IEEE Advancing Technology for Humanity

The 34th IEEE International Symposium on Industrial Electronics
June 20-23, 2025 | Toronto, Canada

IEEE Industrial Electronics Society

Autoregressive DRL for Multi-Robot Scheduling in Semiconductor Cluster Tools

Soo-Hwan Cho, Jean Seong Bjorn Choe, Jong-Kook Kim

Korea University

Paper ID: ISIE25-000262

SEMICONDUCTOR FABRICATION AT A GLANCE

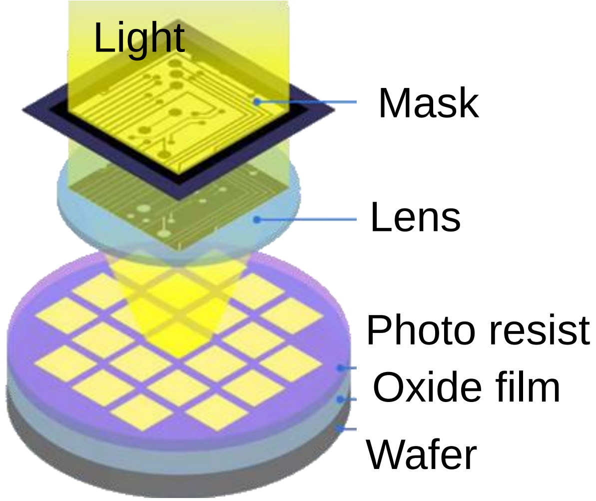
Semiconductor manufacturing involves multi-step processes

– e.g., photolithography, etching, cleaning, deposition

Pattern Drawing Process in Front-end Process

Photolithography → Etch → Clean

Robot



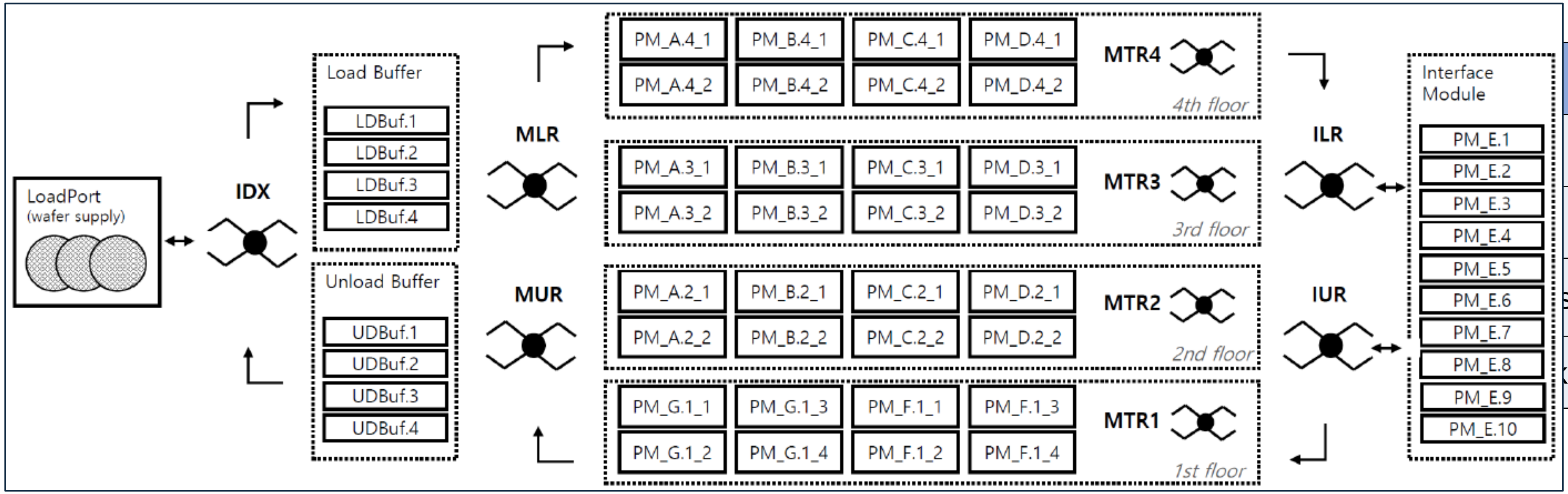
Draw circuits on the wafer

Radial-type Cluster tool

Track-type Cluster Tool

Problem Statement

- Two distinct cluster tool environments with different layouts and robot configurations
- Goal: Learn a scheduling policy that adapts to each environment to maximize throughput



Environment B

	Env B (Track)
	9 (MTRs, ILR/U, MLR/U)
	49
s	4-floor rail system
ks	Coordination across floors

Core Ideas of Our Approach

IEEE ISIE 2025

1. Autoregressive Action Selection

Select robot actions one by one, not jointly.

- Reduces per-step action space
- Enables scaling to many robots

2. Action Masking

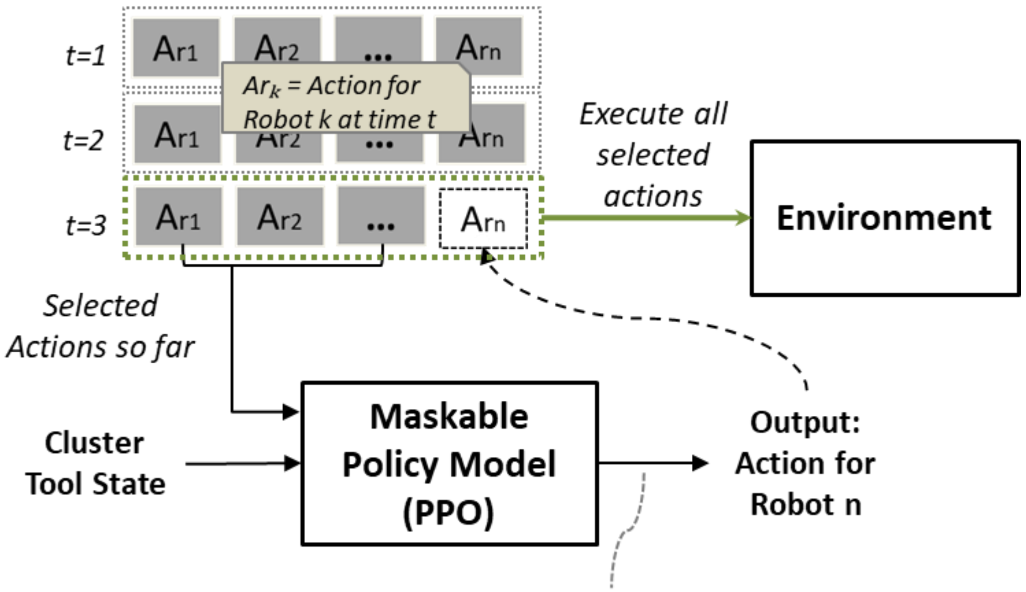
Dynamically mask out invalid actions at each step.

- Improves learning stability
- Helps focus on feasible decisions only

3. Reward Shaping

Combine sparse and dense rewards to guide behavior.

- Penalize idle moves and unproductive actions
- Encourage parallel processing and better throughput



$$a_{t,r_n} \sim \pi(\cdot | f^n(s_t, a_{t,r_1}, \cdots, a_{t,r_{n-1}}))$$

	Joint	AR (Ours)
Decision Complexity	$\mathcal{O}\left(\prod_i \lvert\mathcal{A}_{r_i}\rvert\right)$	$\mathcal{O}\left(\sum_i \lvert\mathcal{A}_{r_i}\rvert\right)$
Env A Action Count	216	35
Env B Action Count	4.4×10^{13}	302

REWARD DESIGN FOR SCHEDULING OPTIMIZATION

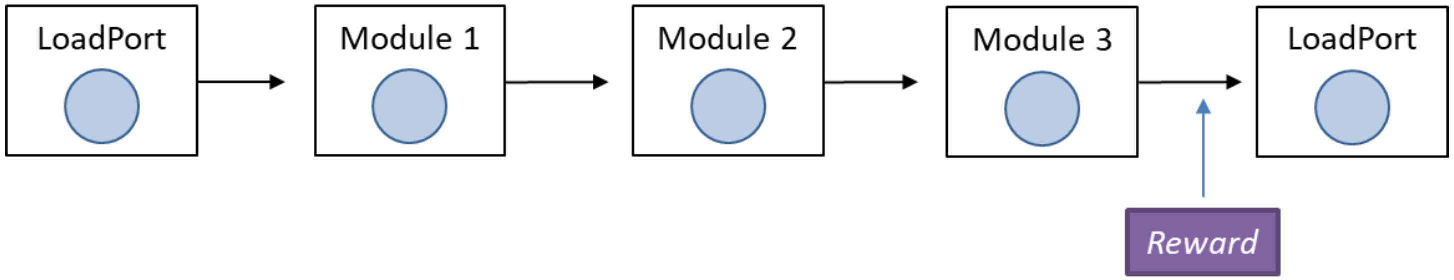
IEEE ISIE 2025

Objective

- Maximize UPEH (Units Per Equipment Hour)
- Provide both sparse but precise and dense but guiding rewards

1. Completion Reward

- $r_t = k_1 \cdot \underline{r_t^{\text{completion}}} + k_2 \cdot r_t^{\text{progress}} - k_3 \cdot r_t^{\text{idle}}$
- +1 when a wafer finishes all processes and returns to the load port
- Sparse but directly aligns with the optimization goal (UPEH)



Reward Design for Scheduling Optimization

2. Wafer Progress Reward

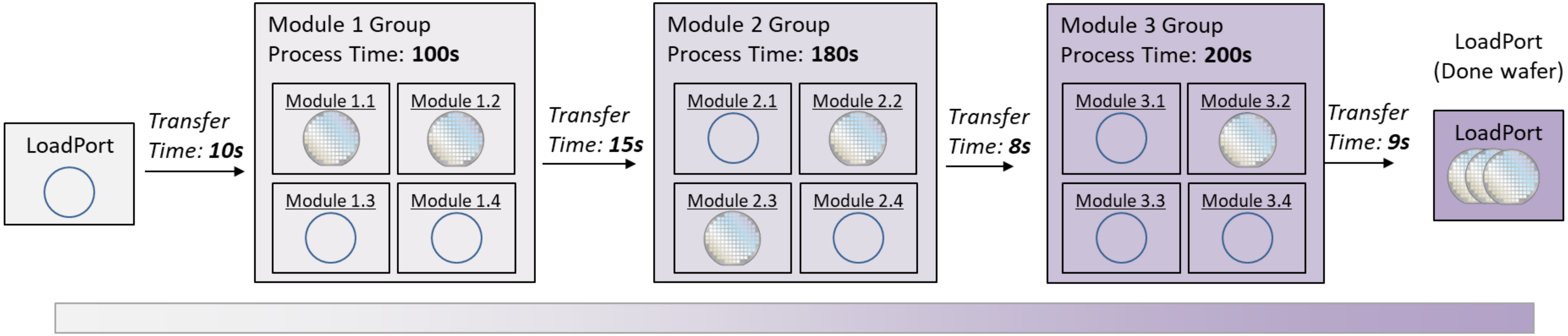
- $r_t = k_1 \cdot r_t^{\text{completion}} + \underline{k_2 \cdot r_t^{\text{progress}}} - k_3 \cdot r_t^{\text{idle}}$
- A dense reward that quantifies how far each wafer has progressed through its processing path.

$$P_t = \sum_{m \in \mathcal{M}} (w_m + \tau_m)$$

$$r_t^{\text{progress}} = P_t - P_{t-1}$$

Wm is the ideal time to reach module **m** without any delays based on physical layout, process time, robot speed, travel distance

Pt is Cumulative wafer progress at time t

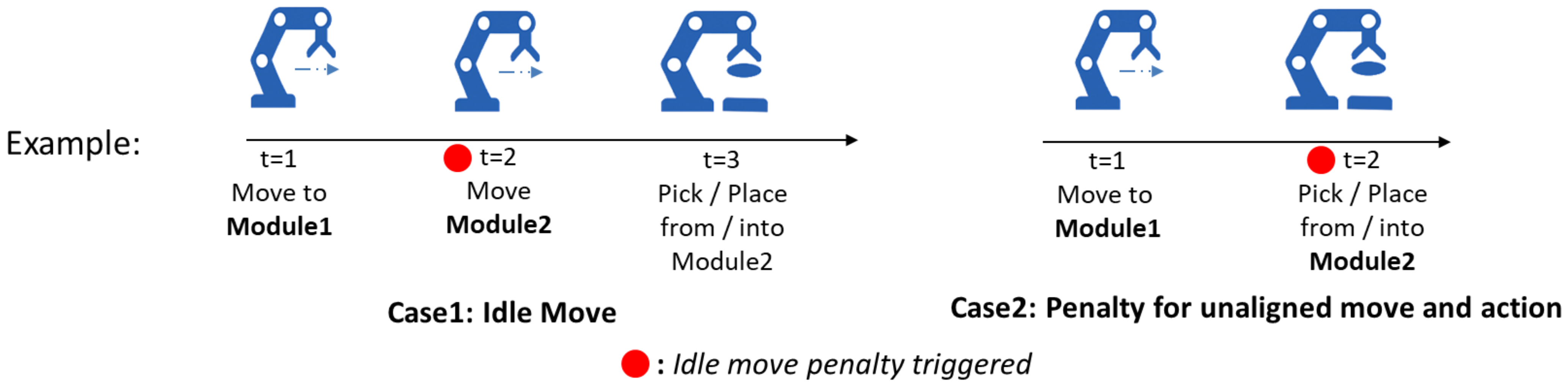


Example of Per-Step Wafer Progress Reward w_m

Reward Design for Scheduling Optimization

3. Idle Move Penalty

- $r_t = k_1 \cdot r_t^{\text{completion}} + k_2 \cdot r_t^{\text{progress}} \underline{- k_3 \cdot r_t^{\text{idle}}}$
- Negative reward for movement actions that do not result in pick/place
  → Discourages unnecessary robot movements



Adaptive Learning Results in Env A and B

Learning Convergence

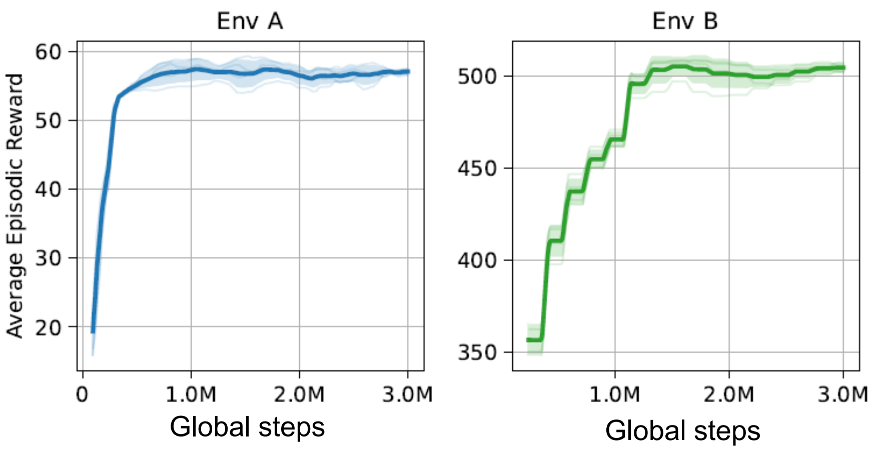
- Stable training curves in Env A & B
- Adaptive policy learned from scratch

Throughput Comparison

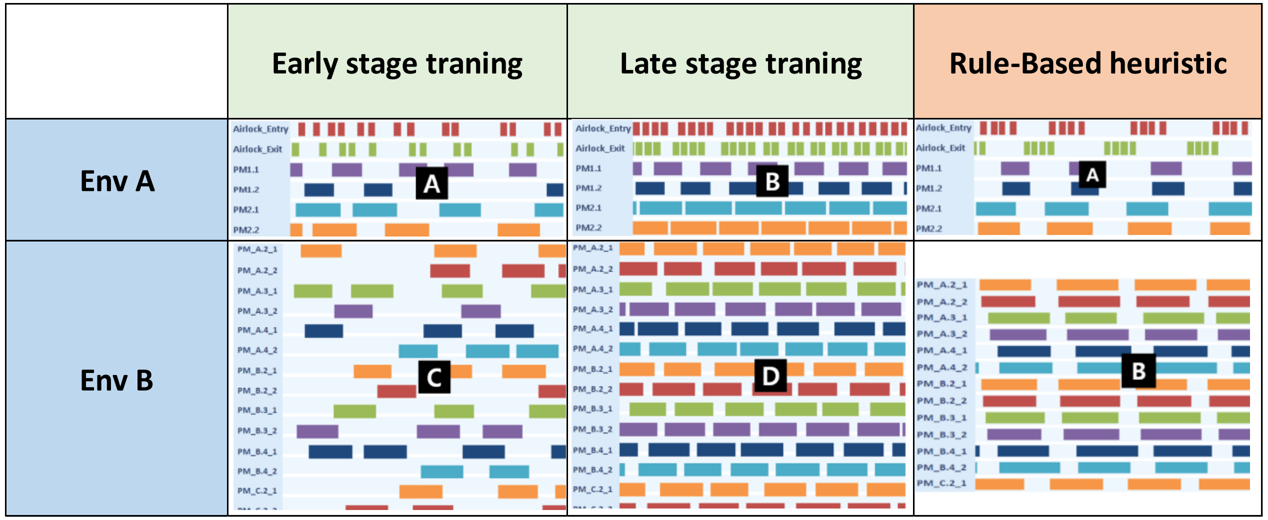
- Higher UPEH than heuristic baseline in both environments

Environment	Ours	Rule-Based
Env A	**96 ± 5**	58
Env B	**304 ± 1**	276

UPEH comparison: Ours vs. rule-based



Training curve showing convergence of average episode reward



Gantt chart visualizations of wafer processing: early vs. late training (DRL) and rule-based

Adaptive Learning Results in Env A and B

Ablation Study – Reward Composition

- Completion reward alone is sparse and insufficient
- Progress reward accelerates learning but lacks final alignment
- Combined rewards yield the best UPEH in both Env A and B

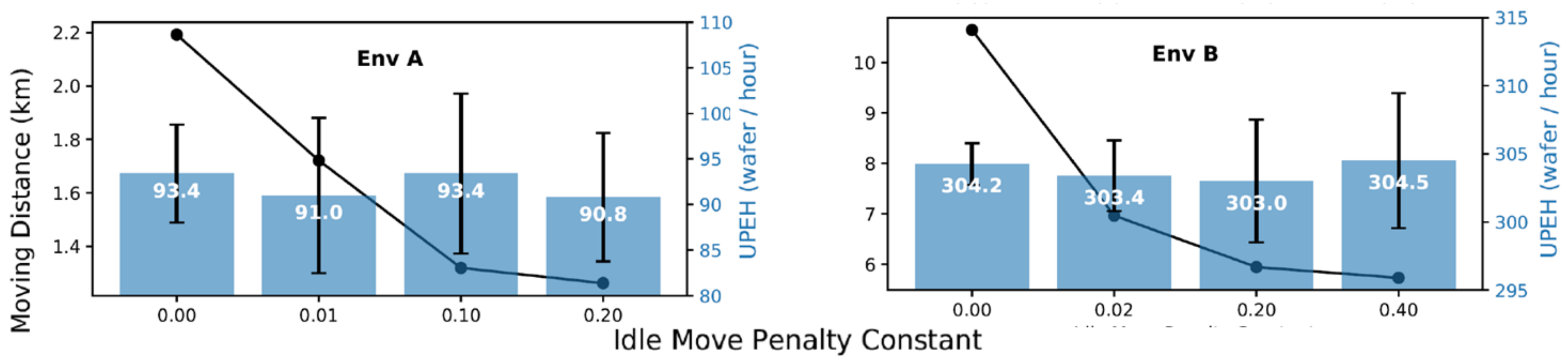
UPEH(Unit per Equipment Hour)

Environment	Completion Only	Progress Only	Both Combined
Env A	81 ± 3	91 ± 6	**96 ± 5**
Env B	52 ± 5	301 ± 2	**304 ± 1**

Combined use of sparse (completion) and dense (progress) rewards leads to optimal throughput

Idle Move Penalty Effect

- Penalizing idle movement reduces total travel distance
- Proper penalty values reduce inefficiency without hurting UPEH



UPEH(Unit per Equipment Hour)

Thank you

🗁 Project code & details:

https://github.com/splendidz/ar_drl_cluster_tool

soohwancho@korea.ac.kr