IT-DSE: Invariance Risk Minimized Transfer Microarchitecture Design Space Exploration

Ziyang Yu[1], Chen Bai[1], Shoubo Hu[2], Ran Chen[2], Taohai He[3], Mingxuan Yuan[2], Bei Yu[1], Martin Wong[1]

[1]The Chinese University of Hong Kong
[2]Huawei Noah's Ark Lab
[3]HiSilicon

Background and Motivation

- Microarchitecture design space exploration (DSE) determines the detailed structures of a microprocessor to obtain desired performance power and area (PPA) values.
- Complexity of design spaces and large evaluation cost are two major challenges.
- Transferring from historical explored tasks to new task reduces the exploration cost.
- Extract the relationship that is approximately invariant among different historical source design tasks, even equipped with different design spaces.

Typical Microprocessor Components

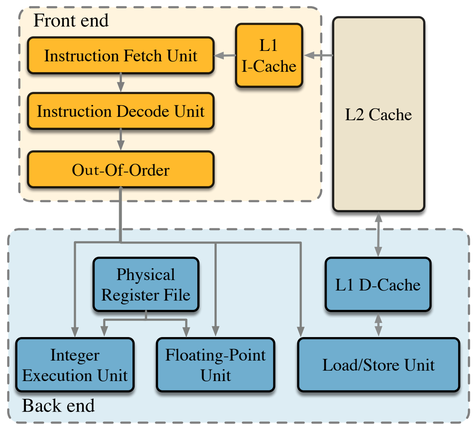

Figure 1. Visualization of typical microarchitecture pipeline.

Invariant Risk Minimization (IRM)

Prediction function $f' : \mathcal{X} \to \mathcal{Y} \Rightarrow \phi_{\boldsymbol{u}}(\cdot)$: feature extractor; $h_{\boldsymbol{w}}(\cdot)$: regressor.

$\boldsymbol{u}$, $\boldsymbol{w}$: parameter.

IRM:

$$\begin{aligned} \min_{\boldsymbol{u},\boldsymbol{w}} \quad & \sum_{s=1}^{S} \mathcal{R}^s(\boldsymbol{u},\boldsymbol{w}), \\ \text{s.t.} \quad & \boldsymbol{w} \in \operatorname*{argmin}_{\boldsymbol{w}^s} \mathcal{R}^s(\boldsymbol{u},\boldsymbol{w}^s), \end{aligned} \tag{1}$$

$\mathcal{R}^s(\boldsymbol{u},\boldsymbol{w}) = \mathbb{E}_{X^s,Y^s}[\mathcal{L}(f'(X^s),Y^s)]$: risk under s-th source task.

$\mathcal{L}(\cdot)$: the loss function.

Simplified IRMv1:

$$\min_{\boldsymbol{u},\boldsymbol{w}} \sum_{s=1}^{S} \mathcal{R}^s(\boldsymbol{u},\boldsymbol{w}) + \lambda ||\nabla_w \mathcal{R}^s(\boldsymbol{u},\boldsymbol{w})||^2, \tag{2}$$

Pareto Optimality

For a multi-objective minimization problem with M objectives, $\boldsymbol{x}_1$ is deemed to dominate solution $\boldsymbol{x}_2$ ($\boldsymbol{x}_1 \succeq \boldsymbol{x}_2$) if

$$\begin{aligned} f_m(\boldsymbol{x}_1) \leqslant f_m(\boldsymbol{x}_2), \forall m \in \{1,...,M\}, \\ f_k(\boldsymbol{x}_1) < f_k(\boldsymbol{x}_2), \exists k \in \{1,...,M\}. \end{aligned} \tag{3}$$

Pareto-optimal set: The collection of solutions that remain non-dominated by others.

Design Tasks

- Source task: dataset $\mathcal{D}^s$ containing n_s parameter vectors $X^s = [\boldsymbol{x}_1^s, \ldots, \boldsymbol{x}_{n_s}^s]^\top \in \mathbb{R}^{n_s \times d_s}$ and the evaluated PPA vectors $Y^s = [\boldsymbol{y}_1^s, \ldots, \boldsymbol{y}_{n_s}^s] \in \mathbb{R}^{n_s \times 3}$.
- Target task: contains only the set of legal parameter configuration vectors $X^t = [\boldsymbol{x}_1^t, \ldots, \boldsymbol{x}_{n_t}^t]^\top \in \mathbb{R}^{n_t \times d_t}$.

Microarchitecture Transferring Design Space Exploration

Given S source tasks with explored datasets $\{\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^S\}$ and the target sample $\mathcal{D}^t$, the objective is to utilize the information from historical source tasks and improve the efficiency of finding a series of microprocessor design configurations X in the target task that forms the Pareto optimality among the associated subset of $Y \subset \mathcal{Y}$, so that $X = \{\boldsymbol{x} | \boldsymbol{x} \succeq \boldsymbol{x}', \forall \boldsymbol{x}' \in X^t\}, Y = \{f(\boldsymbol{x}) | X \in X^t\}$.

Overall flow

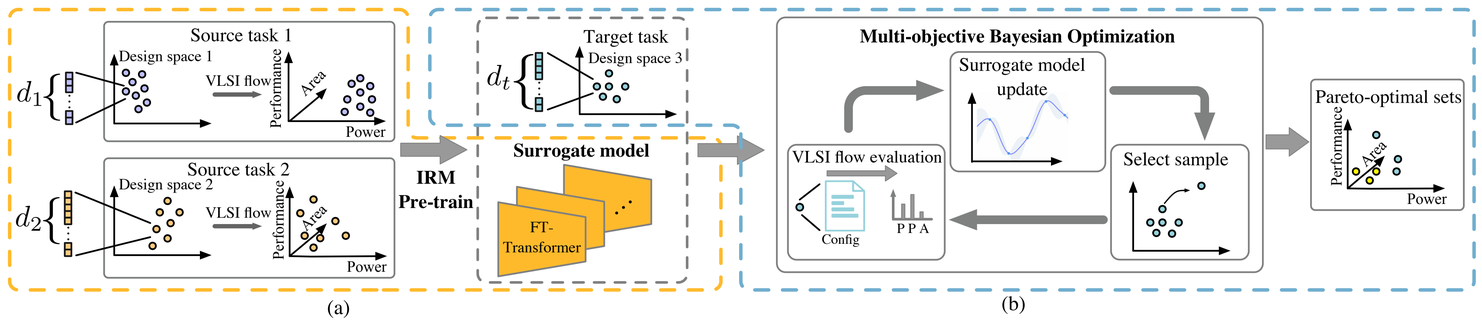

Figure 2. Workflow of proposed IT-DSE.

- Accommodate varying design spaces: FT-Transformer ensemble
- Knowledge fusion and warm start new task: pre-training with invariant risk minimization paradigm

Customized Surrogate Model with Transformer

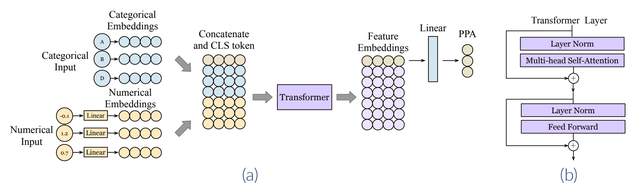

Figure 3. (a) FT-Transformer architecture. Firstly, Feature Tokenizer transforms input parameters into embeddings. The embedding is then processed by the Transformer module. (b) One Transformer layer.

$$\boldsymbol{t}_i^{num} = x_i^{num} \cdot \boldsymbol{w}_i^{num} + b_i^{num}, \tag{4}$$

$$\boldsymbol{t}_j^{cat} = \boldsymbol{e}_j^\top \boldsymbol{w}_j^{cat} + b_j^{cat}, \tag{5}$$

$$\boldsymbol{T} = \text{stack}([\texttt{CLS}], \boldsymbol{t}_1^{num}, \ldots \boldsymbol{t}_{d_n}^{num}, \boldsymbol{t}_1^{cat}, \ldots, \boldsymbol{t}_{d_c}^{cat}). \tag{6}$$

Pre-training on Multiple Source Tasks:

Source task 1: $\boldsymbol{x}^1 = [x_1, x_2]$, source task 2: $\boldsymbol{x}^2 = [x_2, x_3, x_4]$,

Embedding weight: $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \boldsymbol{w}_4]^\top \in \mathbb{R}^{4 \times d}$,

Embedding bias: $\boldsymbol{B} = [b_1, b_2, b_3, b_4]^\top \in \mathbb{R}^{4 \times d}$,

Transformed embeddings:

$\boldsymbol{T}^1 = \text{stack}([\texttt{CLS}], \boldsymbol{t}_1, \boldsymbol{t}_2)$, $\boldsymbol{T}^2 = \text{stack}([\texttt{CLS}], \boldsymbol{t}_2, \boldsymbol{t}_3, \boldsymbol{t}_4)$.

Bayesian Invariant Risk Minimized Feature Extraction

Focusing more on task-invariant features.

For s-th source task, original data: $\mathcal{D}^s = \{X^s, Y^s\} = \{\boldsymbol{x}_i^s, \boldsymbol{y}_i^s\}_{i=1}^{n_s}$,

transferred data: $\mathcal{T}^s = \{\phi_{\boldsymbol{u}}(\boldsymbol{x}_i^s), \boldsymbol{y}_i^s\}_{i=1}^{n_s}$,

collection of transformed datasets of all S source tasks: $\mathcal{T}^c = \cup_{s=1}^S \mathcal{T}^s$.

Objective for Bayesian IRM:

$$\begin{aligned} \max_{\boldsymbol{u}} \sum_s & \mathbb{E}_{g_{\boldsymbol{u}}(\boldsymbol{w})}[\ln p(\mathcal{T}^s|\boldsymbol{w},\boldsymbol{u})] + \lambda \Big(\mathbb{E}_{g_{\boldsymbol{u}}(\boldsymbol{w})}[\ln p(\mathcal{T}^s|\boldsymbol{w},\boldsymbol{u})] \\ & - \mathbb{E}_{g_{\boldsymbol{u}}^s(\boldsymbol{w}^s)}[\ln p(\mathcal{T}^s|\boldsymbol{w}^s,\boldsymbol{u})] \Big). \end{aligned} \tag{7}$$

$g_{\boldsymbol{u}}(\boldsymbol{w}) \approx p(\boldsymbol{w}|\mathcal{T}^c)$, $g_{\boldsymbol{u}}^s(\boldsymbol{w}^s) \approx p(\boldsymbol{w}^s|\mathcal{T}^s)$.

First term: Encouraging $\boldsymbol{u}$ to retain information for data distribution fitting.

Second term: Requires transformed data distribution to be stable among different tasks with feature extractor $\phi_{\boldsymbol{u}}(\cdot)$.

Multi-objective Bayesian Optimization

For new point $\boldsymbol{x}_*^t$, posterior distribution $p(\boldsymbol{y}_*^t|\boldsymbol{x}_*^t)$ is approximated as a Gaussian distributions:

$$p(\boldsymbol{y}_*^t|\boldsymbol{x}_*^t) = \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{y}_*^t|f'(X^t)}, \boldsymbol{\sigma}^2_{\boldsymbol{y}_*^t|f'(X^t)}\right), \tag{8}$$

- FT-Transformer Ensemble as Surrogate Model.
- Pareto EHVI Acquisition Function:
  $EHVI(\hat{\boldsymbol{x}}) = \mathbb{E}_{P(f'(\hat{\boldsymbol{x}}_*)|\mathcal{D})}\left(HVI(f'(\hat{\boldsymbol{x}})|\mathcal{P}(\boldsymbol{y}), \boldsymbol{y}_{ref})\right)$

Transfer Design Space Exploration Experiment Setup

Table 1. Statistics of our microarchitecture design space

Module	# Linear	#Pow	# Categorical	# Combinations
IFU	8	4	0	$\sim 4 \times 10^8$
OoO	11	0	0	$\sim 5 \times 10^9$
IEX	12	0	3	$\sim 1 \times 10^9$
FSU	5	0	0	$\sim 2 \times 10^3$
LSU	6	3	0	$\sim 1 \times 10^7$
L2C	1	2	0	$\sim 8 \times 10^2$
Overall	43	9	3	$\sim 3 \times 10^{40}$

- 55 parameters with more than 3×10^{40} combinations.
- Task A, task B, and Task C share the same design space, for task D, we can only tune 53 parameters out of 55 parameters listed.
- Task A, task B, task C, and Task D contain 1237, 377, 1835 and 3453 evaluated samples, respectively.

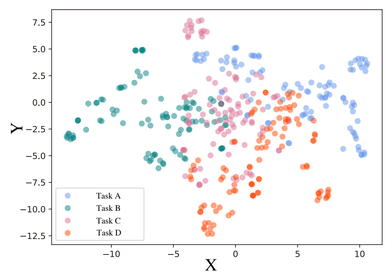

Figure 4. Data distribution for 4 tasks.

Evaluation Metrics:

Hypervolume:

$$HV_{\boldsymbol{y}_{ref}}(\mathcal{P}(\boldsymbol{y})) = \lambda_M(\cup_{\boldsymbol{y} \in \mathcal{P}(\boldsymbol{y})}[\boldsymbol{y}, \boldsymbol{y}_{ref}]), \tag{9}$$

Average distance to reference set (ADRS):

$$ADRS(\mathcal{P}^*, \mathcal{P}) = \frac{1}{|\mathcal{P}^*|} \sum_{\boldsymbol{\alpha} \in \mathcal{P}^*} \min_{\boldsymbol{\beta} \in \mathcal{P}} l_2(\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{10}$$

Transferring Performance

Table 2. Comparison of transfer performance in same design space

Methodologies	A, B → C		A, C → B		B, C → A	
	ADRS	HV	ADRS	HV	ADRS	HV
Ground Truth	0.0	0.0984	0.0	0.0684	0.0	0.0809
ANN-TL	0.069	0.0891	0.045	0.0643	0.031	0.0749
Deep-Ens	0.072	0.0840	0.066	0.0599	0.055	0.0727
ERM	0.080	0.0877	0.064	0.0629	0.043	0.0742
IRMv1	0.025	0.0938	0.023	0.0667	0.027	0.0766
Ours	**0.021**	**0.0944**	**0.017**	**0.0679**	**0.020**	**0.0796**

Table 3. Comparison of transfer performance in different design spaces

Methodologies	A, D → C		A, C → D	
	ADRS	HV	ADRS	HV
Ground Truth	0.0	0.0984	0.0	0.7792
w/o. Pre-train	0.0533	0.0857	0.0853	0.7465
w/o. Ensemble	0.0275	0.0892	0.0722	0.7531
w/o. IRM	0.0293	0.0910	0.0701	0.7569
w/. IRMv1	0.0232	0.0914	0.0687	0.7602
Ours	**0.0217**	**0.0924**	**0.0641**	**0.7624**

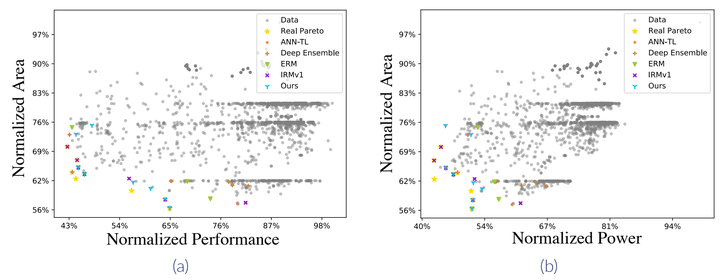

Figure 5. Pareto-optimal sets using source task data from same design space. Left: performance versus area; right: power versus area.