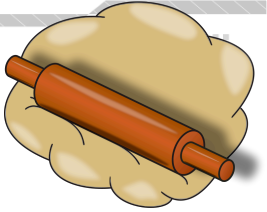


QuickDough: A Rapid FPGA Loop Accelerator Design Framework Using Soft CGRA Overlay

Cheng Liu, Ho-Cheung Ng, **Hayden K. H. So**

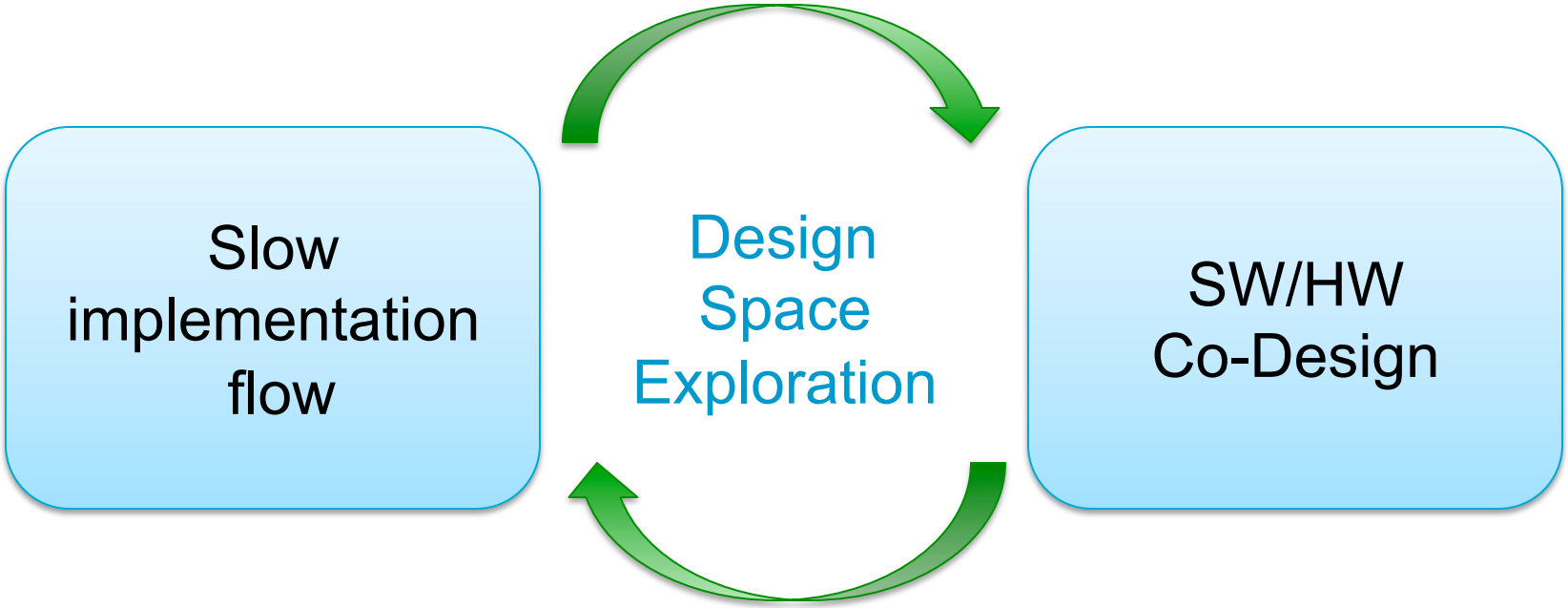
Department of
Electrical and Electronic Engineering
University of Hong Kong

FPT 2015 – 2015/12/07

Motivation

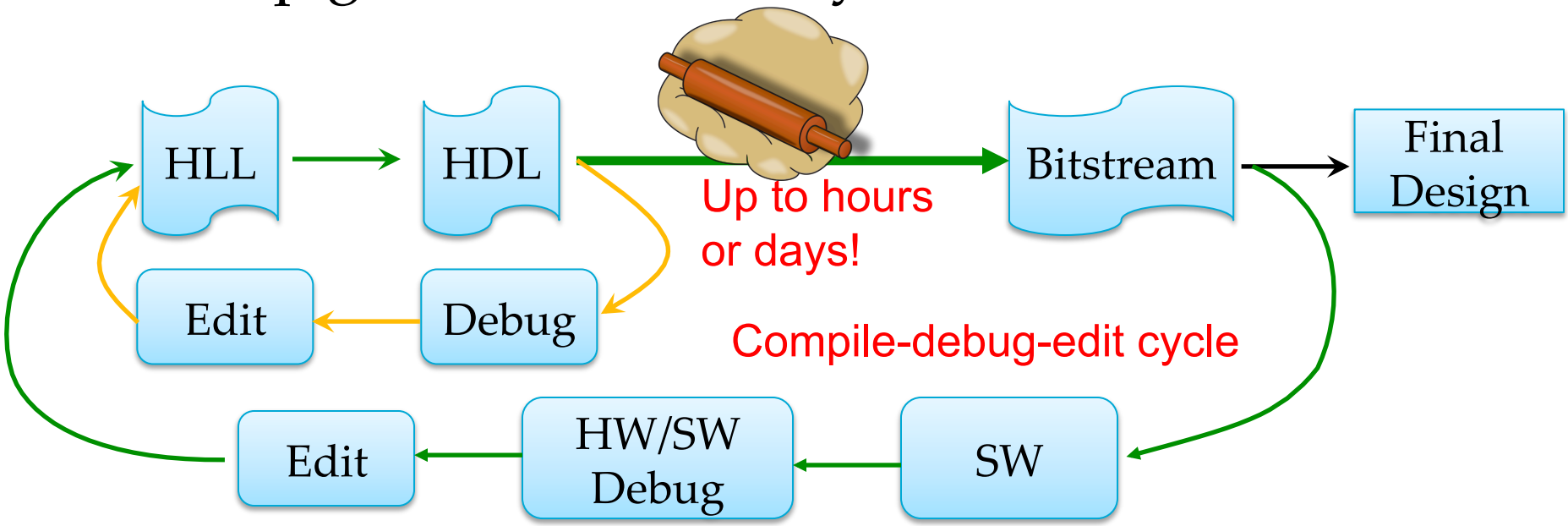
Designing FPGA Accelerators is Hard

Design productivity much lower than sw development



Need for Fast Design Iterations

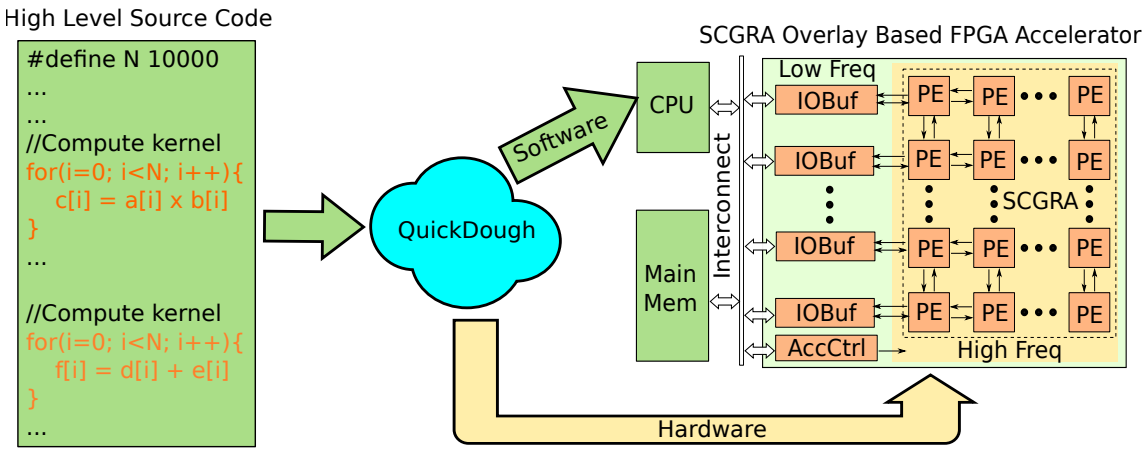
- Multiple levels of **edit-compile-debug** iterations to develop good accelerator systems



QuickDough Overview

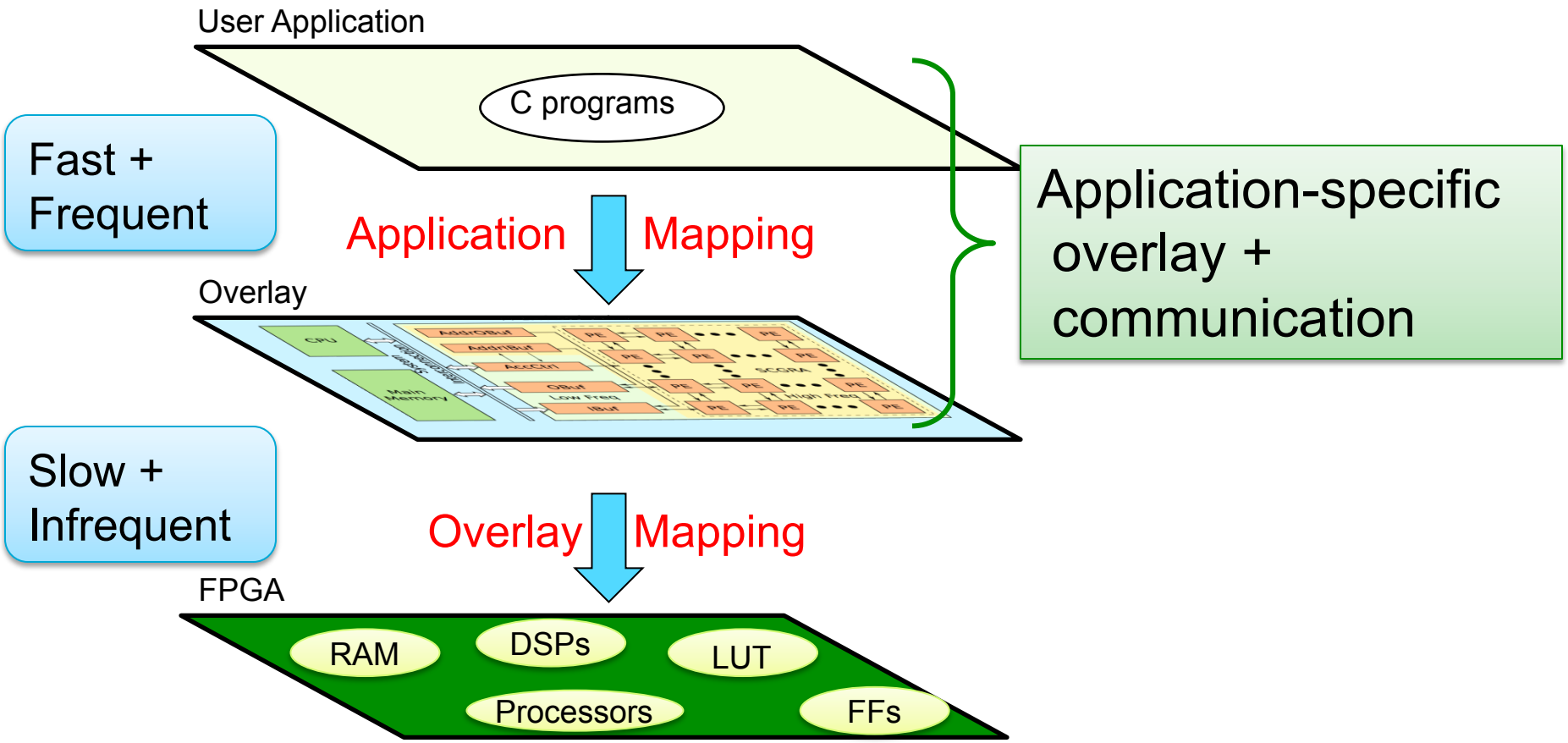
- A **high productivity compilation framework** for high-level applications on CPU-FPGA computers.
  - Compiles **software** and **FPGA accelerator** design within the same framework
  - Overall **SW-like compile speed**

Goals:
- Fast design iterations
- Good performance

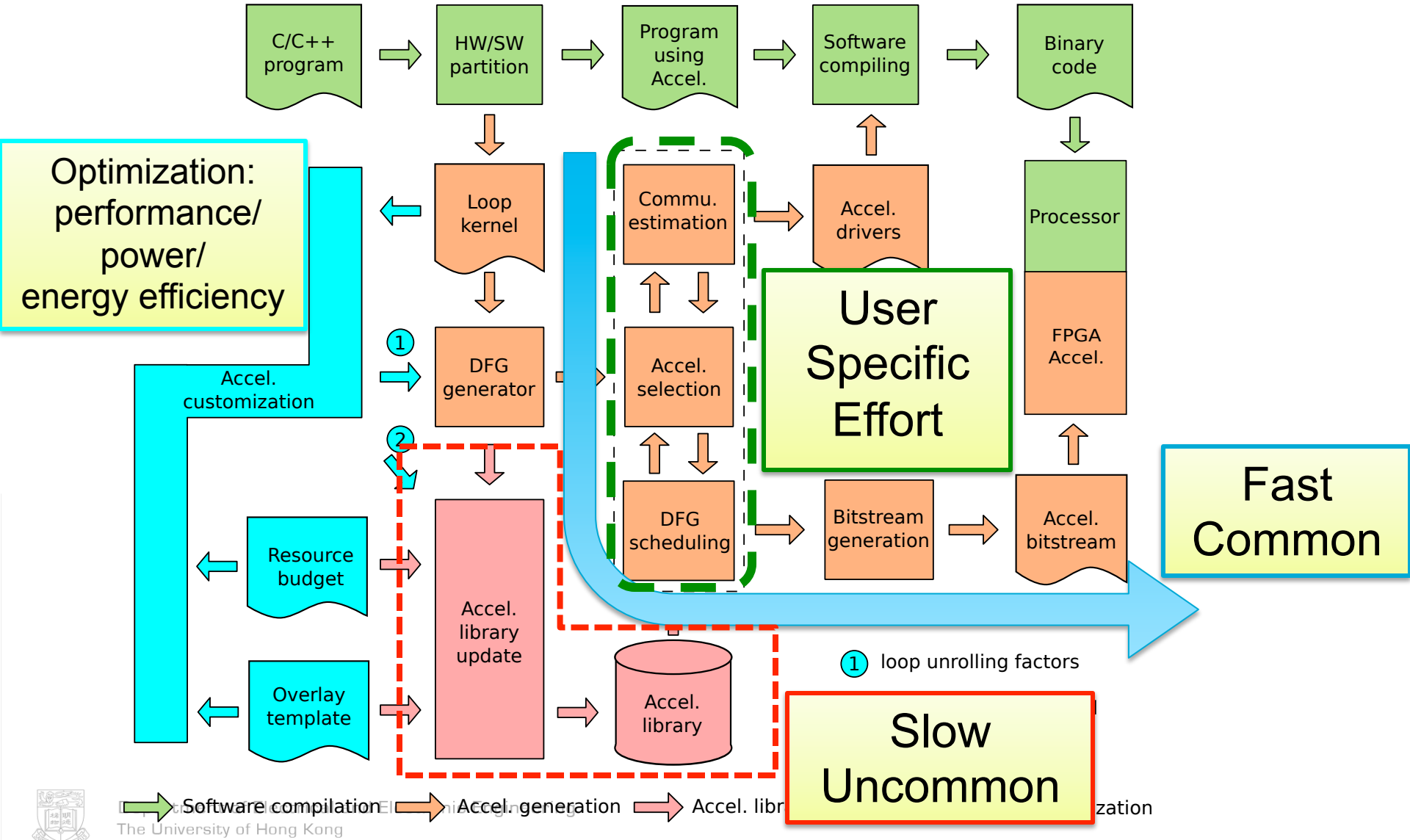


Main Idea: FPGA Overlay

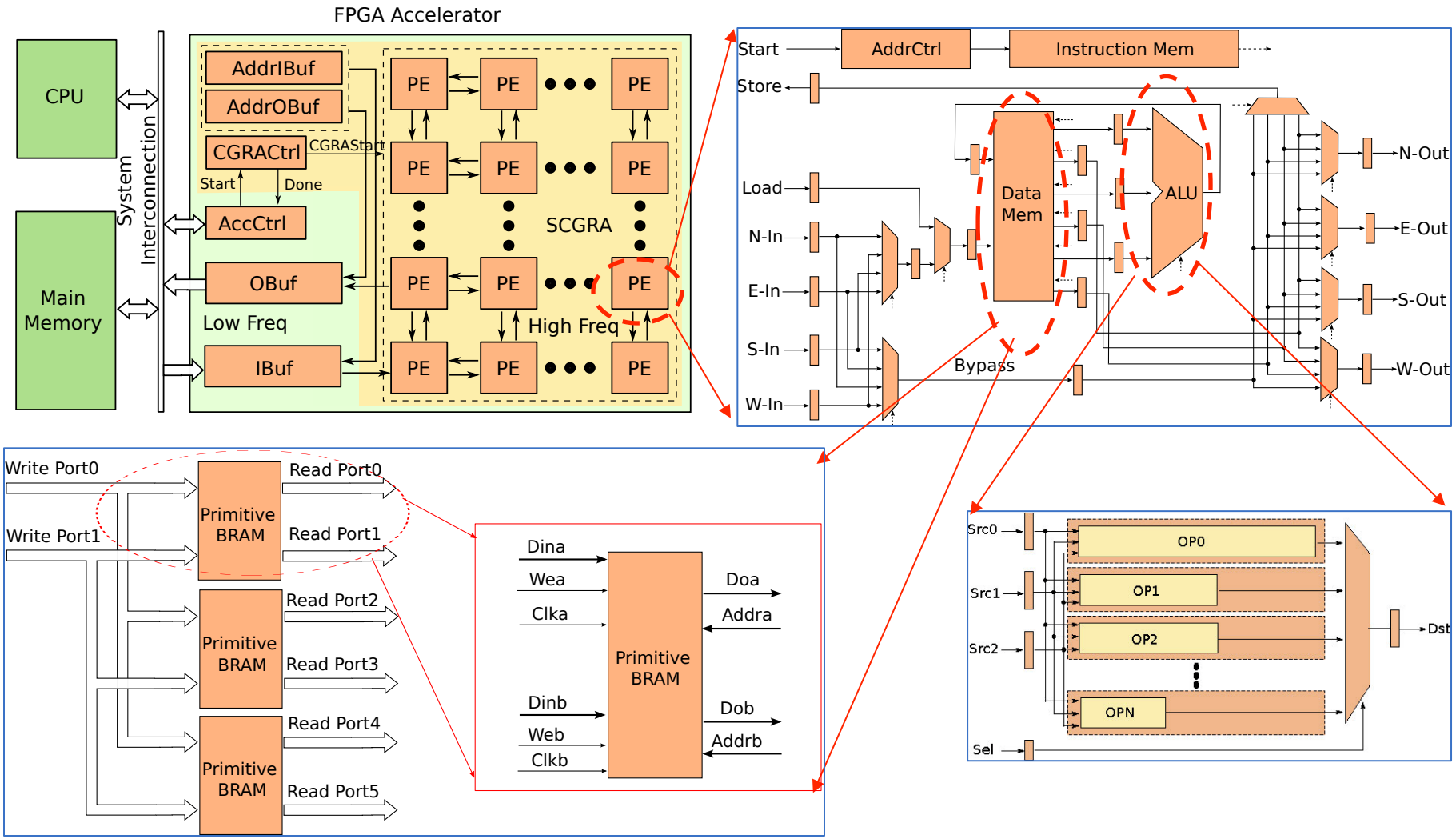
A 2-step HW-SW compilation approach



QuickDough Design Flow



QuickDough Soft CGRA Overlay

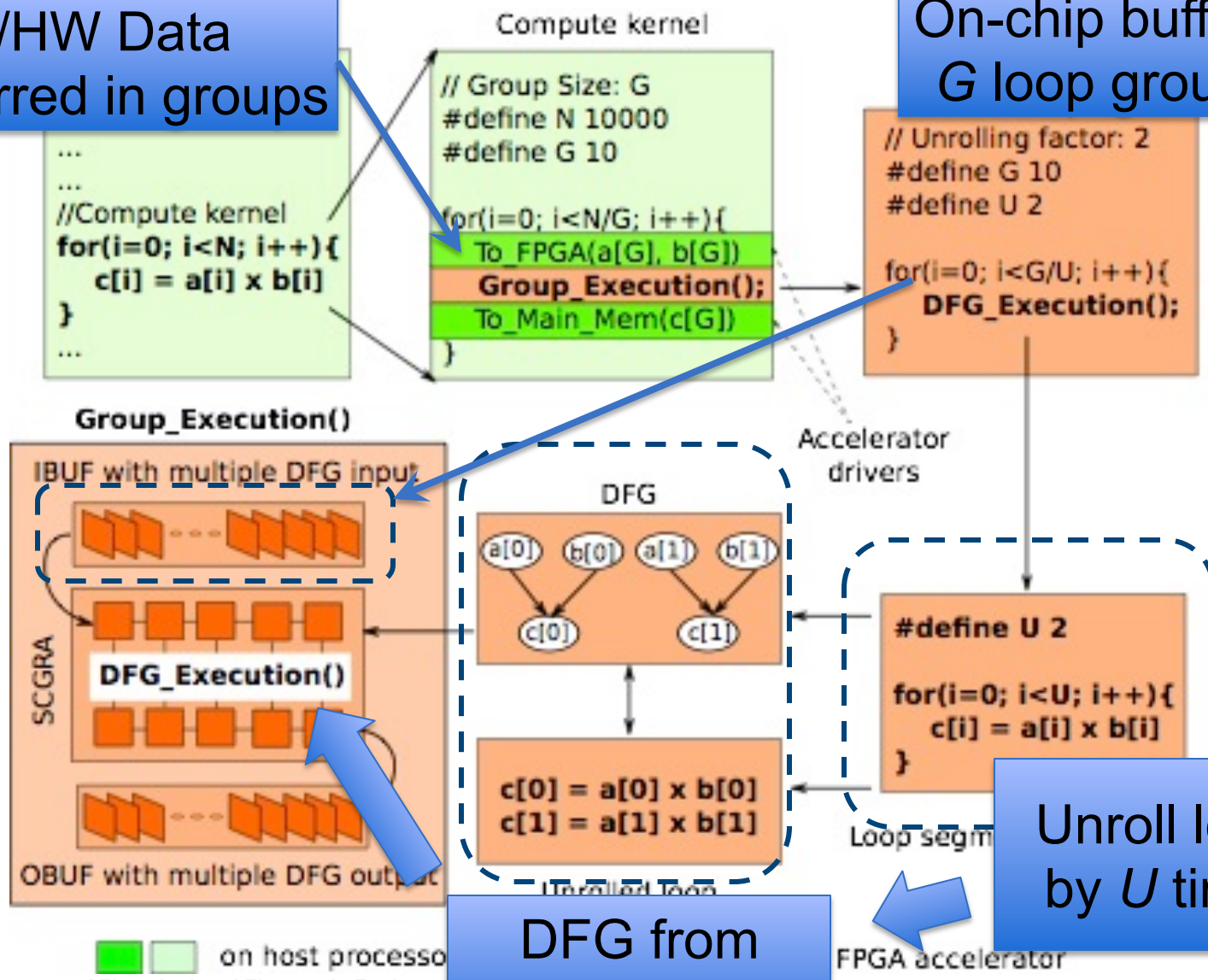


Highly pipelined, good scalability, simple and easy to extend

Loop Execution on the Accelerator

SW/HW Data transferred in groups

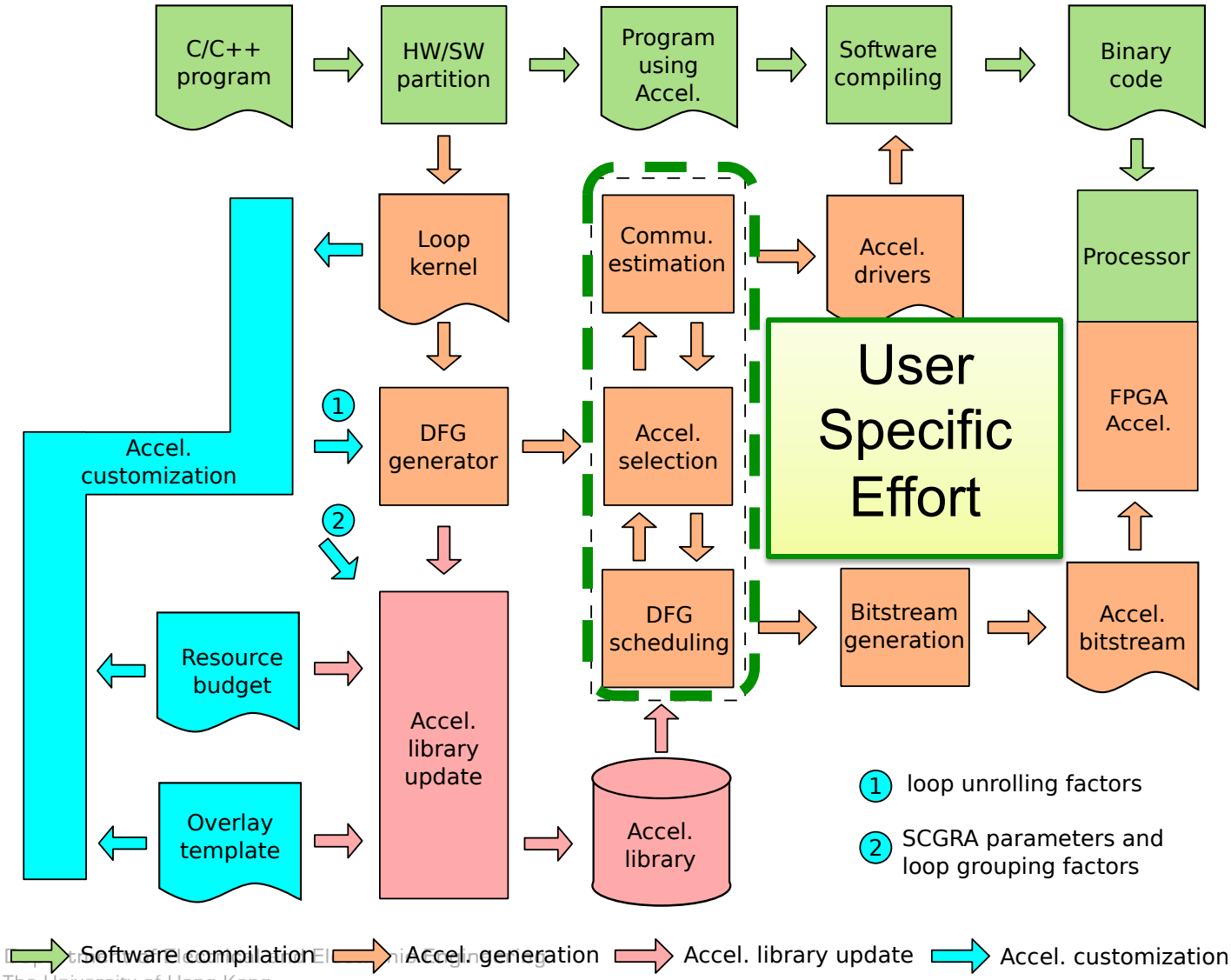
On-chip buffer holds *G* loop groups data



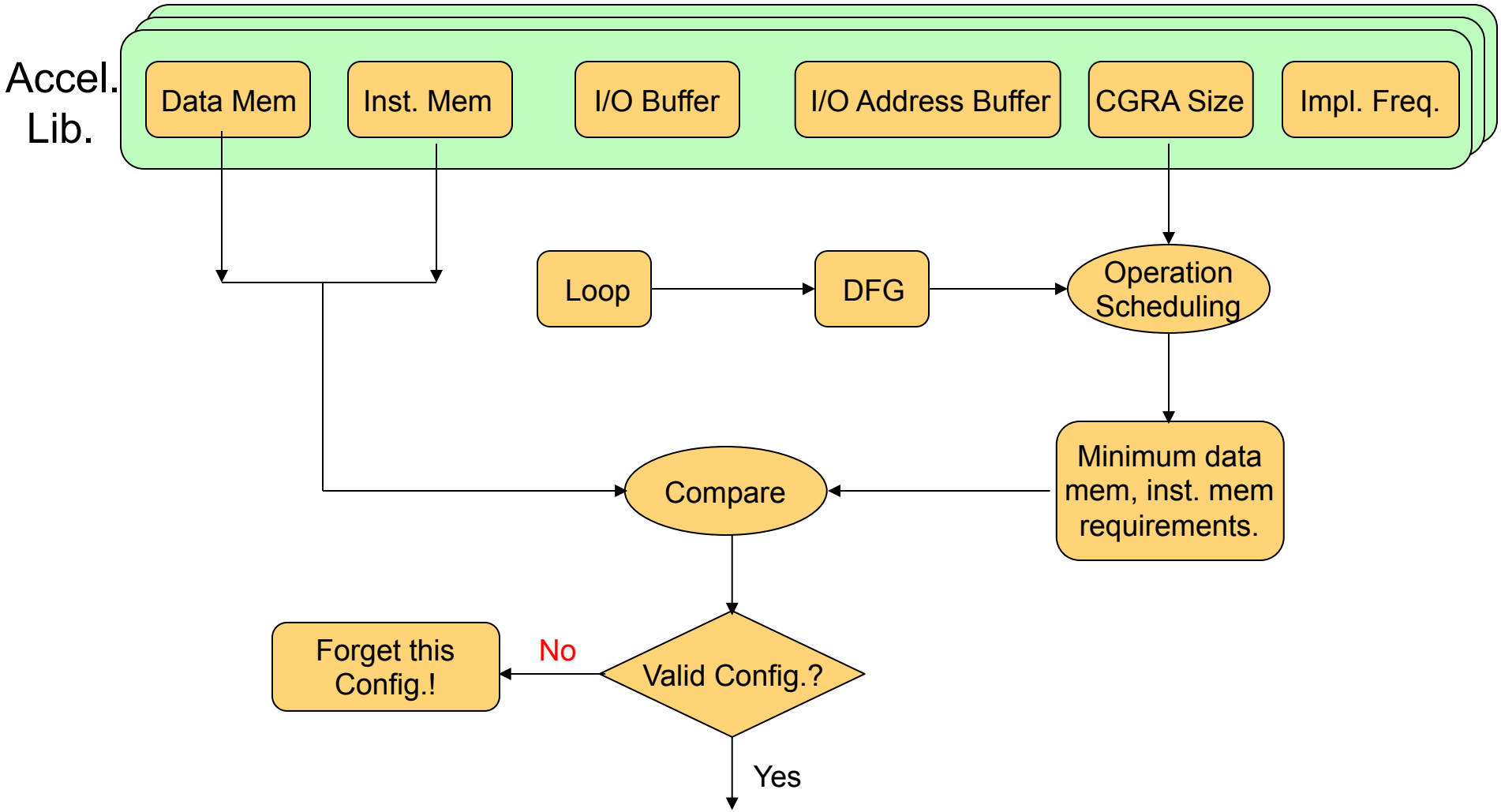
Unroll loop by *U* times

DFG from unrolled loop

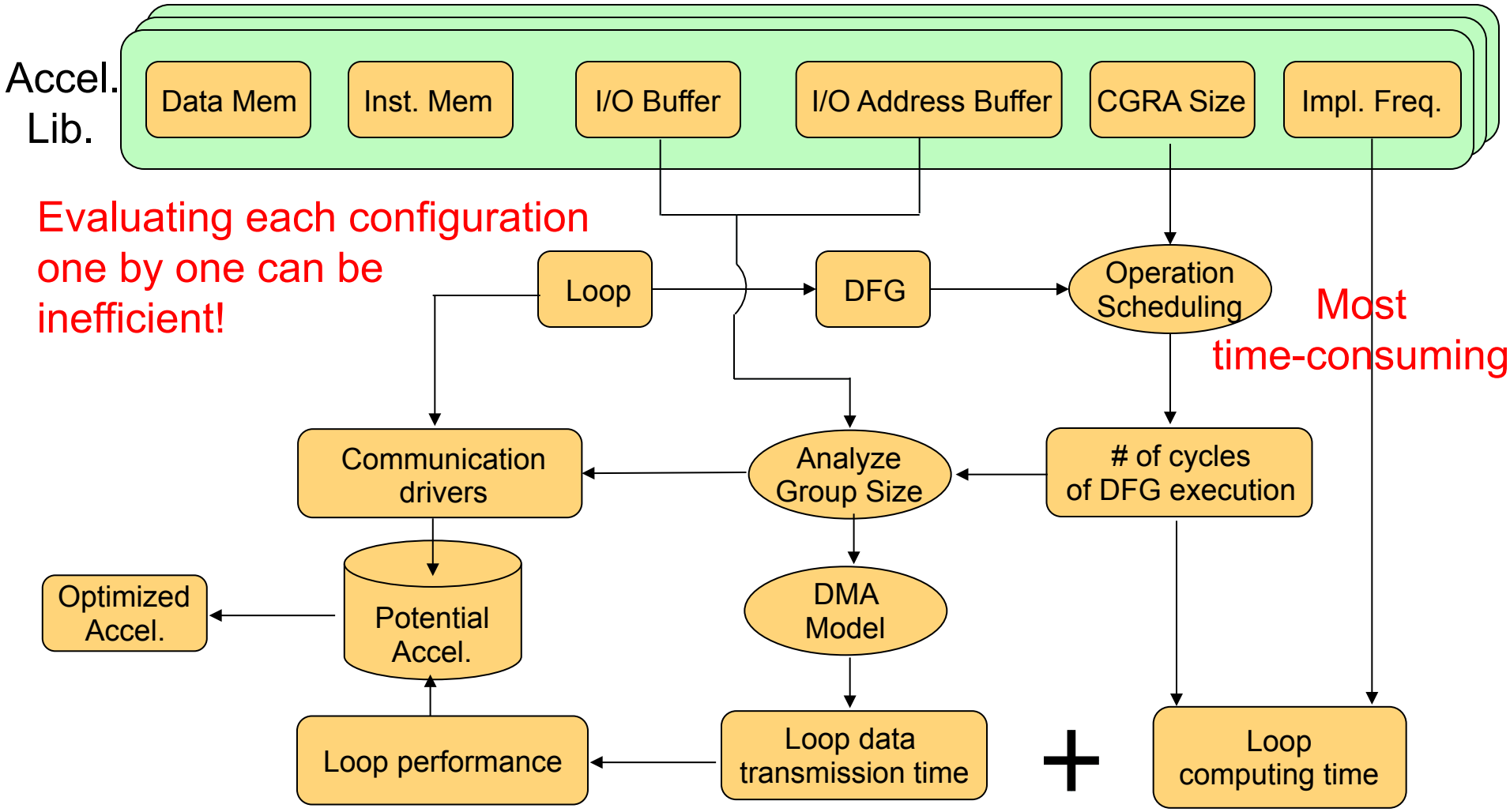
QuickDough Design Flow



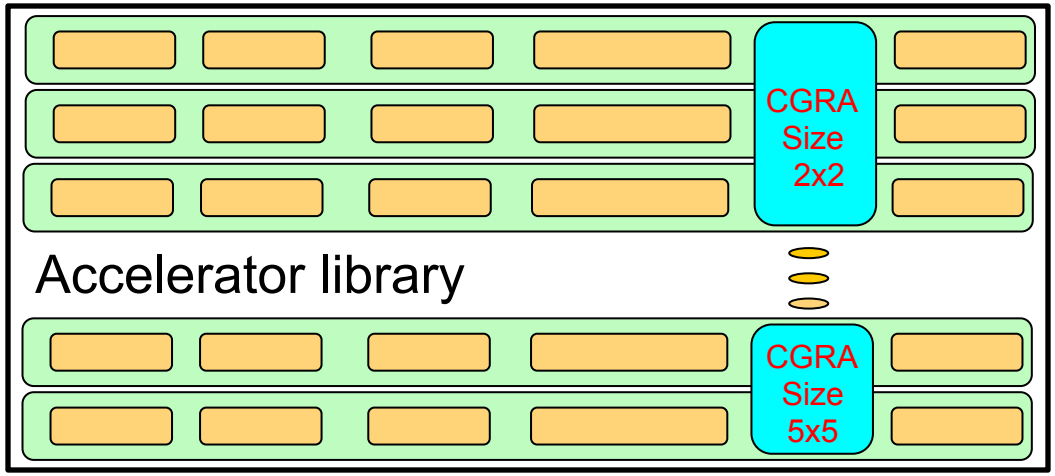
Accelerator Selection(1)



Accelerator Selection(2)



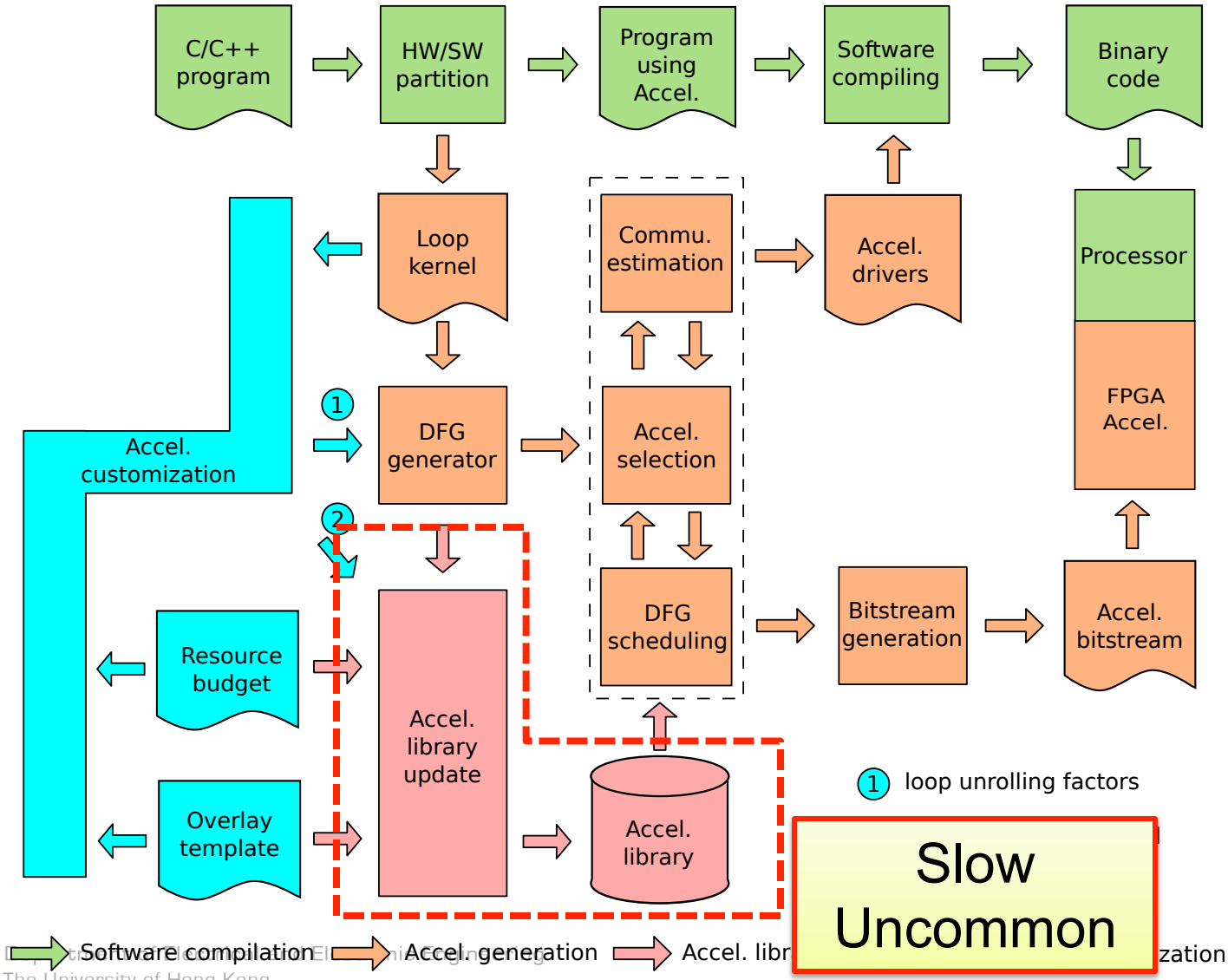
Accelerator Selection(3)



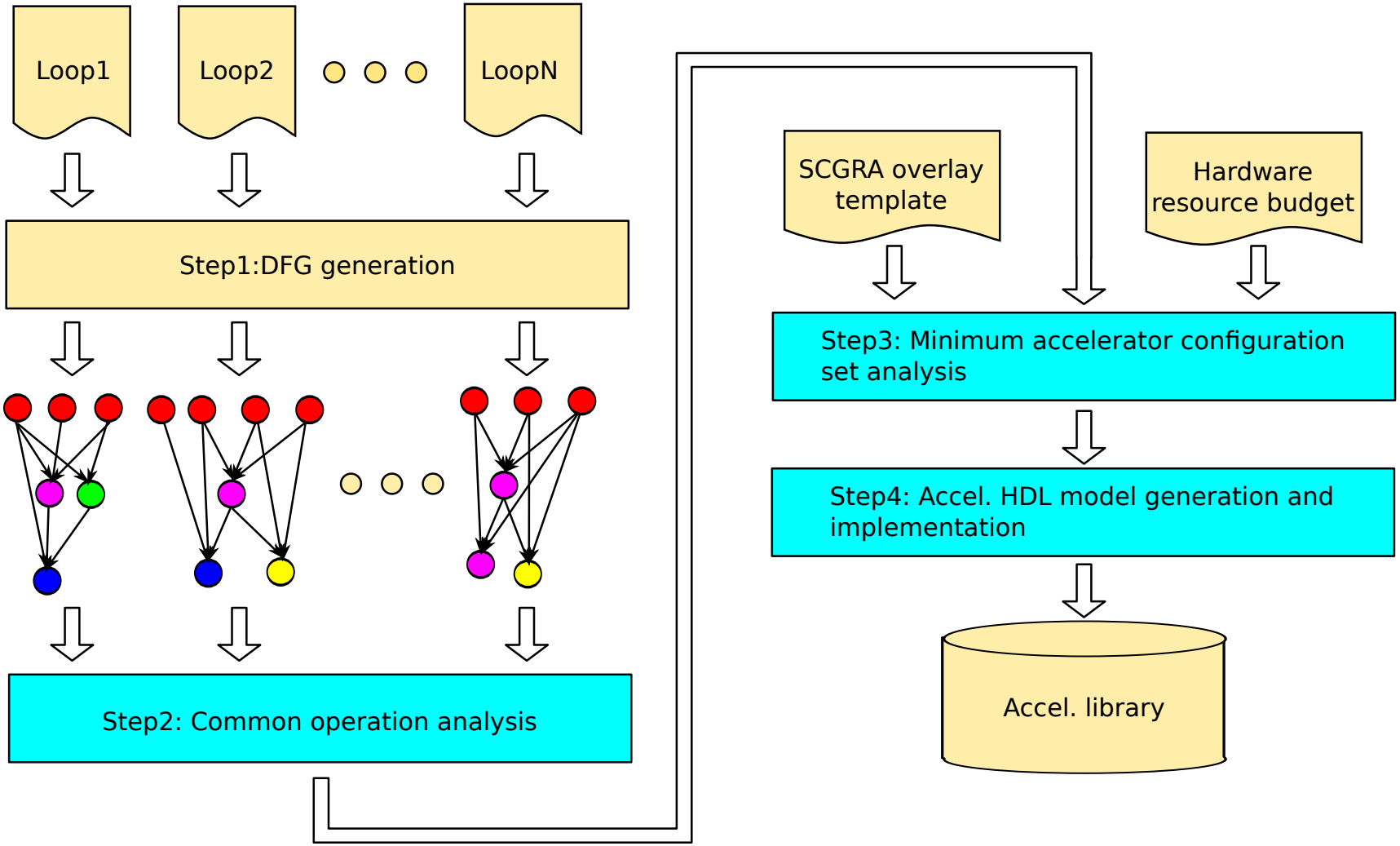
A single operation scheduling can be reused to evaluate a group of accelerator configurations.

- Selection Strategy
  - -O0: accelerators with smallest CGRA size
  - -O1: target three groups of accelerators with different SCGRA size (small, medium, large)
  - -O2: full search

QuickDough Design Flow



Accelerator Library Pre-build



Evaluation

- Two Goals:

Compilation Time

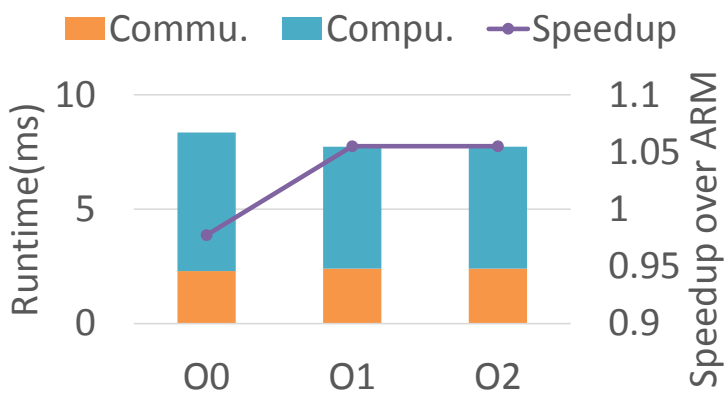
Resulting System Performance

Benchmark

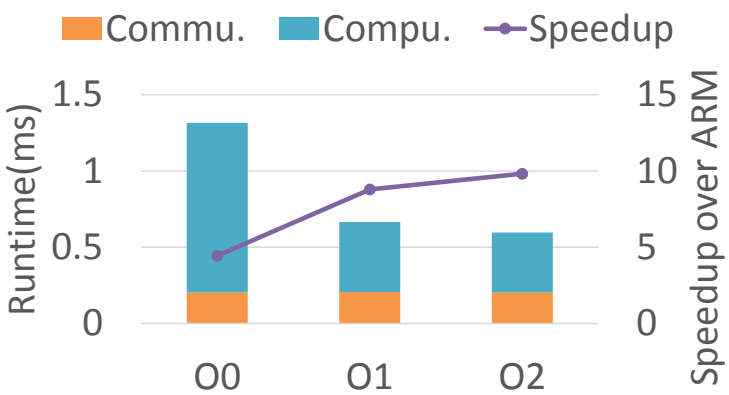
Benchmark	App Parameters	Loop kernel
MM (Matrix multiplication)	matrix size: 100100	100
FIR	# of input: 10000 # of Taps +1: 50	0
SE (Sobel edge detector)	# of vertical pixels:128 # of horizontal pixels: 128	
KM (Kmean)	# of nodes: 5000 # of centroids: 4 # of node dimension: 2	

Benchmark	Unrolling	DFG Size
MM	1100	
FIR	5050	200
SE	163	
KM	2	

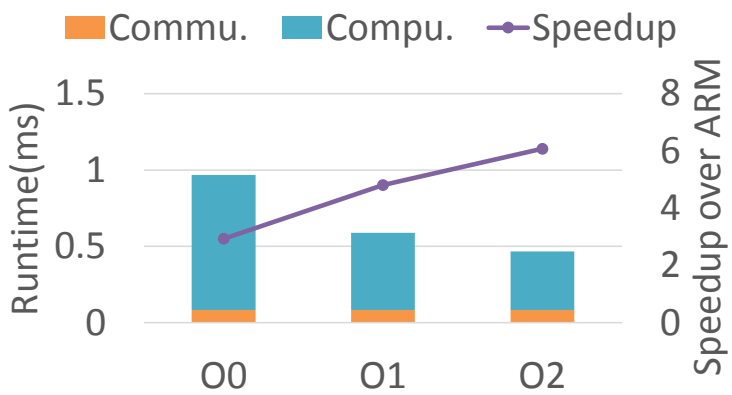
Accelerator Performance



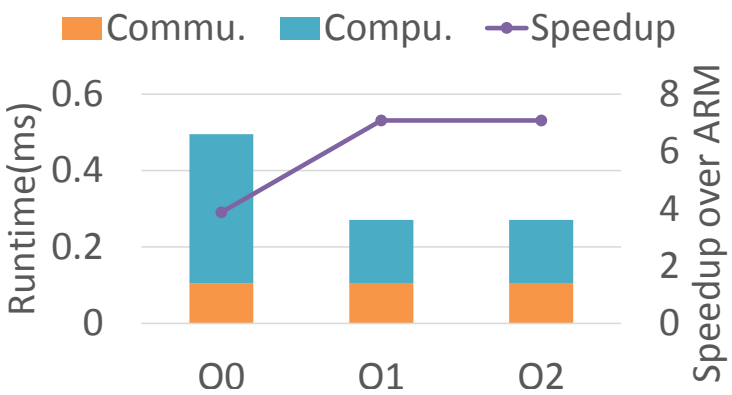
MM



FIR

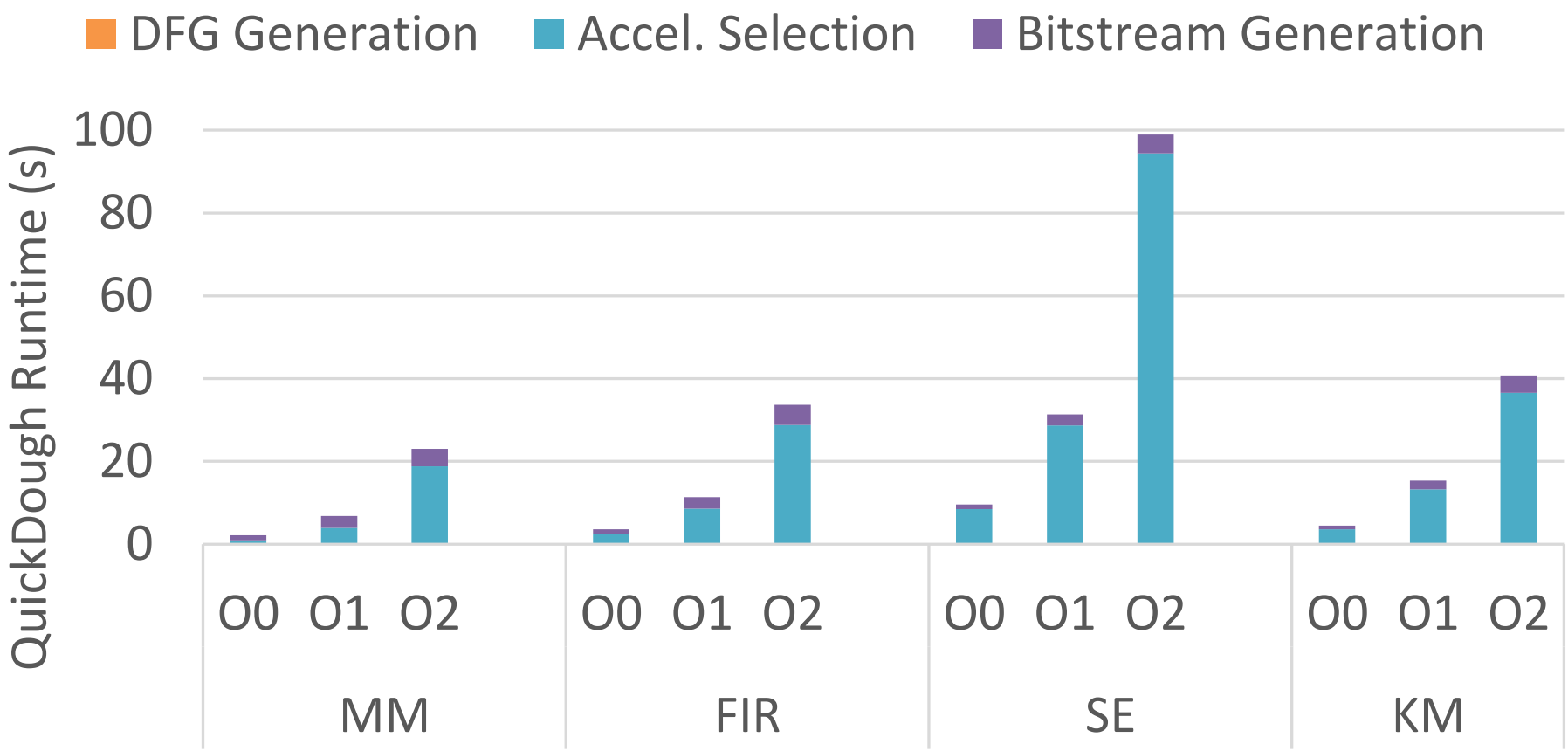


SE



KM

Accelerator Generation Time



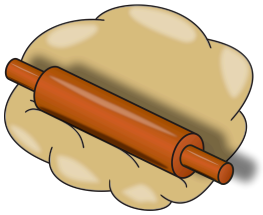
Resulting Accelerator Configurations

Opt. Option	Resulting Config	MM	FIR	SE	KM
O0	SCGRA size	2x2	2x2	2x2	2x2
	Inst. Mem Depth	4K	4K	4K	4K
	I/O Buffer Depth	4K	4K	4K	4K
	Grouping Factor	50x5x100	2500x50	128x64x3x3	1250x4x2
O1	SCGRA Size	3x3	3x3	3x3	5x5
	Inst. Mem Depth	2K	2K	4K	1K
	I/O Buffer Depth	2K	2K	1K	2K
	Grouping Factor	25x5x100	1250x50	64x32x3x3	1000x4x2
O2	SCGRA size	3x3	4x4	4x4	5x5
	Inst. Mem Depth	2K	1K	2K	1K
	I/O Buffer Depth	2K	8K	1K	2K
	Grouping Factor	25x5x100	5000x50	64x32x3x3	1000x4x2

Conclusions

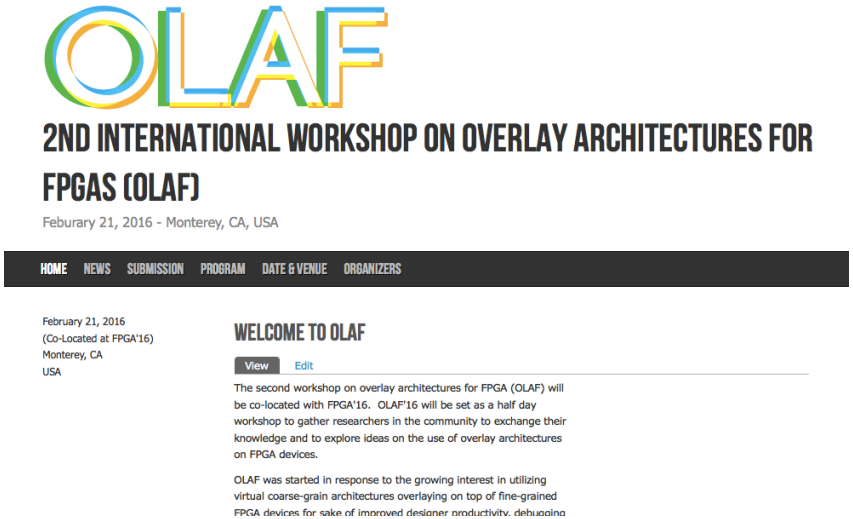
- QuickDough can generate HW/SW accelerator systems **rapidly** and with **good performance**.
- Overlay allows rapid compilation
- A overlay library assist rapid performance tuning
- Data I/O co-optimization with selected overlay

THANK YOU



Happy baking…

Questions?



2nd International Workshop on Overlay Architecture (OLAF)

Co-located with FPGA'16

Deadline: **Jan 11, 2016**

http://olaf.eecs.berkeley.edu

Department of Electrical and Electronic Engineering
The University of Hong Kong