DATE 21
DESIGN, AUTOMATION & TEST IN EUROPE
01 – 05 February 2021 · virtual conference
The European Event for Electronic System Design & Test

Correlated Multi-objective Multi-fidelity Optimization for HLS Directives Design

Qi Sun[1], Tinghuan Chen[1], Siting Liu[1], Jin Miao[2], Jianli Chen[3], Hao Yu[4], Bei Yu[1]

[1]The Chinese University of Hong Kong
[2]Synopsys [3]Fudan University [4]SUSTech

Background

High-level synthesis (HLS)

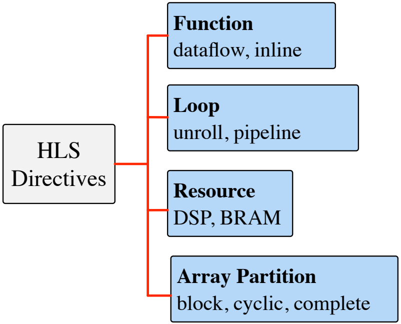
- Translate high-level programming languages (*e.g.*, C/C++) to low-level hardware description languages (HDLs).
- Under the guidance of the HLS directives (pragmas).
- Same high-level descriptions, different HLS directives $\rightarrow$ different hardware implementations.
- For each application, a group of HLS directives is represented as a configuration vector $\boldsymbol{x}$.

```
comp(int in[10], int out[10]):
  #pragma HLS INLINE={ON, OFF}
  for(i = 0; i < 10; i ++) {
    #pragma HLS UNROLL factor={2,5,10}
    in[i] = out[i];
  }
```

Pseudo-codes and HLS directives. The directives are in red. Each directive has some factors, *e.g.*, 2, 5, and 10.

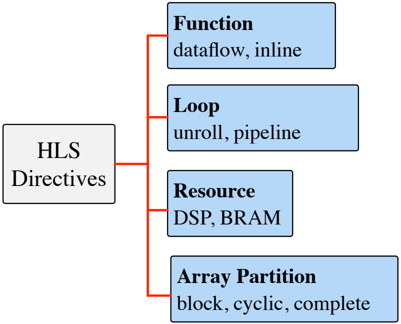
Background

Various types of directives

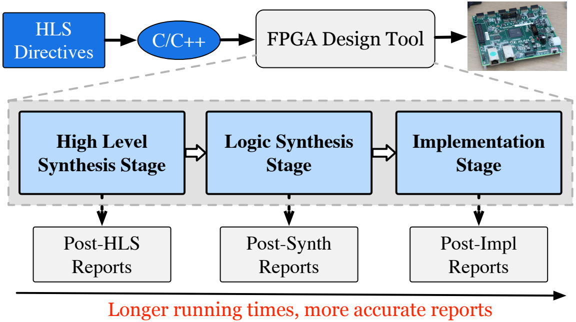


Background

Various types of directives

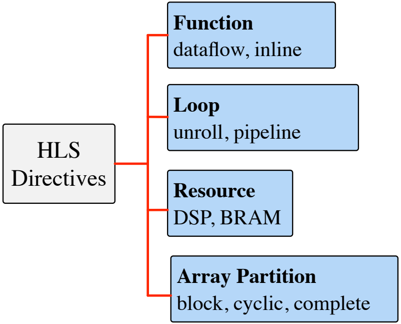


Design flow

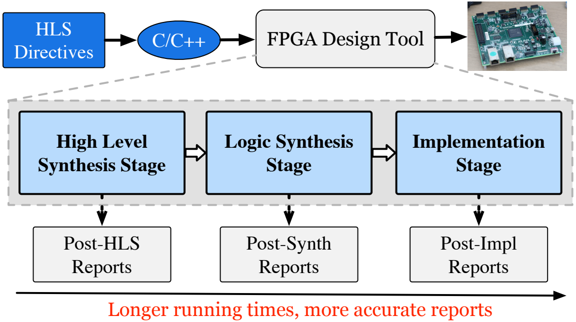


Background

Various types of directives



Design flow



Multiple conflicting design objectives (three fidelities)

- delay, power consumption, and resource consumption

Background

Pareto optimality – find some Pareto-optimal points

- 3 objective functions. $f_m : \mathcal{X} \to \mathbb{R}$, for $m = 1, 2, 3$.

Background

Pareto optimality – find some Pareto-optimal points

- 3 objective functions. $f_m : \mathcal{X} \to \mathbb{R}$, for $m = 1, 2, 3$.

- A value point $y = [\, f_1(x), f_2(x), f_3(x) \,]$, in the value space $\mathcal{Y}$.

Background

Pareto optimality – find some Pareto-optimal points

- 3 objective functions. $f_m : \mathcal{X} \rightarrow \mathbb{R}$, for $m = 1, 2, 3$.
- A value point $y = [\, f_1(x), f_2(x), f_3(x) \,]$, in the value space $\mathcal{Y}$.
- For $y_i, y_j \in \mathcal{Y}$, y_i dominates y_j when $y_{i,m} \geq y_{j,m}$, for $\forall\, m \in \{1, 2, 3\}$, represented as $y_i \succeq y_j$.

Background

Pareto optimality – find some Pareto-optimal points

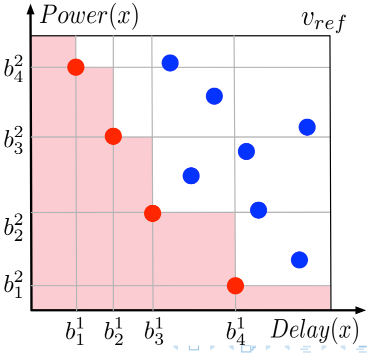
- 3 objective functions. $f_m : \mathcal{X} \rightarrow \mathbb{R}$, for $m = 1, 2, 3$.
- A value point $y = [\, f_1(x), f_2(x), f_3(x) \,]$, in the value space $\mathcal{Y}$.
- For $y_i, y_j \in \mathcal{Y}$, y_i dominates y_j when $y_{i,m} \geq y_{j,m}$, for $\forall\ m \in \{1, 2, 3\}$, represented as $y_i \succeq y_j$.
- The non-dominated points are called Pareto-Optimal Set, $\mathcal{P}(\mathcal{Y}) \in \mathcal{Y}$.

Background

Pareto optimality – find some Pareto-optimal points

- 3 objective functions. $f_m : \mathcal{X} \to \mathbb{R}$, for $m = 1, 2, 3$.
- A value point $y = [\, f_1(x), f_2(x), f_3(x)\,]$, in the value space $\mathcal{Y}$.
- For $y_i, y_j \in \mathcal{Y}$, y_i dominates y_j when $y_{i,m} \geq y_{j,m}$, for $\forall\, m \in \{1, 2, 3\}$, represented as $y_i \succeq y_j$.
- The non-dominated points are called Pareto-Optimal Set, $\mathcal{P}(\mathcal{Y}) \in \mathcal{Y}$.

- Blank cells are dominated
- Pareto hyper-volume $\mathrm{PV}_{v_{ref}}(\mathcal{P}(\mathcal{Y}))$.



Background

Target

- Find the **Pareto-optimal** points in HLS design problem

Background

Target

- Find the **Pareto-optimal** points in HLS design problem

Challenges

- Hard to predict the performance values according to the directives
- Hard to characterize the complicated relationships between the multiple objectives
- Hard to balance the consumption of running time and accuracy of results

Background

Target

- Find the **Pareto-optimal** points in HLS design problem

Challenges

- Hard to predict the performance values according to the directives
- Hard to characterize the complicated relationships between the multiple objectives
- Hard to balance the consumption of running time and accuracy of results

Requirements

- Develop a flexible and general method
- Strike a balance between optimization workloads and accuracy of results
- Able to characterize the complicated relationships between the HLS directives and multiple objectives

Our Solution

Optimization strategy

- Bayesian optimization
- Acquisition function: expected improvement

Multi-fidelity model

- Non-linear Gaussian process model

Multi-objective model

- Pareto learning
- Correlated Gaussian prorcess model

Multi-Fidelity Model

Traditional linear correlation model

$$f_m^h(\boldsymbol{x}) = \rho^h \times f_m^l(\boldsymbol{x}) + f_m^e(\boldsymbol{x}).$$

- ρ^h: a scaling factor. $f_m^e(\boldsymbol{x})$: error term.

Our non-linear correlation model

The reports of the low fidelity are concatenated as part of the inputs to the next high fidelity.

$$f_m^h(\boldsymbol{x}) = z_m^h(\, f_m^l(\boldsymbol{x}), \boldsymbol{x}\ ) + f_m^e(\boldsymbol{x}).$$

- $z_m^h(\cdot)$: correlation term, modelled by a GP model.

Multi-Objective Model – Pareto Learning

Acquisition function: expected improvement of Pareto hyper-volume

- At step $t+1$ of Bayesian optimization, we already have data set $D = \{\boldsymbol{x}_s, \boldsymbol{y}_s\}_{s=1}^{t}$, with $\mathcal{P}(\mathcal{Y}) = \{\boldsymbol{y}_s\}_{s=1}^{t}$. Sample a new point $\boldsymbol{x}_{t+1}$, the predicted value is $\boldsymbol{y}(\boldsymbol{x}_{t+1})$.

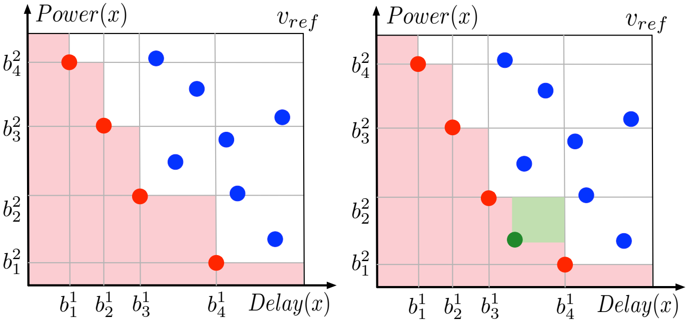
$$\mathrm{EIPV}(\boldsymbol{x}_{t+1}|\mathcal{D}) = \mathbb{E}_{p(\boldsymbol{y}(\boldsymbol{x}_{t+1})|\mathcal{D})}\left[\mathrm{PV}_{\boldsymbol{v}_{ref}}\left(\mathcal{P}(\mathcal{Y} \cup \boldsymbol{y}(\boldsymbol{x}_{t+1}))\right) - \mathrm{PV}_{\boldsymbol{v}_{ref}}\left(\mathcal{P}(\mathcal{Y})\right)\right].$$

Multi-Objective Model – Pareto Learning

Acquisition function: expected improvement of Pareto hyper-volume

- At step $t+1$ of Bayesian optimization, we already have data set $D = \{\boldsymbol{x}_s, \boldsymbol{y}_s\}_{s=1}^{t}$, with $\mathcal{P}(\mathcal{Y}) = \{\boldsymbol{y}_s\}_{s=1}^{t}$. Sample a new point $\boldsymbol{x}_{t+1}$, the predicted value is $\boldsymbol{y}(\boldsymbol{x}_{t+1})$.

$$\mathrm{EIPV}(\boldsymbol{x}_{t+1}|\mathcal{D}) = \mathbb{E}_{p(\boldsymbol{y}(\boldsymbol{x}_{t+1})|\mathcal{D})}\left[\mathrm{PV}_{\boldsymbol{v}_{ref}}\left(\mathcal{P}(\mathcal{Y} \cup \boldsymbol{y}(\boldsymbol{x}_{t+1}))\right) - \mathrm{PV}_{\boldsymbol{v}_{ref}}\left(\mathcal{P}(\mathcal{Y})\right)\right].$$



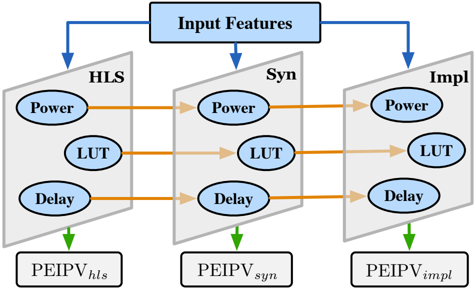
Combined Model

- Two dimensions: one for the multi-objective functions, one for the multi-fidelities.
- Augment acquisition function:

$$\mathrm{PEIPV}_i(\boldsymbol{x}_{t+1}|\mathcal{D}) = \mathrm{EIPV}_i(\boldsymbol{x}_{t+1}|\mathcal{D}) \cdot \frac{T_{impl}}{T_i}, i \in \{hls, syn, impl\},$$

$$\max_i \mathrm{PEIPV}_i, \; i \in \{hls, syn, impl\}$$

- Select the largest one, and run the compilation flow to that fidelity.



Experiments and Results

Experimental settings

- 5 traditional benchmarks, 1 DNN benchmark
- All HLS code are compiled via Vivado HLS to get the reports (for validation of results of various algorithms).

Quality metric – average distance to reference set (ADRS)

- Γ reference set (real Pareto set).
- Ω learned Pareto set.

$$ADRS(\Gamma, \Omega) = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \min_{\omega \in \Omega} f(\gamma, \omega)$$

Results

All algorithms use the same input features.

- Bayesian methods: 8 initial samples, at most 40 optimization steps.
- Other methods, each training set has 48 points.

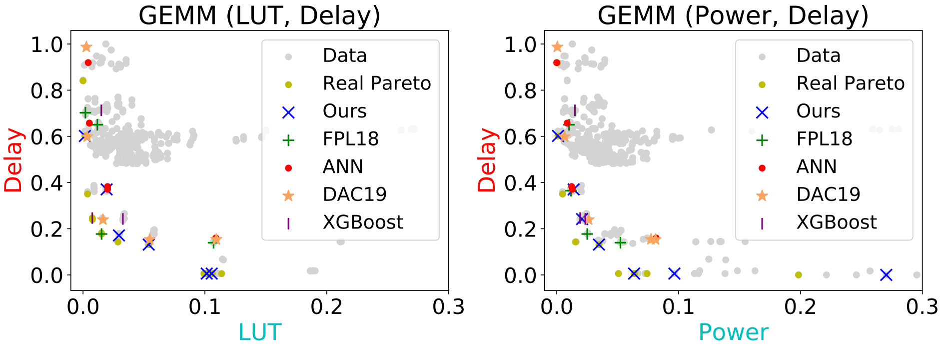
Table: Normalized Experimental Results

Model	Normalized ADRS					Normalized Standard Deviation of ADRS					Normalized Overall Running Time				
	Ours	FPL18	ANN	BT	DAC19	Ours	FPL18	ANN	BT	DAC19	Ours	FPL18	ANN	BT	DAC19
GEMM	**0.27**	0.50	1.00	0.65	1.08	**0.12**	0.46	1.00	0.37	0.90	**0.68**	0.83	1.00	1.00	7.00
iSmart2	**0.65**	0.68	1.00	1.28	1.49	**0.20**	0.75	1.00	1.10	1.24	**0.42**	0.88	1.00	1.00	7.00
SORT_RADIX	**0.64**	0.72	1.00	1.09	0.94	**0.48**	0.57	1.00	1.72	2.28	**0.34**	0.47	1.00	1.00	7.00
SPMV_ELLPACK	**0.19**	0.47	1.00	0.22	1.21	0.09	0.24	1.00	**0.06**	0.99	0.65	**0.42**	1.00	1.00	7.00
SPMV_CRS	**0.22**	0.29	1.00	2.09	1.15	**0.03**	0.26	1.00	2.09	1.52	**0.72**	0.90	1.00	1.00	7.00
STENCIL3D	**0.39**	0.41	1.00	0.40	0.41	0.03	0.57	1.00	**0.00**	0.05	0.44	**0.41**	1.00	1.00	7.00
Average	**0.39**	0.51	1.00	0.96	1.05	**0.16**	0.47	1.00	0.89	1.16	**0.54**	0.65	1.00	1.00	7.00

Example – GEMM

Directives

- INLINE, PIPELINE, UNROLL, Mul_LUT, DSP48, ARRAY_PARTITION, BRAM.



Thank you!