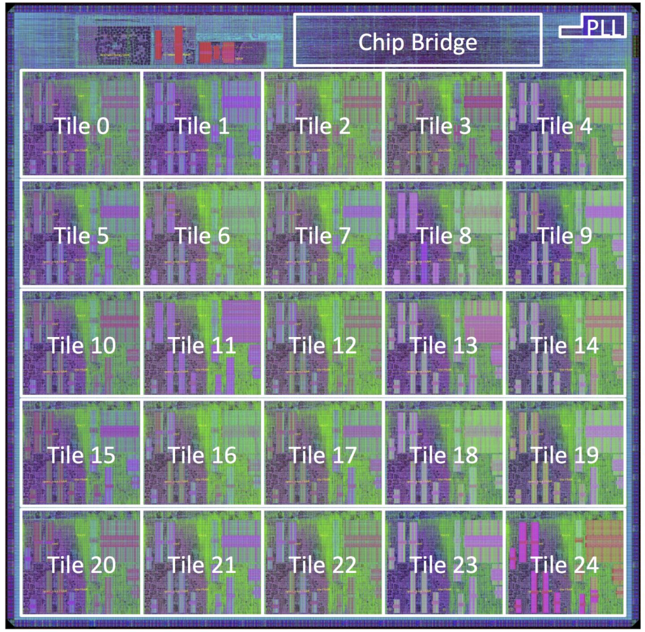
EMULATION & PROTOTYPING

- Take DECADES architecture to FPGA
- Continued design refinement throughout program



- Multi-FPGA emulation infrastructure

- Prototype chip to de-risk architecture



- Recent 25-core manycore system built by our team

SUMMARY & IMPACT

Language/Compiler/Runtime:

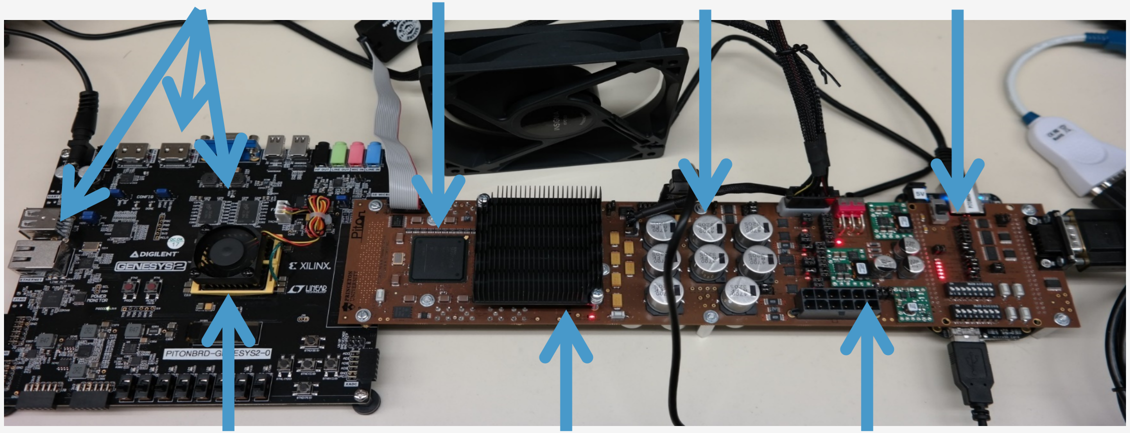
- Latency: >4X per thread performance benefits from memory data supply decoupling
- Bandwidth: Granularity management and Multiplicative outer-loop parallelism up to bandwidth limit
- Total of 50X over single-thread from software

Configurable Hardware Platform:

- Hardware speedups from accelerators for address calculation, memory fetch, or compute
- Fine-grained, low-overhead measurements drive adaptation and module depowering
- 10-20X multiplicative power/performance benefits

TECHNOLOGY TRANSFER PLANS

- Outputs:
  - Software ecosystem
  - Chip design
  - FPGA emulation system
- Technology transfer plans:
  - Release of software, hardware, and data where possible
  - Commercialization and licensing



DRAM + I/O

Bridge FPGA Spartan 6

Bulk Decoupling

Misc. Configuration

Chipset FPGA Kintex 7

Test Chip + Heat Sink

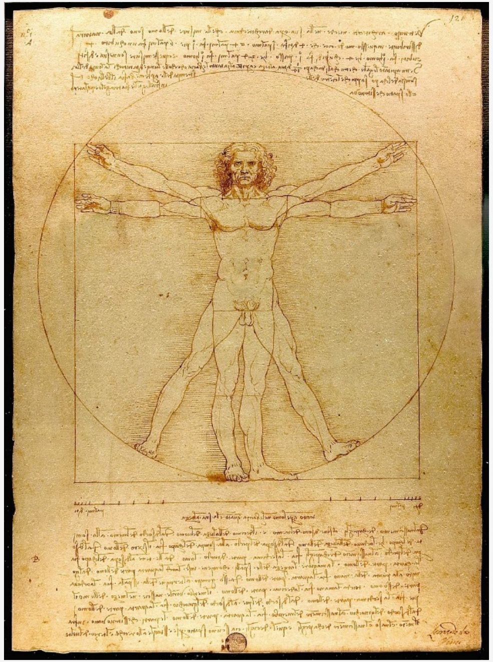
Power Supply

- Leverage extensive past experience:
  - Widely-used open-source software (Wattch, *Check tools, scalable QEMU)
  - Patents licensed to major companies (Power-efficient ALUs)
  - Technology transferred from academia to startups (Tilera)
  - Open-Source Hardware (OpenPiton)

TOWARDS A COMPUTER DESIGN RENAISSANCE

- The end of silicon dimensional scaling and the rise of heterogeneous reconfigurable computing bring an opportunity for a Computer Design Renaissance

- …by supporting the creativity of application developers to realize innovative architectures,
  chips,
  systems
  and products



- … through richly reconfigurable substrates and intelligent compilation and mapping

Microsystems Technology Office

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | JULY 23-25