A 4x4 CGRA with Local Register File and 2-D Mesh Interconnect

Coarse Grained Reconfigurable Arrays (CGRAs)

- Array of Processing Elements (PEs); each PE has ALU-like functional unit that works on an operation every cycle.
- Power-efficiency of several 10s of GOPS/Sec per Watt!
  - ADRES [HiPEAC '08]
  - HyCUBE [DAC '17]

Performance Impact of Ad-Hoc Routing Strategies

- Iterative Modulo Scheduling
  - [B. Rau, MICRO '94]
- Software Pipelined Execution

Mapping Loops on CGRAs

- B = 0;
- for(i=0; i<1000; i++)
- a = B - 4;
- b = A + L;
- c = A + 3;
- d = C + 7;
- time

- Modulo Schedule

- Sample Loop

- Data Dependency Graph (DDG)

- Challenge with Code Generation Heuristics Employing Ad-hoc Routing Strategies

- DDG, Arch Description, Target II
- Scheduling (IMS)
- Place & Route
- Mapping

- Routing Data Dependency via PEs
- - EMS [H. Park et al., PACT '08]
- - EPIMap [M. Hamzeh et al., DAC '12]

- Routing via Registers
- - REGImap [M. Hamzeh et al., DAC '13]
- - GraphMin [L. Chen et al., TRETSS '14]

- Routing via Memory
- - MEMMap [S. Yin et al., TVLSI '16]

- Cannot efficiently utilize distributed RFs to route e3 → a1

- Since a, is not rescheduled, cannot route a → e

RAMP: Resource-Aware Mapping Technique

- Partition Mapping Problem in 3 Sub-Problems
- - Systematically and Flexibly Explore Resources to Achieve Mapping, Adapting to the Application Needs

- E.g. we can choose to first map the DDG with routing via registers. Then, for any unmapped data dependency, explore different routing options, per failure analysis.

- Architecture Configuration
- - Increase in Resources

Acknowledgements

This work was partially supported by funding from NSF grants CNS 1153855 and CCF 1734516 - NSF/Intel joint research center for Computer Assisted Programming for Heterogeneous Architectures (CAPA).

Experimental Setup

- 8 MiBench benchmarks (top performance critical loops)
- RAMP modeled using CCF
- Compilation & Simulation Framework
- Available at: https://github.com/cmlasu/ccf
- CGRA modeled as a separate core coupled with ARM Cortex-like core
- Evaluation over 12 architectural configurations
- - PEs connected in a 2D torus, perform fixed-point computations
- - CGRA accesses 4 kB data memory and 4 kB instruction memory
- - Configurations vary in terms of array size, PE functionality, registers etc.

RAMP Improves CGRA’s Acceleration Capability by 2.13x

- Speedup
- - RAMP
- - REGImap
- - MEMMap
- - REGImap
- - MEMMap
- - RAMP

RAMP: Resource-Aware Mapping for CGRAs

Shail Dave, Mahesh Balasubramanian, Aviral Shrivastava
Compiler Microarchitecture Lab, Arizona State University

Performance Critically Depends on the Obtained Mapping
- Mapping Problem = Routing Problem
- Routing is needed when the dependent operations are scheduled at a distant time, or operations cannot be mapped due to resource constraints.