# ML²Tuner: Efficient Code Tuning via Multi-Level Machine Learning Models

JooHyoung Cha[1], Munyoung Lee[2], Jinse Kwon[2], Jubin Lee[3], Jemin Lee[2], Yongin Kwon[2]

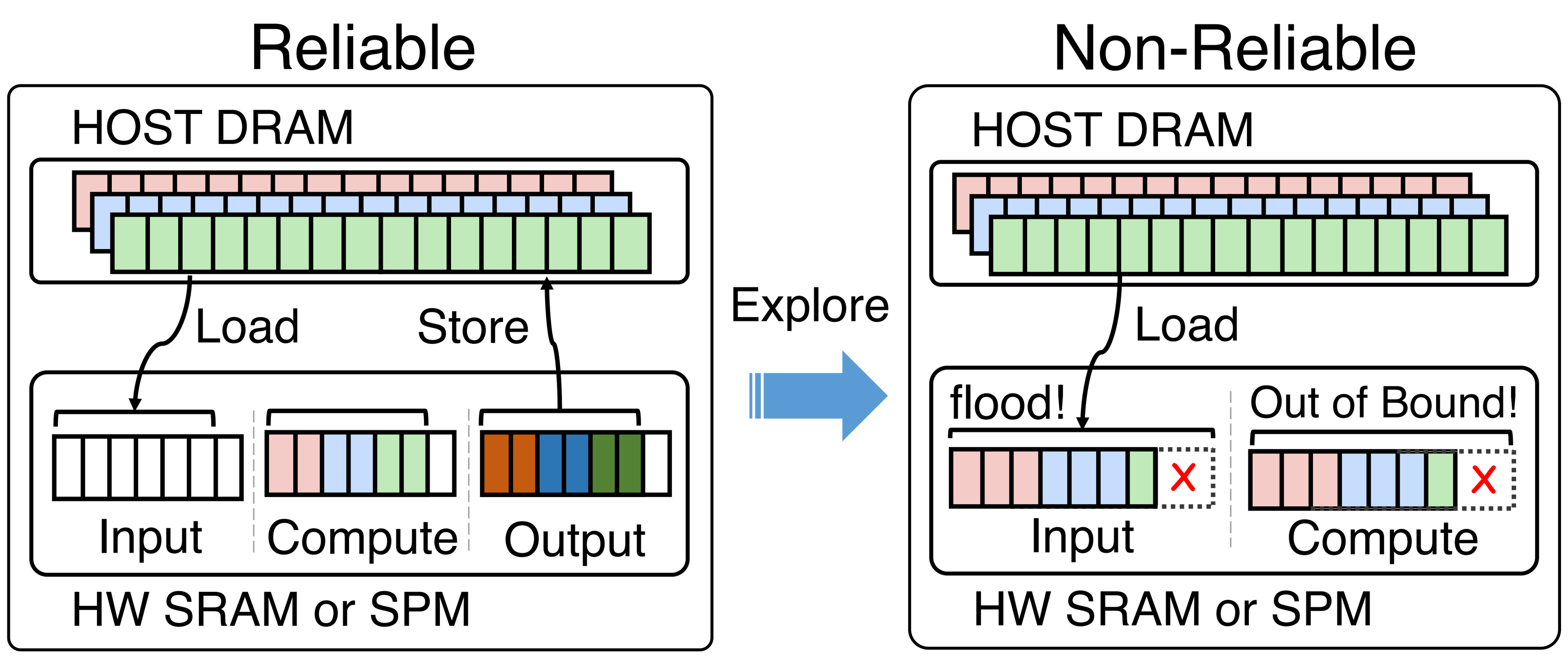
[1]University of Science of Technology, [2]Electronics and Telecommunications Research Institute, [3]Neubility

{jh.cha, munyounglee, kwonse, leejaymin, yongin.kwon}@etri.re.kr [1,2], jubin0927@gmail.com [3]

## Motivation

### Challenges in ML Code Auto-tuning

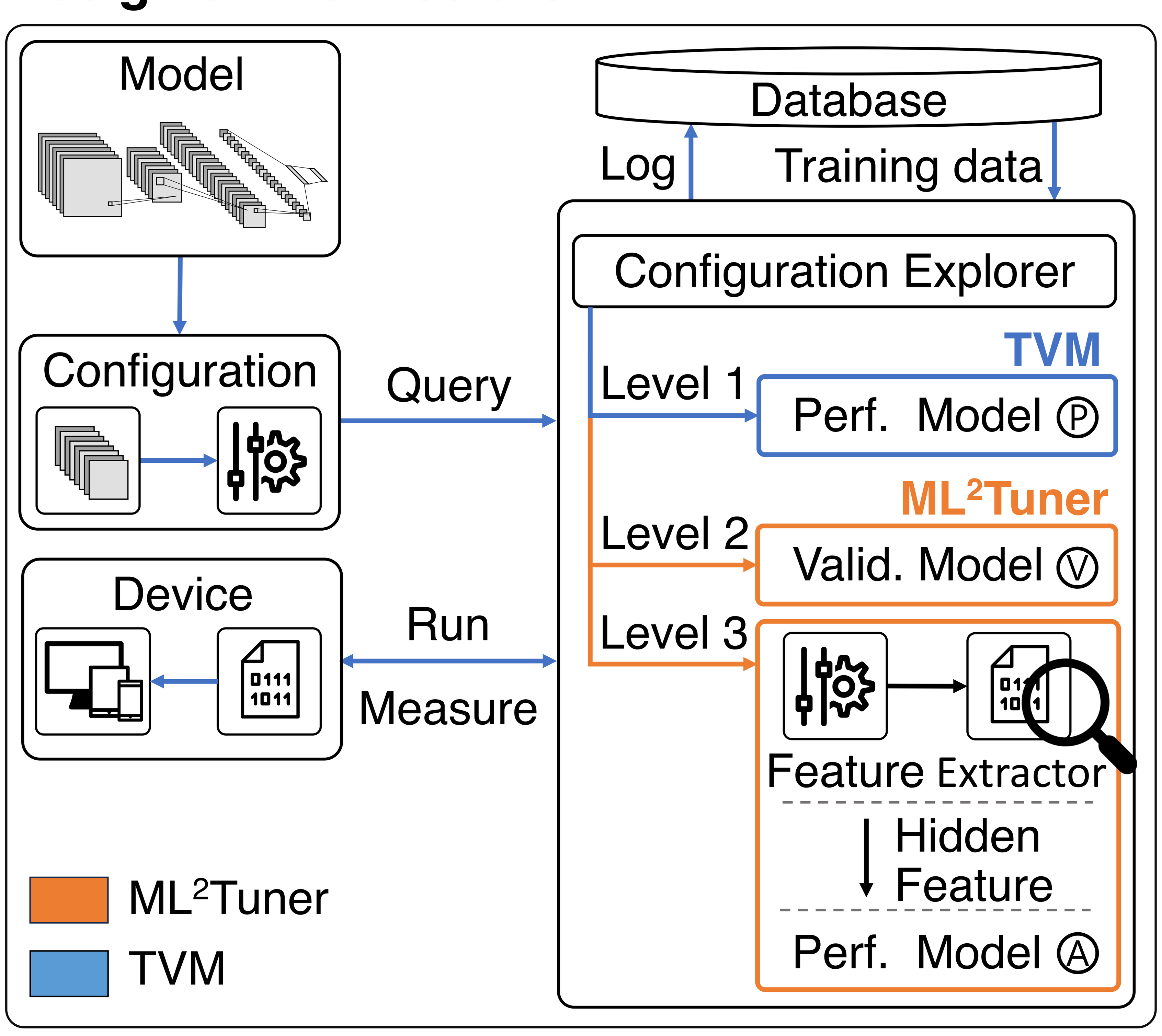
- Need to explore a vast search space
- Aggressive optimization increases the chance of runtime errors.




Example of Runtime Error in HW

## Proposed Method

### Design of Architecture




Ⓟ : Performance prediction

Ⓥ : Validity prediction

Ⓐ : Advanced performance prediction

**Level 1 :** Model Ⓟ

- Predict the Highest Performance from Configurations (such as Kernel, Layer Information)

**Level 2 :** Model Ⓥ

- Validate Configuration from Configurations (Based on Configuration used in Model P)

**Level 3 :** Model Ⓐ

- Predict the Highest Performance from Configurations (Kernel, Layer Information and w/ Hidden Feature)

Feature Extractor :

- Find and Collect Hidden Feature in **"Compiler"** (such as decisions, loop sizes and tiling strategy)

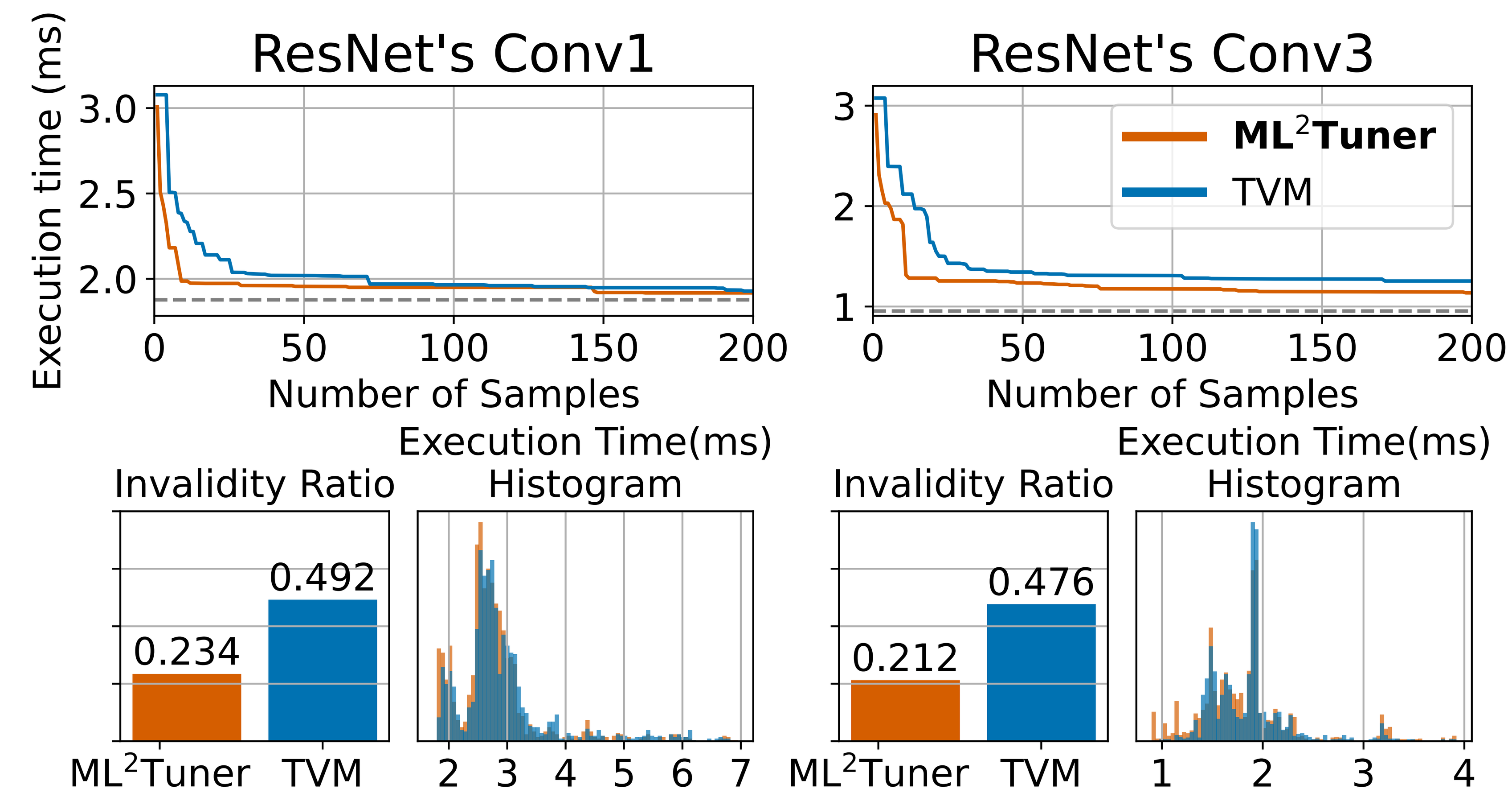
## Experimental

### Experimental Setup

- Hardware Platform : VTA on XILINX ZCU102
- Compiler : NEST-C (based on GLOW)
- Optimizer : XGBoost (v2.1.1)

---

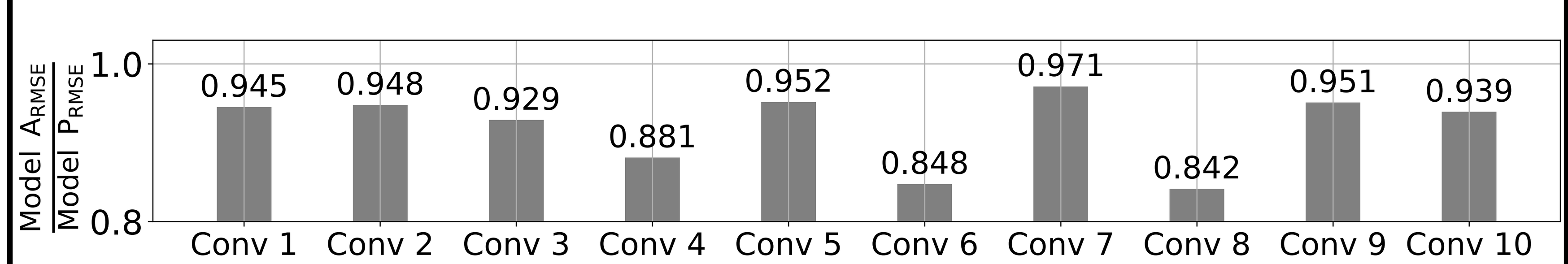
- Equivalent performance with 12.3% fewer Samples
- Reducing the invalidity ratio to 60.8%
- Provide increased opportunities for exploration




[ Result for Conv1 and Conv2 of ResNet18 on VTA ]

---

- Model A's average RMSE achives 91.6% of Model P's.




[ Evaluate Robustness Between Model P and Model A ]

## Conclusion and Discussion

### Our Contribution:

Error-Aware Search Refinement

### Future Research

**1. Validate Evaluation on Diverse Environment**

- Testing Diverse hardware to assess generalizability.
- Testing Algorithm to further refine the tuning process.

**2. Auto-identifying features from binaries**

- Study on improved accuracy and probabilistic error detection systems.

## References

Paper

Video

Drive