Pacific Northwest NATIONAL LABORATORY

Northeastern University

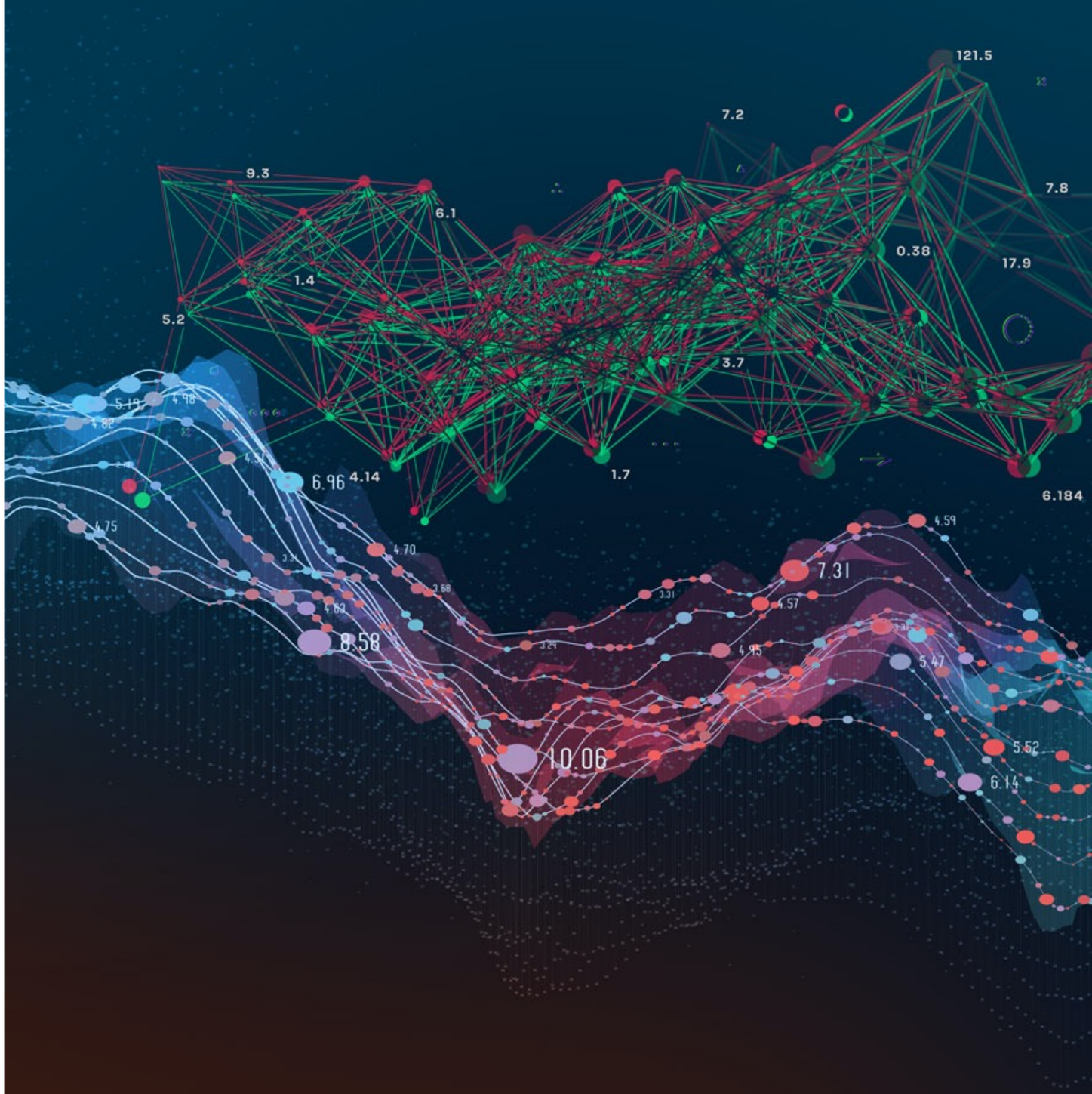
SODA-OPT
Enabling System-Level Design in MLIR for High-Level Synthesis and Beyond

March 21, 2022

Nicolas Bohm Agostini

PhD Student at Northeastern University

PhD Intern at PNNL

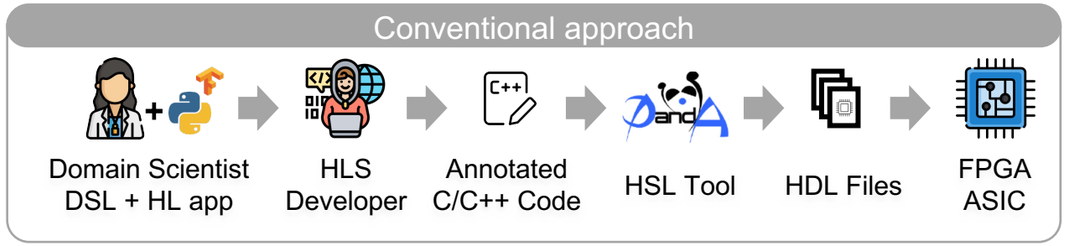


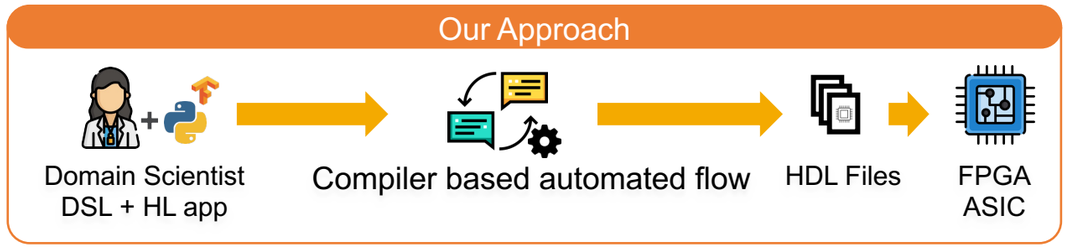
U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

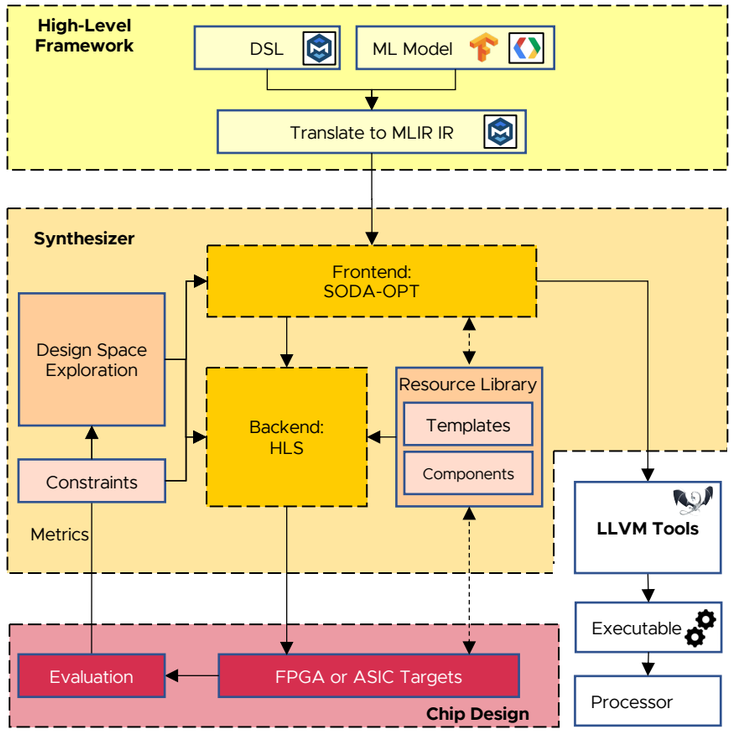
PNNL-SA-171156

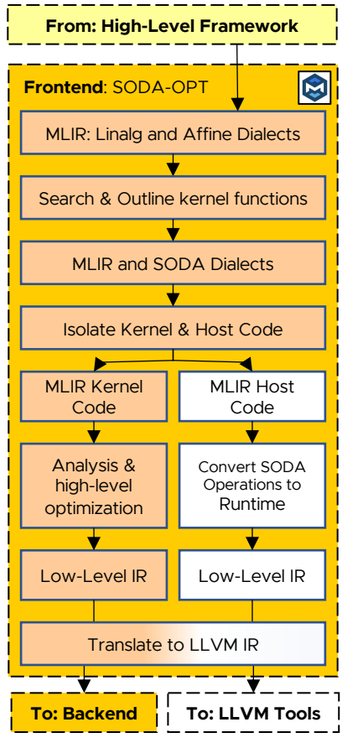
Motivation

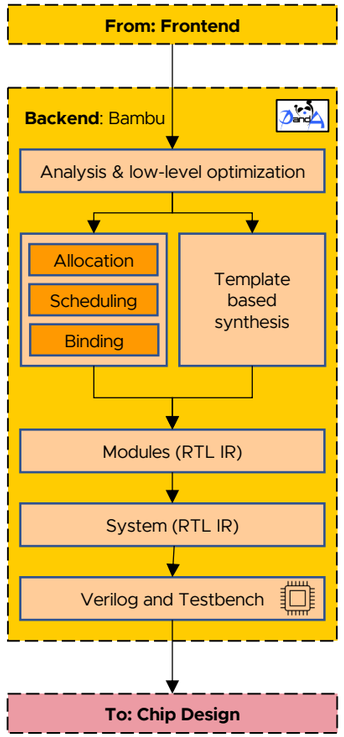

Conventional approach
Domain Scientist
DSL + HL app
HLS
Developer
Annotated
C/C++ Code
HSL Tool
HDL Files
FPGA
ASIC


Our Approach
Domain Scientist
DSL + HL app
Compiler based automated flow
HDL Files
FPGA
ASIC

SODA End-To-End Toolchain







Pacific Northwest NATIONAL LABORATORY

Northeastern University

Goals

1. Learn how to use the SODA toolchain and the SODA-OPT compiler to perform end-to-end synthesis of high-level applications
2. Experiment with different high-level optimization strategies provided by SODA-OPT

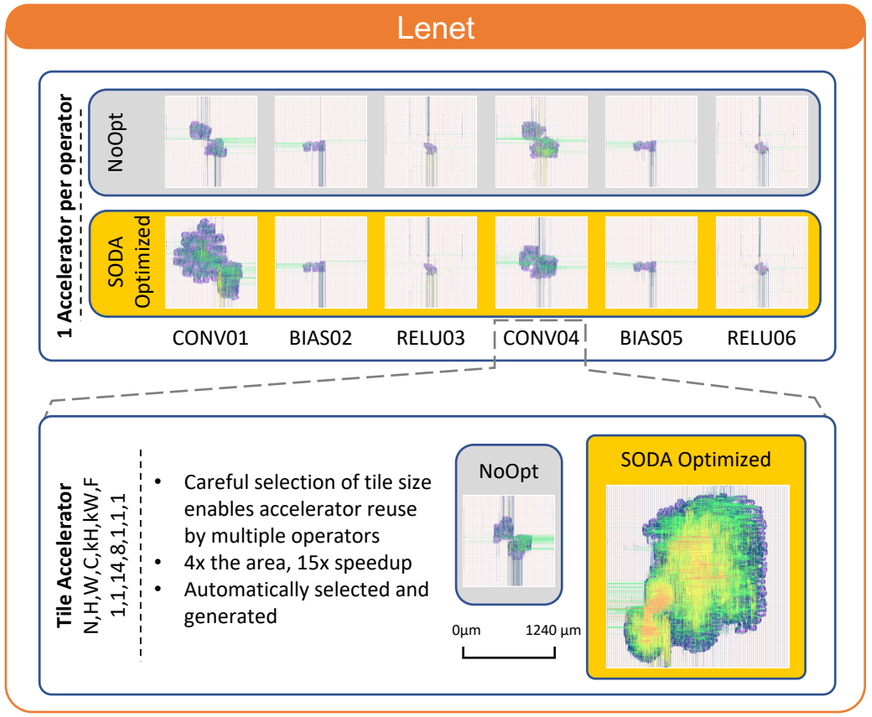
Why are HL code transformations important?

PolyBench

EXECUTION TIME (IN CLOCK CYCLES) FOR POLYBENCH KERNELS WITH ASIC TARGET - OPENPDK 45NM @ 500MHZ. SPEEDUP SHOWN IN PARENTHESIS.

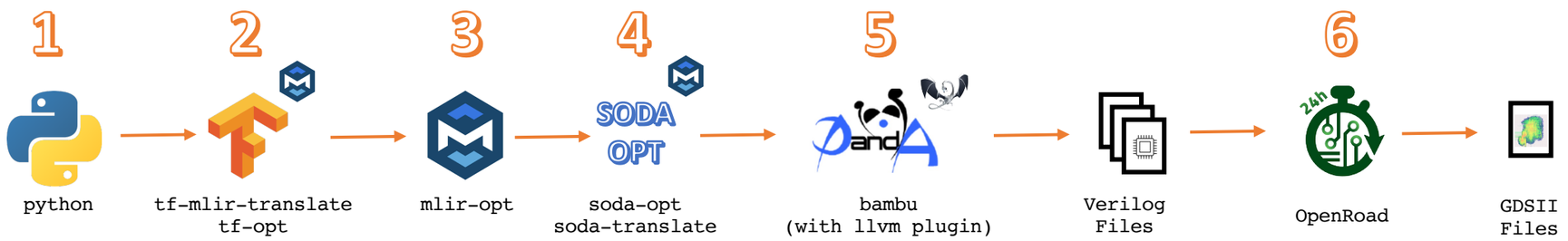
Opt. Strategy	No High Level Opts.				SODA-OPT Pipeline			
Kernel Size	**2**	**4**	**8**	**16**	**2**	**4**	**8**	**16**
symm	421	2,928	21,400	163,368	31 (13.6x)	34 (86.1x)	325 (65.8x)	2,600 (62.8x)
three_mm	388	3,087	25,010	211,298	47 (8.3x)	82 (37.6x)	656 (38.1x)	5,248 (40.3x)
two_mm	315	2,475	20,258	167,490	52 (6.1x)	86 (28.8x)	688 (29.4x)	5,504 (30.4x)
gemm	186	1,446	11,922	95,376	31 (6.0x)	56 (25.8x)	448 (26.6x)	3,584 (26.6x)
doitgen	277	4,282	67,666	999,698	29 (9.6x)	258 (16.6x)	2,064 (32.8x)	16,512 (60.5x)
bicg	129	518	2,058	8,482	26 (5.0x)	43 (12.0x)	85 (24.2x)	340 (24.9x)
mvt	130	514	2,051	8,195	26 (5.0x)	45 (11.4x)	89 (23.0x)	356 (23.0x)
gemver	283	1,118	4,393	17,617	77 (3.7x)	106 (10.5x)	424 (10.4x)	1,696 (10.4x)
gesummv	162	578	2,178	8,722	39 (4.2x)	56 (10.3x)	105 (20.7x)	420 (20.8x)
atax	132	523	2,067	8,227	44 (3.0x)	73 (7.2x)	292 (7.1x)	1,168 (7.0x)
syr2k	186	1,310	9,018	68,986	38 (4.9x)	567 (2.3x)	3,033 (3.0x)	24,264 (2.8x)
syrk	142	990	6,714	49,250	31 (4.6x)	453 (2.2x)	2,581 (2.6x)	20,648 (2.4x)
trmm	46	532	4,402	34,018	24 (1.9x)	532 (1.0x)	4,402 (1.0x)	34,018 (1.0x)

Lenet



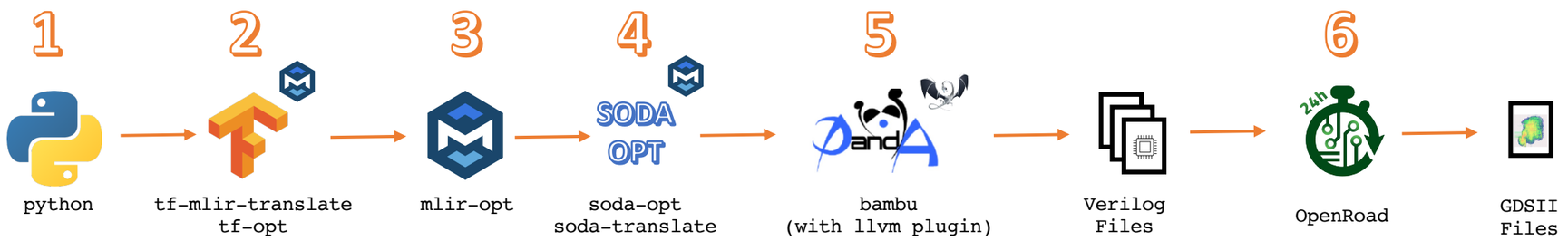
* Bambu target: OpenPDK 45nm ASIC @500MHz, 2 memory channels

Tutorial Structure



https://gitlab.pnnl.gov/sodalite/soda-opt/-/tree/main/docs/tutorials/date2022

Tutorial Structure



1
- Write a simple DNN model in python
- Translate model to `protobuf` file

2
- Convert model to MLIR (`tf` dialect)
- Lower model to `TOSA` MLIR dialect

3
- Lower model to `linalg` MLIR dialect
- Visualize the intermediate representation

4
- Select MLIR code for custom accelerator generation
- Optimize kernel code and generate IR for Bambu

5
- Synthesize baseline and optimized code into Verilog
- Compare results

6
- Place and route synthesized code
- Visualize final GDSII files

https://gitlab.pnnl.gov/sodalite/soda-opt/-/tree/main/docs/tutorials/date2022

Pacific Northwest NATIONAL LABORATORY

Northeastern University

Thank you