Dynamo: A Runtime Codesign Environment

Heather Quinn[1], Dr. Miriam Leeser
Northeastern University
[1]hquinn@ece.neu.edu

Dr. Laurie Smith King
College of the Holy Cross

Motivation: Accelerate image processing tasks through efficient use of FPGAs. Combine already designed components at runtime to implement series of transformations (pipelines)

Fast, Flexible Image Processing

Run this pipeline:

On this Environment:

- Which component implementations to use?
- How to minimize overall latency?
- When to use FPGA?
- How to change the pipeline or interfaces dynamically?

Reconfigurable Systems

- Using reconfigurable hardware incurs execution costs not present in software or ASIC-based systems
  - Hardware initialization
  - Communication
  - Reprogramming

Efficient Use of FPGAs

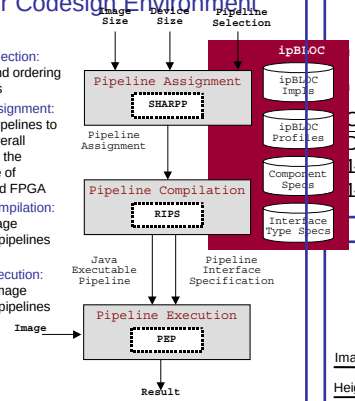
- Software algorithm's runtime for small images less than the hardware costs
  - Profiling the hardware and software runtimes for different image sizes determines the crossover point
  - Deciding at runtime to execute in software or hardware is simple for one algorithm processing one image

Image Processing Pipelines

- Series of image processing algorithms applied to an image
- Each algorithm has a software and hardware implementation
- Finding the optimal pipeline assignment is complicated
  - Exponential number of implementations
  - Coupling costs differ for each pipeline assignment
- Need a strategy to find a fast pipeline implementation at runtime

Our Codesign Environment

- Pipeline Selection: choosing and ordering components
- Pipeline Assignment: assigning pipelines to minimize overall latency with the efficient use of software and FPGA
- Pipeline Compilation: creating image processing pipelines dynamically
- Pipeline Execution: executing image processing pipelines



Goal: If pipeline selection is left to the image analyst, can the other three steps be performed automatically at runtime?

Four Shortcomings in Codesign

- Applications are configured statically
  - Design is not sensitive to user changes
- FPGA-based tools do not account for overhead costs
  - Latency is underestimated
- Partition bound too early
  - Interface between HW and SW is hard coded
- Interface changes too costly
  - System code needs extensive rewrites

Four Challenges to Codesign

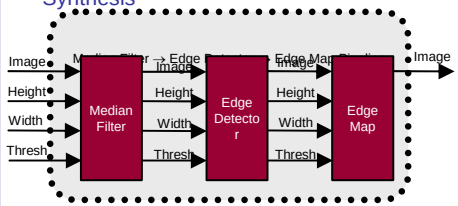
- Combining two design processes
  - Unify implementation languages
- Partitioning design
  - The pipeline assignment problem
- Interfacing hardware and software
  - Abstract communication layer and runtime interface synthesis
- Choosing a target architecture
  - One FPGA and one GPP

SW/HW Runtime Procedural Partitioning Tool

- Solves PA within either fixed or adaptive time limit based on user's choice
- Chooses an algorithm to solve PA based on pipeline size

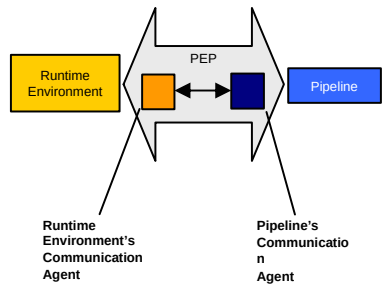
Optimization Method	Fixed	Adaptive
Dynamic Programming	1-7	1-15
1-Opt Tabu Search with Greedy	8,9	--
1-Opt Tabu Search with All Hardware	10-20	16-20

Runtime Interfacing for Pipeline Synthesis



- Builds executable pipeline from PA solution
- Connects the appropriate implementations so that the coupling costs are satisfied

Packet Exchange Platform



- Separates pipeline from runtime environment
- Makes communication abstract and generic

*Absolute Relative Error (ARE) = |(measured latency - predicted latency) / measured latency|

A Two Component Pipeline

- Median Filter→ Histogram
- Image size of 40185 pixels

PA	Predicted Latency w/ overhead (ms)	Predicted Latency w/o overhead (ms)	Actual Latency (ms)	ARE* w/ overhead	ARE* w/o overhead
sw/sw	2509	2509	2141	0.1719	0.1719
sw/hw	4905	2516	3967	0.2365	0.3658
hw_1/sw	2864	376	2975	0.0373	0.8736
hw_2/sw	2852	392	3141	0.0920	0.8752
hw_3/sw	3036	577	3004	0.0107	0.8079
hw_1/hw	2896	399	3042	0.0480	0.8688
hw_2/hw	2884	584	2803	0.0289	0.7917

Random Pipeline Test

- Forty test pipelines of different lengths were run in the Dynamo system for the best latency solution
- Image size of 40185 pixels
- Average ARE: 23% with overhead, 70% without

Test	Predicted Latency w/ overhead (ms)	Predicted Latency w/o overhead (ms)	Actual Latency (ms)	ARE* w/ overhead	ARE* w/o overhead
1	1111	1111	1309	0.1513	0.1513
5	2902	375	3169	0.0843	0.8817
10	3362	743	3571	0.0585	0.7919
15	4509	1411	4789	0.0585	0.7054
20	4849	1701	5955	0.1857	0.7144
25	4928	1785	6012	0.1803	0.7031
30	5297	2114	8560	0.3812	0.7530
35	6006	2575	10922	0.4501	0.7642
40	7289	3450	12217	0.4034	0.7176

Future Work

- Extend the pipeline assignment problem for FPGA devices:
  - in a network of workstations
  - with embedded processors
- Extend the pipeline assignment problem's objectives to include power minimization
- Extend the latency model to include an estimation of the error for better accuracy

Publications

Dynamo: A Runtime Partitioning System, L. A. Smith King, Miriam Leeser, and Heather Quinn, The 2004 International Conference on Engineering of Reconfigurable Systems and Algorithms (ERSA'04), pp. 145-151, Las Vegas, Nevada, June 21-24, 2004.

Runtime Assignment of Reconfigurable Hardware Components for Image Processing Pipelines, Heather Quinn, L.A. Smith King, Miriam Leeser, and Waleed Meleis, 2003 IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM), pp. 173-182, Napa, CA, April 8-11, 2003.