Pyramid: Machine Learning Framework to Estimate the Optimal Timing and Resource Usage of a High-Level Synthesis Design

Hosein Mohamamdi Makrani, Farnoud Farahmand, Hossein Sayadi, Sara, Bondi, Sai Manoj PD, Houman Homayoun,

And Setareh Rafatirad

Presenter: Rashmi Agrawal

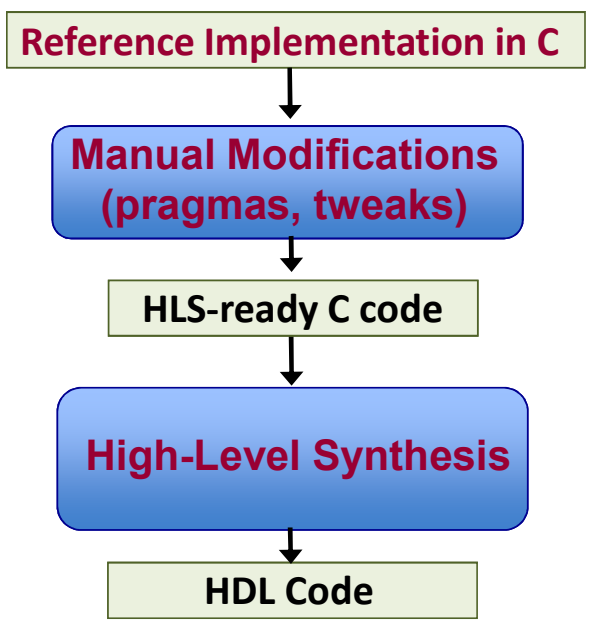
September, 2019

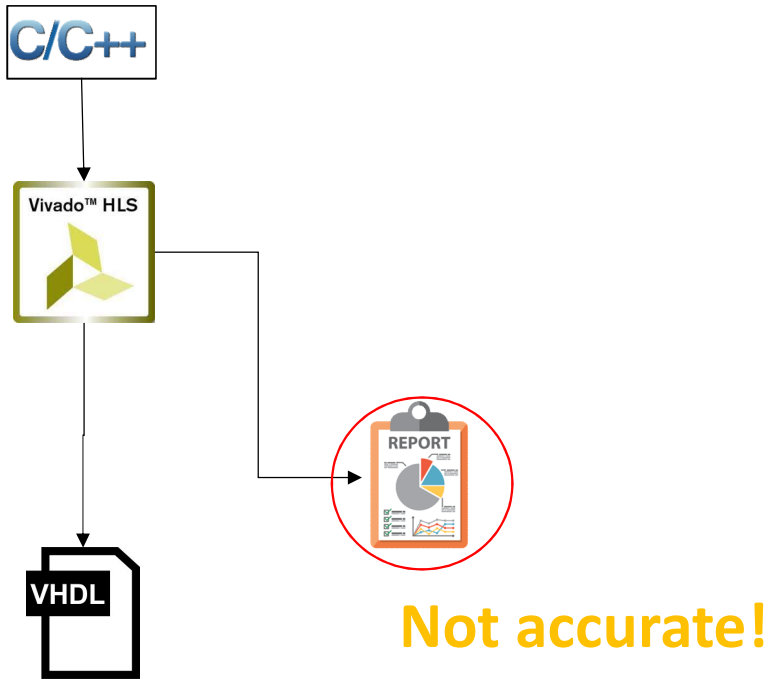
International Conference on Field-Programmable Logic and Applications (FPL 2019)

GEORGE MASON UNIVERSITY

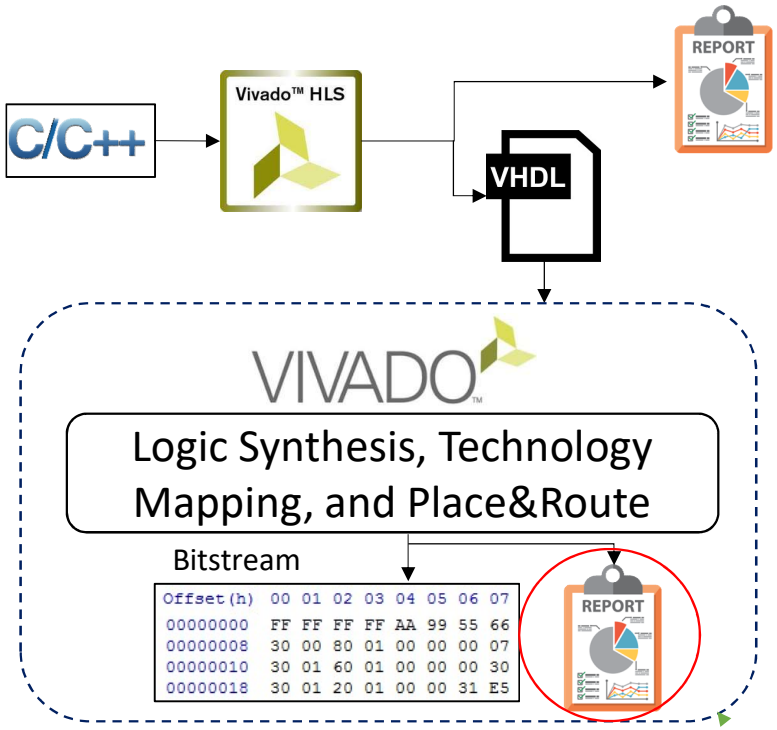
ACCELERATED, SECURE & ENERGY-EFFICIENT COMPUTING LAB (ASEEC)

HLS-Based Development and Benchmarking Flow



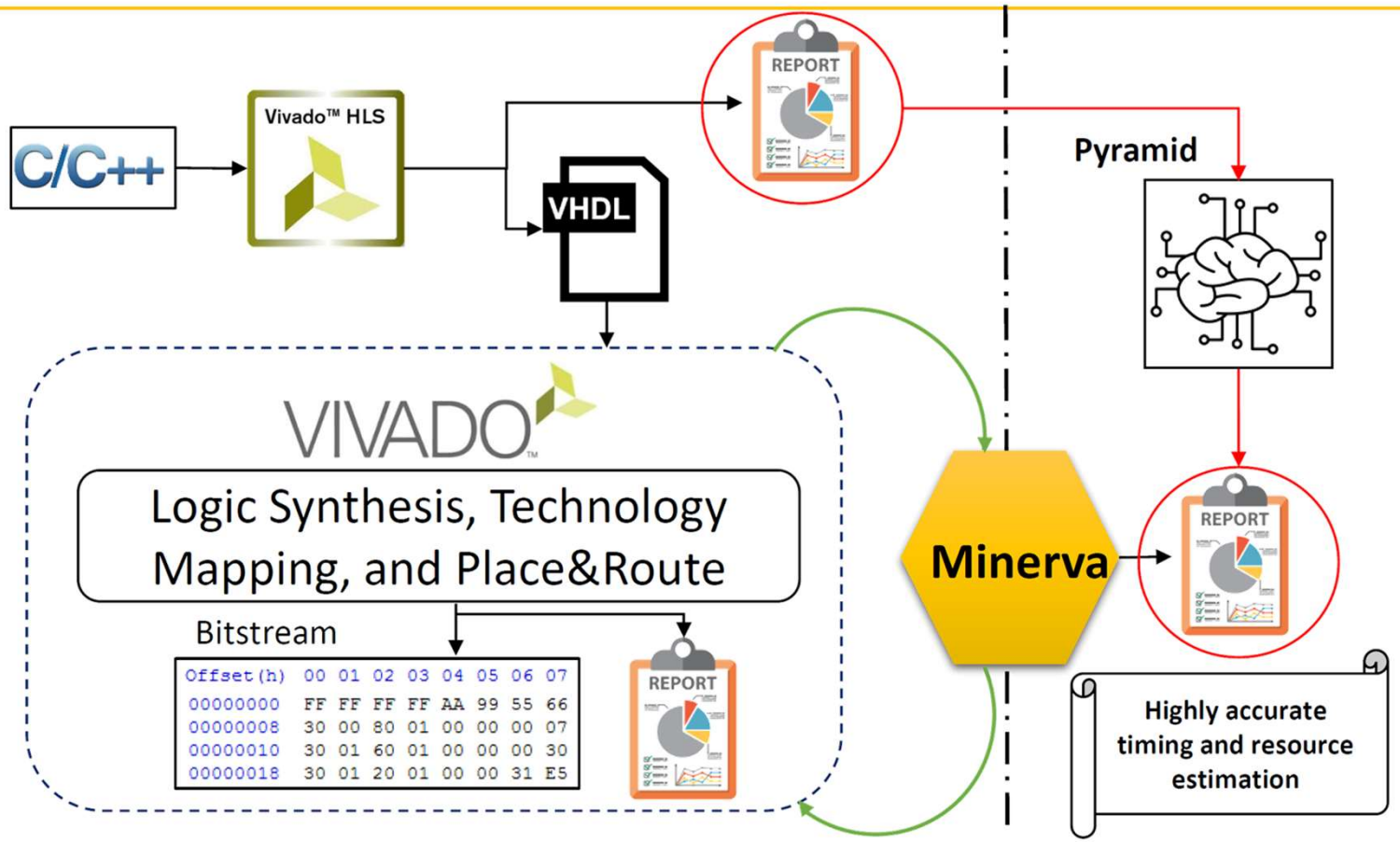


Static Timing Analysis using CAD Tools



Time consuming!
Not Optimal!

Solution


C/C++
Vivado™ HLS
VHDL
REPORT
Pyramid
VIVADO
Logic Synthesis, Technology Mapping, and Place&Route
Bitstream
Offset(h) 00 01 02 03 04 05 06 07
00000000 FF FF FF FF AA 99 55 66
00000008 30 00 80 01 00 00 00 07
00000010 30 01 60 01 00 00 00 30
00000018 30 01 20 01 00 00 31 E5
REPORT
Minerva
REPORT
Highly accurate timing and resource estimation

Goal of this Research

- We propose **Pyramid,** a framework that uses ML to accurately estimate the optimal performance and resource utilization of an HLS design
- For this purpose:
  - we first create a database of C-to- FPGA results from a diverse set of benchmarks using Minerva to find maximum clock frequency
  - We use the database to train an **ensemble** machine learning model to map the HLS-reported features to the results of Minerva

General ML Models

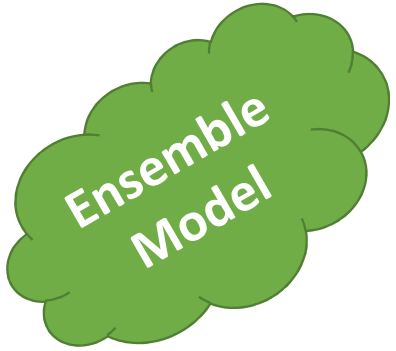
ML ESTIMATION ERRORS

Device		Artix7					Kintex7					Virtex7				
Resource		LUT	FF	DSP	BRAM	Timing	LUT	FF	DSP	BRAM	Timing	LUT	FF	DSP	BRAM	Timing
TP	LR	27%	19%	22%	29%	18%	23%	25%	17%	21%	20%	29%	20%	21%	19%	24%
	ANN	12%	17%	13%	14%	16%	10%	11%	19%	16%	15%	11%	14%	18%	14%	13%
	SVM	22%	16%	19%	18%	19%	20%	22%	18%	15%	17%	24%	21%	16%	17%	18%
	RF	9%	17%	19%	14%	16%	15%	18%	12%	11%	17%	16%	20%	19%	15%	14%

ML models: regression model, artificial neural network (ANN), support vector machine (SVM), and random forest (RF)

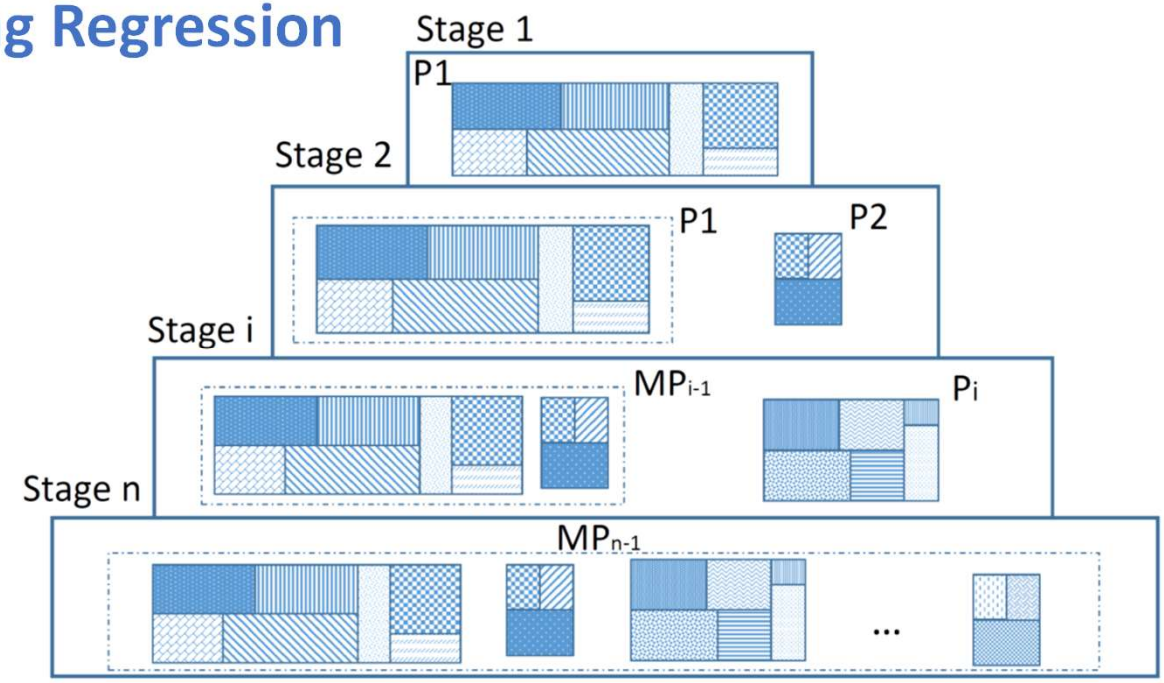
Tuning technique: Grid Search

Average Error: LR=23%, ANN=14%, SVM=19%, and RF=15%



Ensemble Model

Stacking Regression



Key Idea: Cross Validation and Batch Training

Model Parameters: threshold for the accuracy and the maximum number of iterations

Evaluation Results

AVERAGE ERROR OF PYRAMID ESTIMATIONS

Devices	Artix7		Kintex7		Virtex7	
Tragets	Resource	Timing	Resource	Timing	Resource	Timing
Pyramid-TP	6.3%	3.8%	5.2%	4.1%	4.9%	4.4%
Pyramid-TPA	4.8%	3.5%	4.7%	4.6%	4.8%	4.9%

AVERAGE ERROR OF ML TECHNIQUES FOR DIFFERENT BENCHMARKS

Benchmark's category	Machine Learning		Img/Vid Processing		Crypto.		Mathe.	
Targets	*Res*	*Tim*	*Res*	*Tim*	*Res*	*Tim*	*Res*	*Tim*
LR	29%	25%	17%	16%	38%	22%	11%	8%
ANN	17%	14%	13%	11%	19%	14%	8%	7%
SVM	22%	19%	18%	17%	23%	18%	10%	7%
RF	16%	16%	14%	12%	20%	15%	9%	7%
Ensemble	6%	5%	4%	3%	5%	4%	4%	3%

Questions

Thank you!

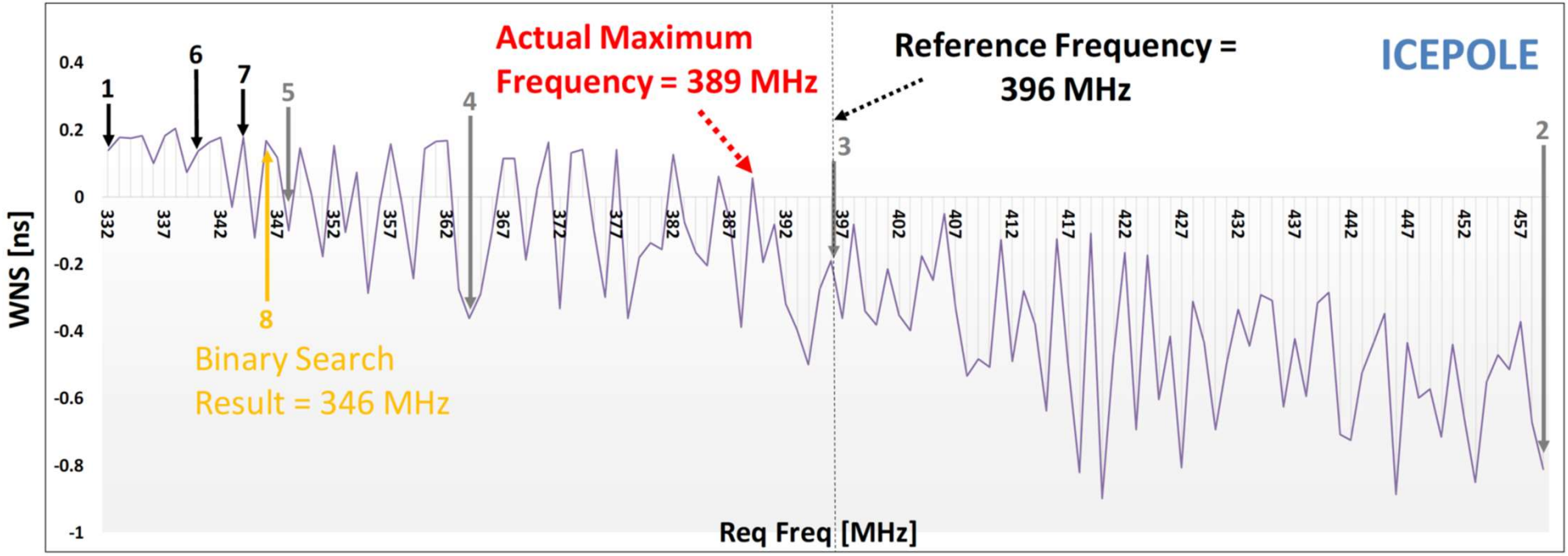
Supporting Slides

Inaccuracy of HLS Report

$$Relative\ RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\frac{p_i - a_i}{a_i})^2} \times 100$$

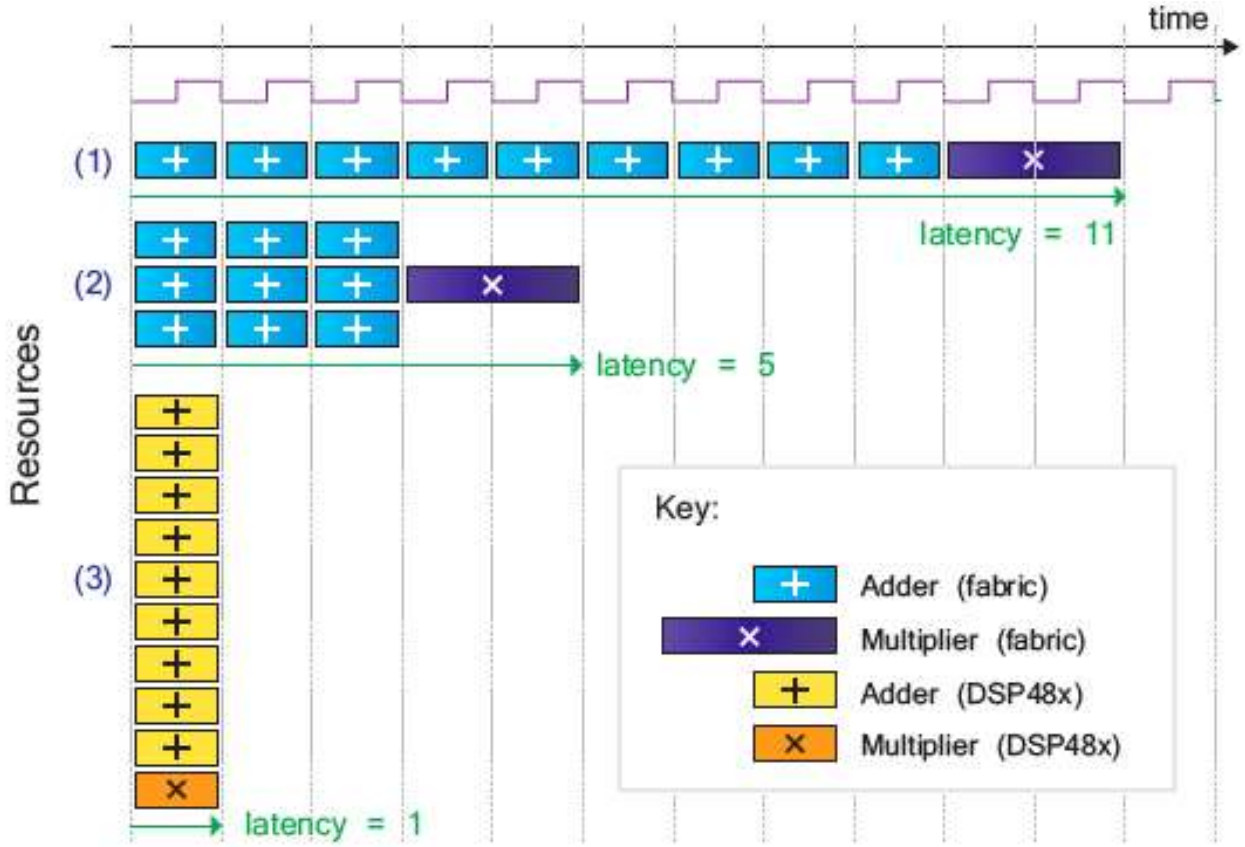
Devices	Artix7		Kintex7		Virtex7	
Tragets	Resource	Timing	Resource	Timing	Resource	Timing
HLS Estimate	91.7%	23.6%	112.5%	28.1%	88.4%	17.2%

Challenge of Hardware Evaluation

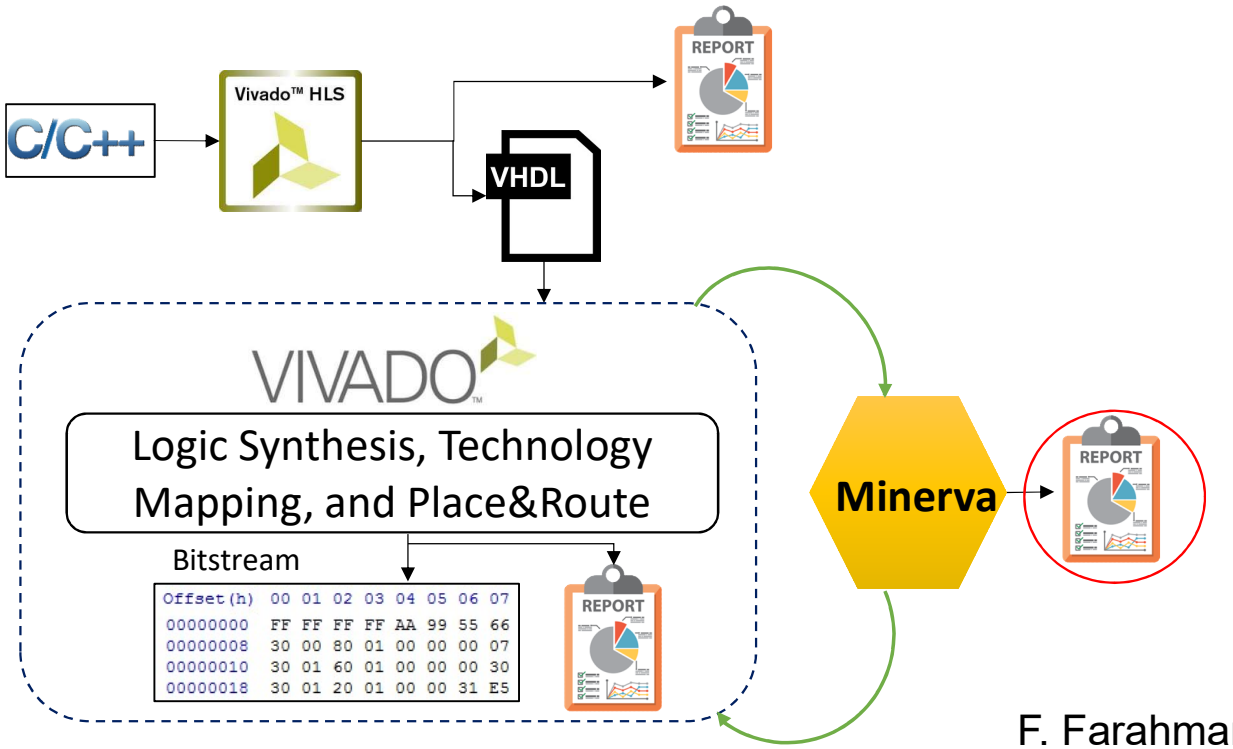


$$Minimum\ Clock = Target\ Clock - WNS$$

Impact of Optimization on the Outcome of HLS: Average of 10 numbers



Using Minerva to Tackle Finding Optimal Results



Extremely time consuming!

F. Farahmand et al., “Minerva: Automated hardware optimization tool,”
in ReConFig, 2017.

Studied Applications

- Benchmarks: **Machsuit, S2CBench, CHStone, Rosetta, and xfOpenCV**
- Development suites:

  Vivado and Vivado HLS version 2017.2
- Total number of applications: **90**
- Applications' types:

  Machine learning, Image/Video Processing Cryptography, and Mathematical

Data Collection

Feature Category	Brief Description	# features
Performance	Requested clock period, estimated clock period by HLS, Uncertainty	3
Resources	Utilization and availability of LUT, FF, DSP, and BRAM	36
Logic and arithmetic operation	Bitwidth/resource statistics of operations	29
Memory	Number of memory words/banks/bits;	2
Multiplexer	Multiplexer input size/bitwidth	2

- Total features: 72
- Data set: 60% for training, 20% for Validation, and 20% for testing (unseen data)