2019 Design Automation Conference

A Learning-Based Recommender System for Autotuning Design Flows of Industrial High-Performance Processors

Jihye Kwon*, Matthew M. Ziegler†, Luca P. Carloni*

*Department of Computer Science, Columbia University, New York, NY, USA

†IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

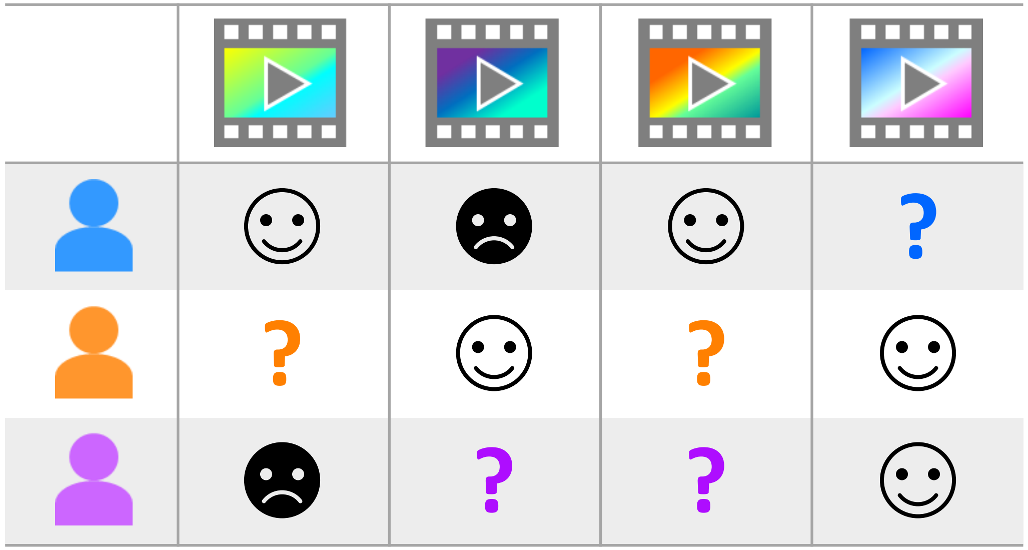
Recommender Systems

- Diverse application areas
  - ✓Movies, music, SNS posts, online shopping items, personalized tips
- Two main paradigms
  - ✓**Content** filtering

User profile / preferences

Item content / information

✓**Collaborative** filtering



A Design Flow of Industrial Processors

- VLSI design with CAD tools for **Logic Synthesis and Physical Design** (LSPD)
  - ✓Hierarchy of a high-performance processor:
    Chip → Processor core → Unit → **Macro** (10,000 – 100,000+ logic gates)

Macro specification ↓

LSPD parameter configuration (scenario) ↓

LSPD Phase-1: Logic synthesis → Physical placement → Clock tree synthesis → Post-placement optimization

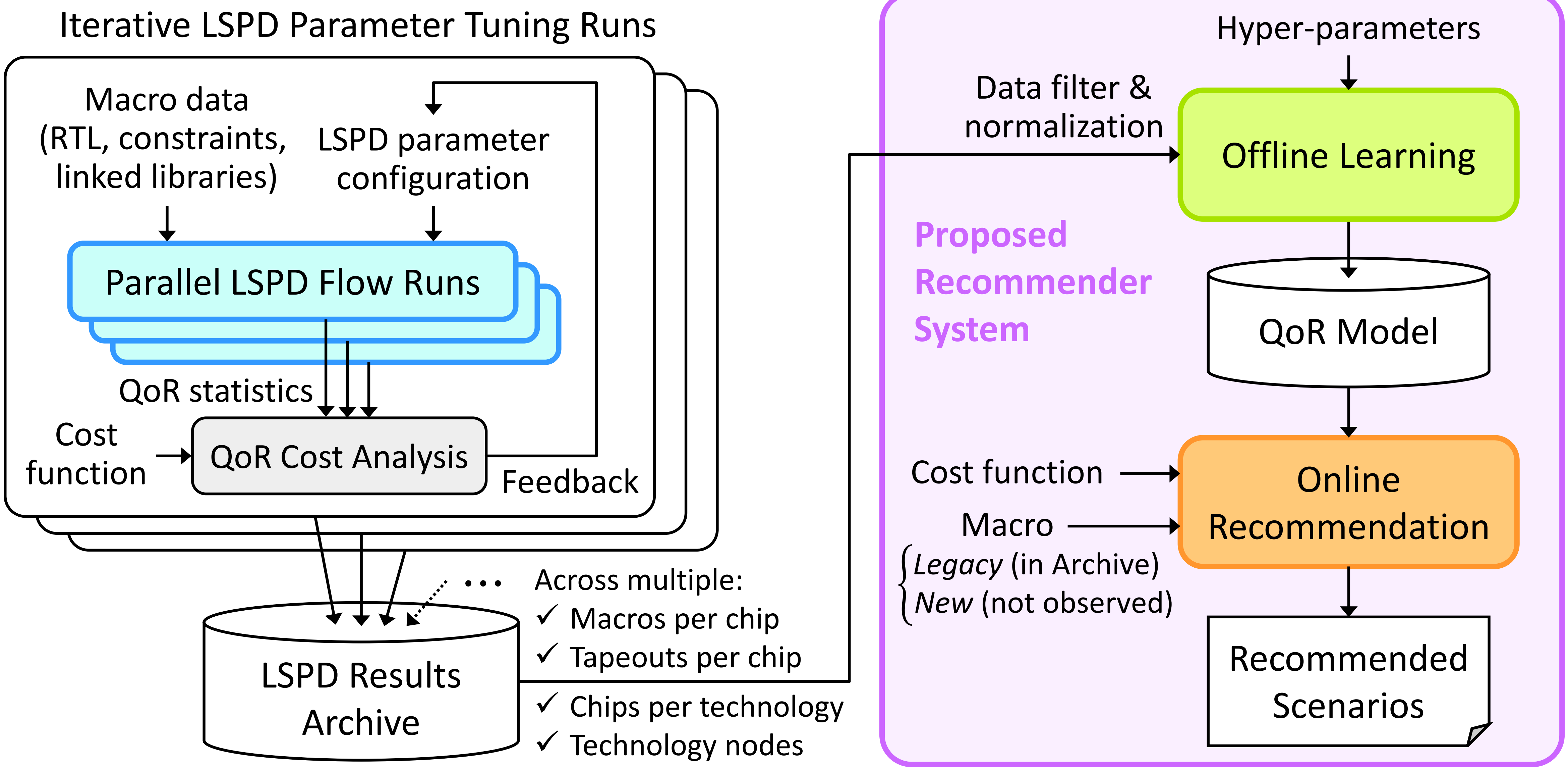
→ **LSPD Phase-2**: Routing → Post-route optimization

LSPD Phase-1 ↓ Estimated Quality-of-Result (QoR) e.g., timing, power, routability

Design-Space Exploration ◂······ Estimated Quality-of-Result (QoR)

LSPD Phase-2 ↓ Layout for fabrication

Overview of the Proposed System



Offline Learning Module

- The Archive contains sparse records of (Input: *Macro, Scenario*; Output: *QoR*)

Input		Output (normalized QoR)				
Macro	*Scenario*	*Slack 1*	*Slack 2*	*Slack 3*	*Power*	*Congestion*
	$1000\cdots0$	0.42	0.56	0.34	0.88	0.76
	$0110\cdots0$	0.89	0.87	0.68	0.75	0.60
m_1	$1010\cdots1$	0.92	0.84	0.56	0.65	0.54
	$0101\cdots1$	0.27	0.30	0.40	0.45	0.63
	⋮	⋮	⋮	⋮	⋮	⋮
	$1000\cdots0$	0.34	0.22	0.50	0.56	0.83
m_2	$1011\cdots0$	0.51	0.63	0.74	0.66	0.77
	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮

LSPD Results Archive

✓ *Macro*: RTL description, timing and physical constraints, linked libraries

✓ *Scenario*: configuration of binary meta-parameters for tuning LSPD flows

✓ *QoR*: normalized QoR scores for each of the d metrics (e.g., 5) for each macro

- Goal: to build a QoR prediction model F

$$F(Macro, Scenario) = (QoR_1, \cdots, QoR_d)$$

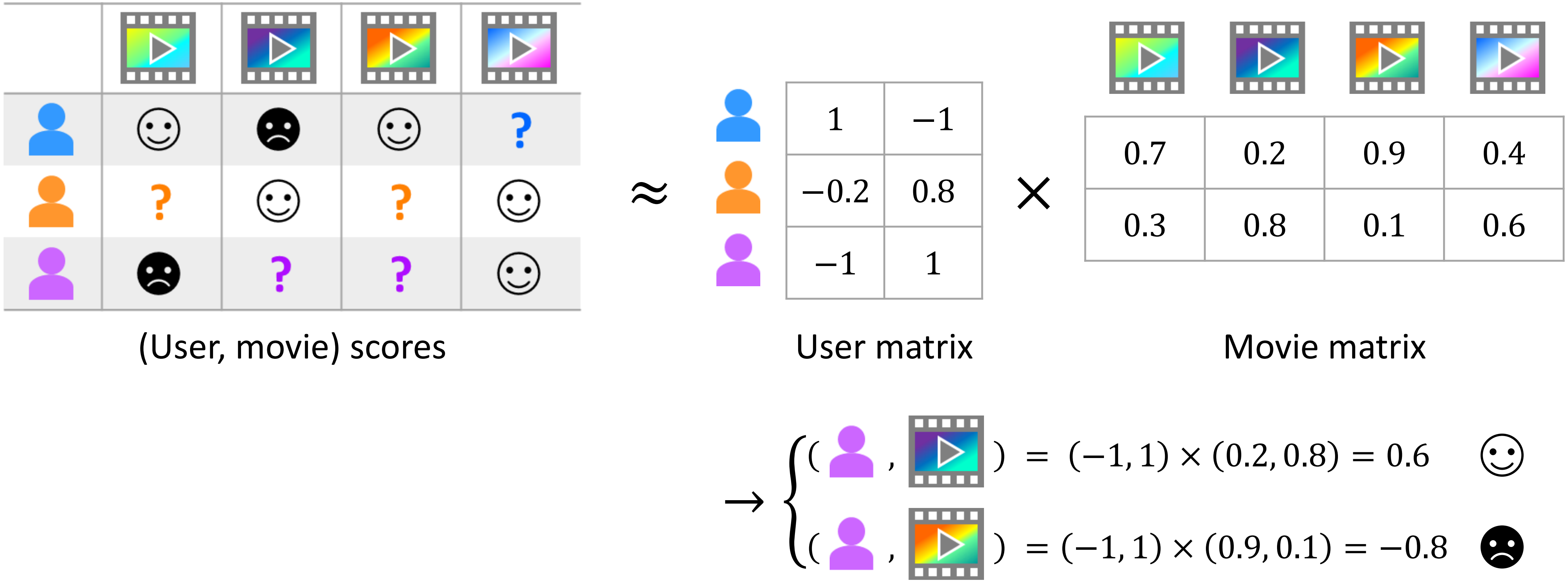
$\uparrow$ $\uparrow$

NOT easily available or quantifiable

→ A collaborative filtering approach

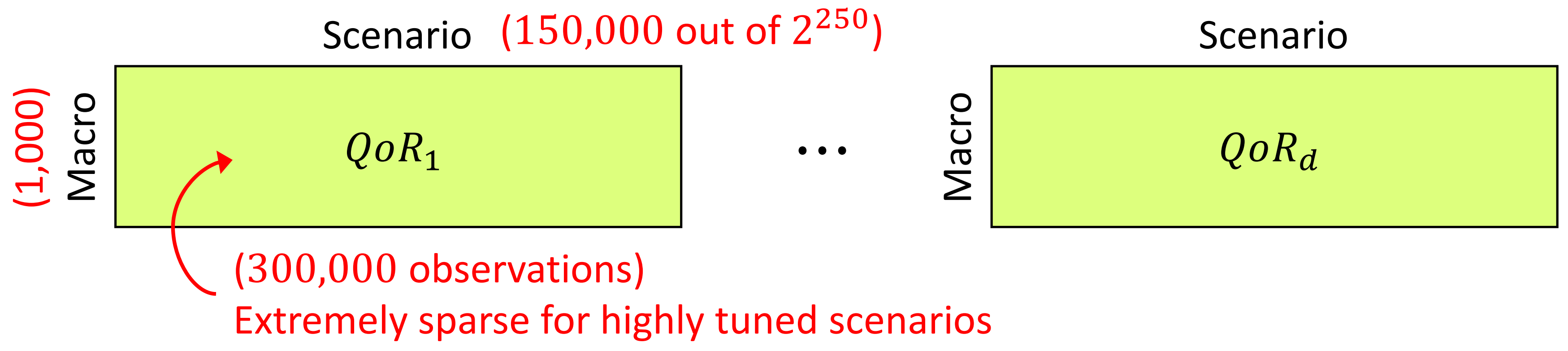
Offline Learning Module

- Goal: to build a QoR prediction model F
  - ✓ A collaborative filtering approach
    - E.g., Matrix factorization for a movie recommender system

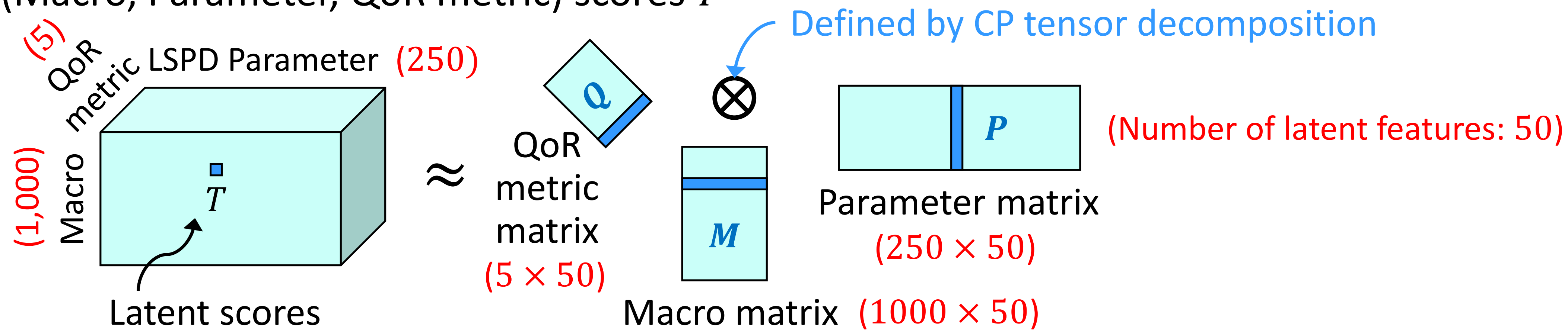


Offline Learning Module

- Goal: to build a QoR prediction model F
  - ✓ (Macro, Scenario) scores

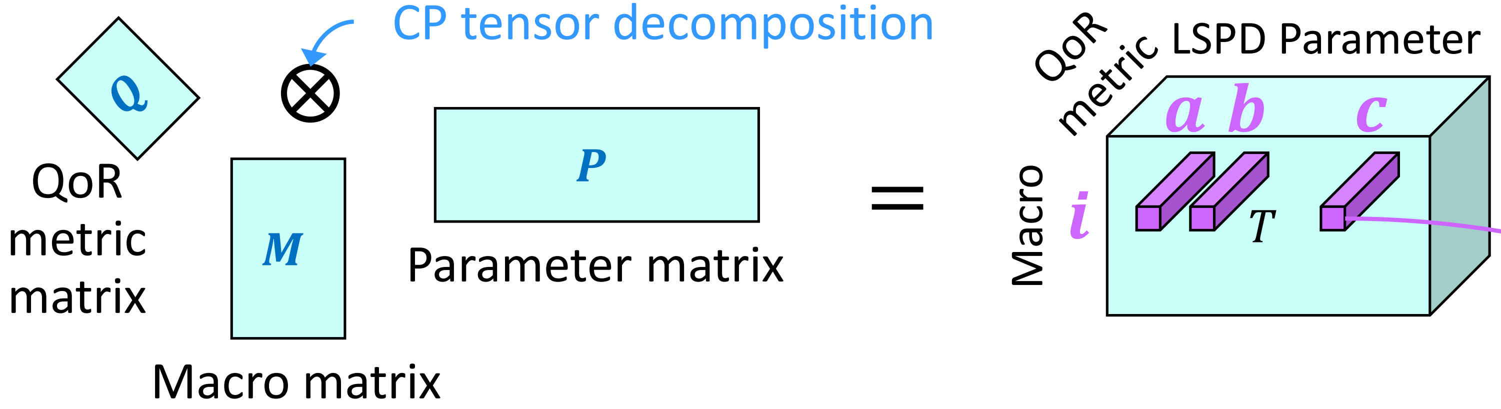


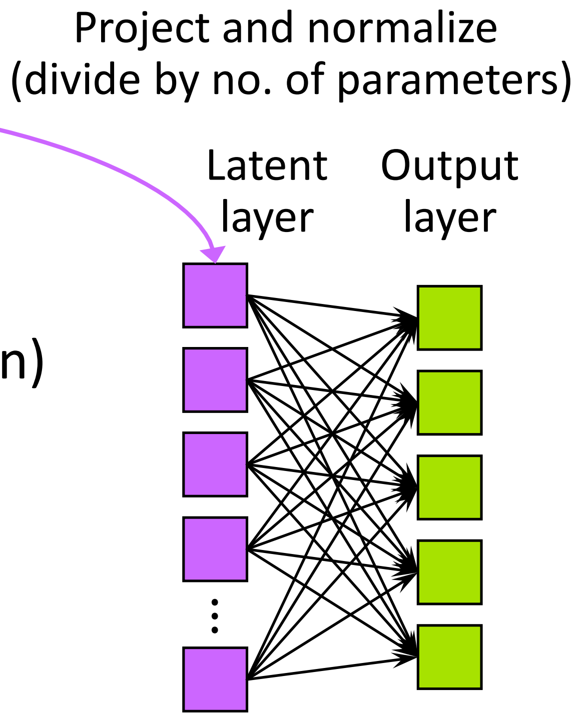
  - ✓ (Macro, Parameter, QoR metric) scores T



Offline Learning Module

- Goal: to build a QoR prediction model F
  - ✓ Macro matrix $\boldsymbol{M}$, Parameter matrix $\boldsymbol{P}$, QoR matrix $\boldsymbol{Q}$ → Latent tensor T



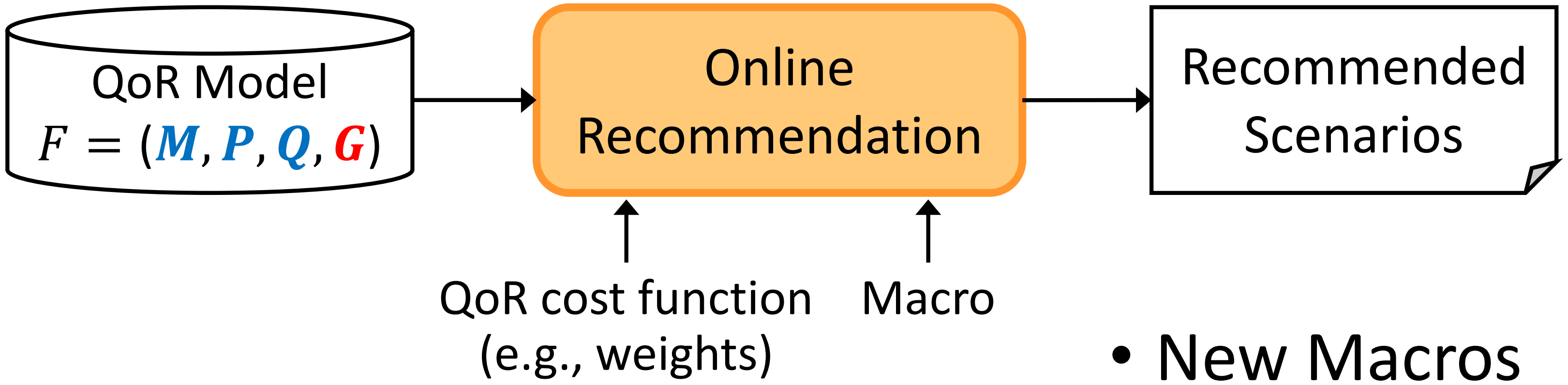


  - ✓ A single-layer perceptron network $\boldsymbol{G}$ for QoR prediction (regression)

$$F(\text{Macro } \boldsymbol{m_i}, \text{Scenario } \boldsymbol{p_a} \cdot \boldsymbol{p_b} \cdot \boldsymbol{p_c}; T) = \boldsymbol{G}(\boldsymbol{T_{ia:}}, \boldsymbol{T_{ib:}}, \boldsymbol{T_{ic:}})$$

  - ✓ Learn $(\boldsymbol{M}, \boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{G})$ by a stochastic gradient descent (SGD) method
    - ➢ Objective: to minimize the prediction error (RMSE)

Online Recommendation Module



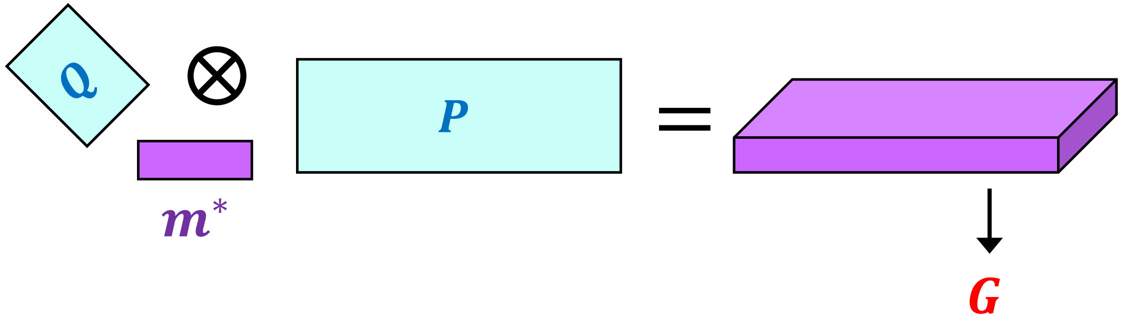
+ Designer's parameter settings

- Legacy Macros
  - ✓ Target macro m_i in the archive
  - ✓ Use $F = (\boldsymbol{M}[\boldsymbol{i}], \boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{G})$

    for making an inference in minutes

    instead of applying an LSPD flow taking hours

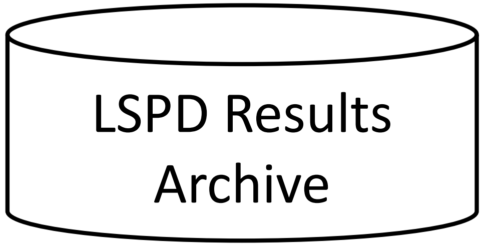
- New Macros
  - ✓ Sample LSPD results for a new macro
  - ✓ Train $F^* = (\boldsymbol{m}^*, \boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{G})$



with $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{G}$ fixed (to learn $\boldsymbol{m}^*$)

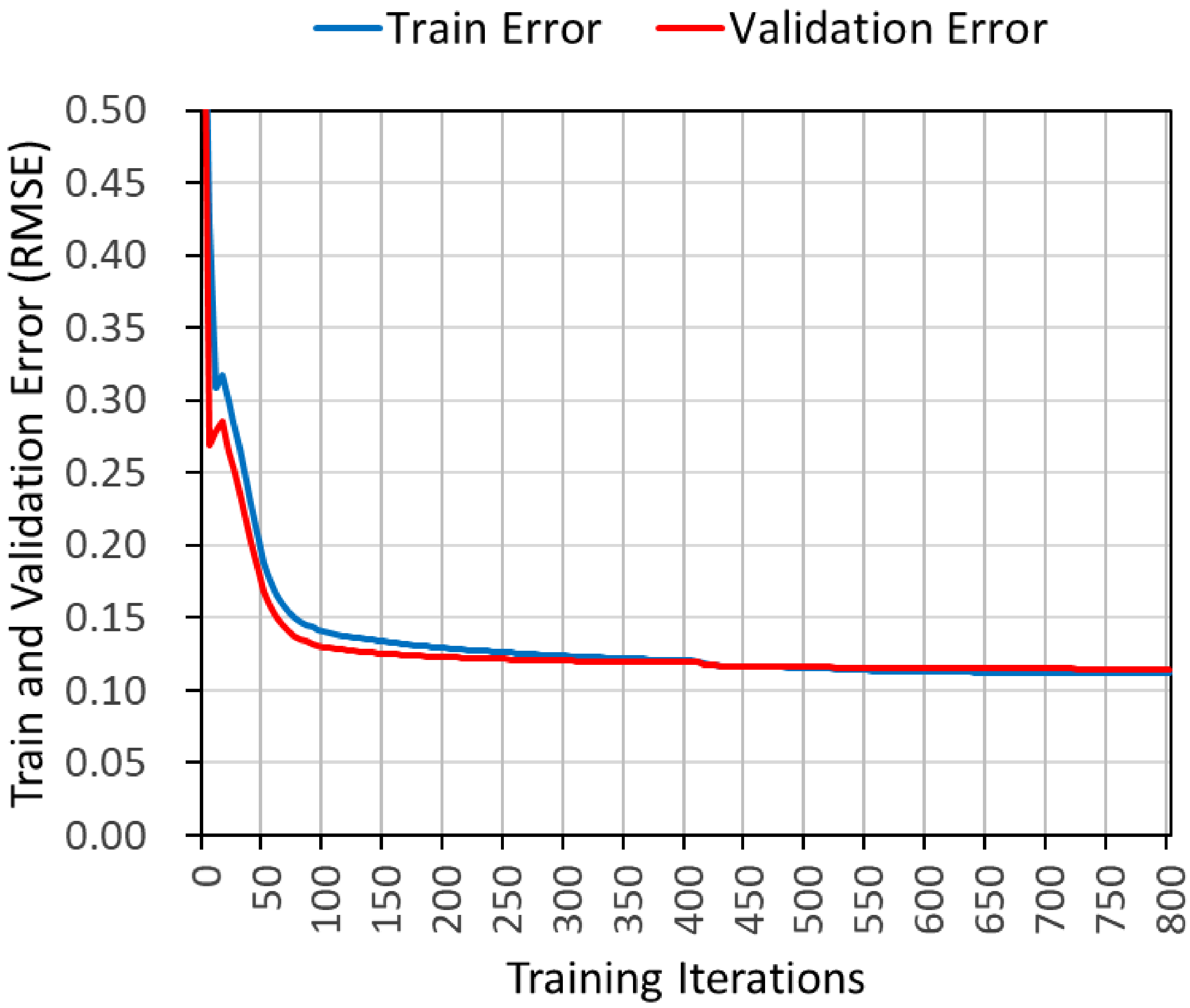
→ Use $F^* = (\boldsymbol{m}^*, \boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{G})$ for inference

Experimental Results



- ✓1,000 macros in 14 nm chip designs and tapeouts
- ✓250 binary meta-parameters
- ✓300,000 LSPD flow results
- ✓150,000 distinct scenarios

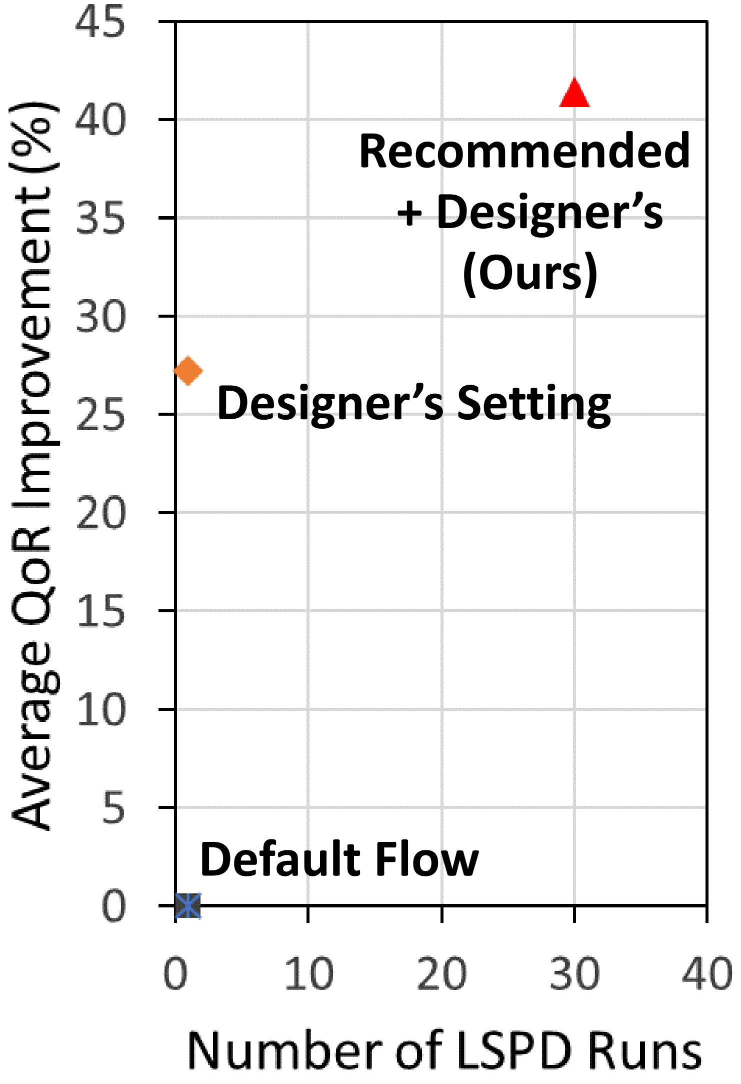
✓**80% train set, 20% validation set**



Experimental Results: Legacy Macros

Macro name	Logic function	Logic gates	Runtime (hours)
FP	Floating-point pipeline	75 K	8.0
ECDT	Execution control & data transfer	45 K	6.2
IDEC	Instruction decode	210 K	21.6
ISC	Instruction sequencing control	77 K	13.1
LSC	L2 cache control & FSM	195 K	12.3

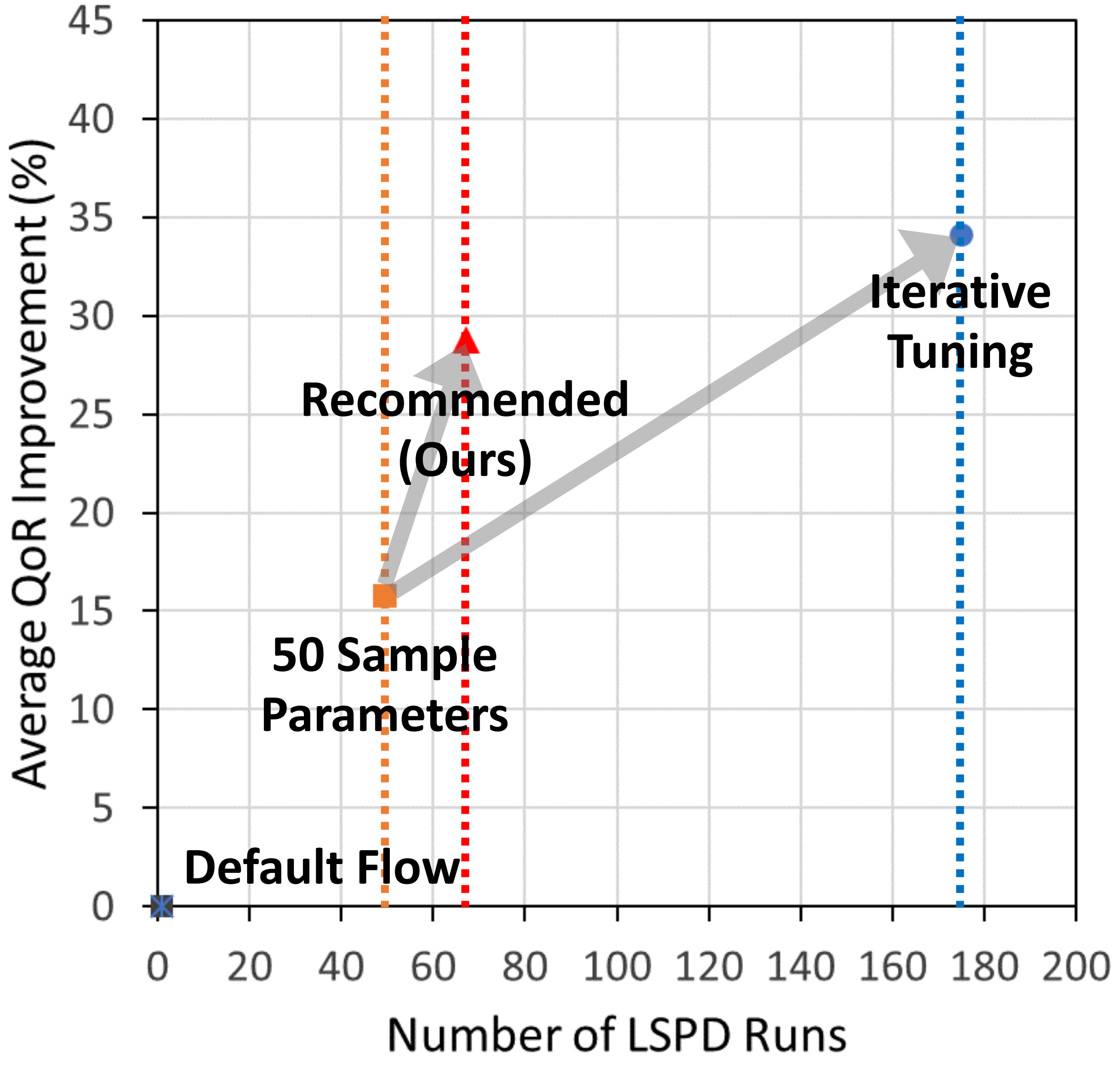
5 Macros from Industrial 14nm Processors



Experimental Results: New Macros

Macro name	Logic function	Logic gates	Runtime (hours)
FP	Floating-point pipeline	75 K	8.0
ECDT	Execution control & data transfer	45 K	6.2
IDEC	Instruction decode	210 K	21.6
ISC	Instruction sequencing control	77 K	13.1
LSC	L2 cache control & FSM	195 K	12.3

5 Macros from Industrial 14nm Processors



Concluding Remarks

- Collaborative recommendation for VLSI design
- Data from LSPD flow runs of industrial high-performance processors
- Reduced computational (LSPD) cost for design-space exploration
- Many unique and unobserved scenarios recommended
- The model learned for 14nm designs used for a 7nm design in progress