생성형 AI 기반 강화학습을 활용한 반도체 Fab 스케줄링 최적화

최재영, 안효준, 김중헌
고려대학교

{cielblue0522, hyojun, joongheon}@korea.ac.kr

Generative AI-Based Reinforcement Learning for Semiconductor Fab Scheduling Optimization

Jaeyoung Choe, Hyojun Ahn, Joongheon Kim
Korea Univ.

요 약

현대 반도체 제조 시스템은 재진입 공정과 높은 공정 변동성이 공존하는 복잡한 환경으로, 생산 효율을 극대화하기 위한 정교한 스케줄링이 필수적이다. 그러나 규칙 기반 알고리즘은 급변하는 공정 상황에 취약하며, 기존 강화학습 기법 또한 방대한 연속 행동 공간으로 인해 고부하 Fab 환경에서 최적 정책으로 수렴하는 데 어려움을 겪는다. 이에 본 연구는 생성형 AI 가 유망한 행동 후보를 생성하고 강화학습이 이를 평가·선택함으로써 탐색 공간을 효율적으로 축소하여 기존 기법의 구조적 한계를 극복하고 공정 안정성을 향상시키는 하이브리드 스케줄링 프레임워크를 제안한다.

Ⅰ. 서 론

현대 산업 중 반도체 제조 공정은 재진입 흐름, 높은 공정 변동성, 복잡한 설비 제약이 동시에 작용하여 가장 복잡한 스케줄링 문제 중 하나로 평가된다. 이에 따라 반도체 Fab 스케줄링에 강화학습을 적용하여 동적으로 의사결정을 수행하려는 시도가 최근 활발히 진행되고 있다. 그러나 기존 강화학습 기법은 높은 환경 불확실성과 방대한 연속 행동 공간, 초기 탐색의 비효율성으로 인해 실제 Fab 수준의 복잡성에서 안정적으로 수렴하는 데 한계를 보인다 [1].

본 연구에서는 생성형 AI 가 유망한 행동 후보를 생성하고 강화학습이 이를 평가·선택하는 하이브리드 구조를 도입함으로써, 대규모 연속 행동 공간을 효율적으로 탐색하고 급변하는 공정 상황에도 실시간으로 대응할 수 있는 고도화된 스케줄링 방법론을 제안한다.

Ⅱ. 이론적 배경

본 연구는 다수의 제조 장비와 버퍼로 구성된 반도체 Fab 환경을 고려한다. 각 장비는 재진입 공정 특성에 따라 다양한 종류의 Lot 을 처리하며, 이 과정에서 제품 유형 변경에 따른 셋업 시간과 같은 물리적 제약이 발생한다. 공정 내에는 확률 분포에 따라 Lot 이 지속적으로 유입되며 Agent 는 매 시점 대기 중인 후보군 중 최적의 Lot 을 선택하여 장비에 할당하는 의사결정을 수행한다.

이러한 동적 환경에서의 의사결정 문제는 Markov Decision Process (MDP)으로 모델링 가능하며, 최적 정책 학습을 위한 이론적 토대로 벨만 최적 방정식과 정책 경사 및 어드밴티지 함수 등이 활용된다. [2].

(1) Markov Decision Process

MDP 는 상태와 행동의 상호작용을 기반으로 의사결정을 모델링하는 구조로, 강화학습에서 공정 스케줄링과 같은 동적 최적화 문제를 표현하는 데 활용된다.

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma) \quad (1)$$

여기서 $\mathcal{S}$ 는 상태 공간, $\mathcal{A}$ 는 행동 공간, $\mathcal{P}$ 는 상태 전이 확률 함수, $\mathcal{R}$은 보상 함수, γ는 할인율을 나타낸다.

(2) 벨만 최적 방정식

벨만 최적 방정식은 특정 상태-행동 쌍 (s, a) 에서의 최적 Q 값이, 현재 보상과 이후 상태에서 취할 수 있는 최적 행동의 기대 가치의 합으로 표현됨을 나타낸다.

$$Q^*(s,a) = \mathbb{E}[r + \gamma \max_{a'} Q^*(s', a')] \quad (2)$$

여기서 $Q^*(s,a)$ 는 최적 Q 값, r 은 즉각적인 보상, s' 은 다음 상태, γ는 미래 보상의 할인율이며, $\max_{a'} Q'(s', a')$는 다음 상태에서 가능한 행동 중 가장 높은 가치를 의미한다 [3].

(3) 재파라미터화 기법

재파라미터화 기법은 생성형 AI 의 확률적 샘플링 과정을 고정된 노이즈와 결정적 변환으로 분해하여 미분 가능한 형태로 만든다.

$$a = \mu_\theta(s) + \sigma_\theta(s) \odot \epsilon \quad (3)$$

여기서 a는 행동, s는 상태를 나타내며 θ는 가중치, ϵ은 노이즈, $\mu_\theta(s)$ 는 행동의 평균값, $\sigma_\theta(s)$ 는 행동의 표준편차를 나타낸다.

Table I: 스케줄링 방식에 따른 성능 비교

Method	Throughput (lots /day)	Cycle Time (days)	Tardiness (%)	Utilization (%)	WIP Level (lots)	Setup Time (hrs /day)	Queue Time (days)
FIFO	0.20	53.62	100.00	88.77	184.00	1763.59	53.60
Selector-only RL	2.40	40.72	79.00	86.76	90.20	1416.00	40.70
Proposer-only RL	0.08	56.99	100.00	88.73	198.60	1748.06	56.97
DQN	2.10	38.56	62.88	81.09	92.40	1331.75	38.54
GenAI-RL (Proposed)	**2.60**	**33.45**	**38.13**	**79.07**	**81.80**	**1189.21**	**33.43**

Ⅲ. 시스템 모델 설계

제안하는 GenAI-RL 은 생성형 AI 기반의 Proposer 와 강화학습 기반의 Selector 가 상호작용하는 계층 구조를 기반으로 한다. Proposer 는 주어진 Fab 상태를 입력 받아 현재 상황에 적합한 다수의 행동 후보를 확률적으로 생성한다. Selector 는 생성된 각 행동 후보에 대한 Q 값을 평가하여 최적의 행동을 선택한다. 이러한 구조를 기반으로 공정의 생산 효율성을 높이고 비효율적 요소를 억제하기 위해 정의된 보상 함수는 다음과 같다.

$$r_t = \tanh\left(\frac{1}{2}\left(w_{th}\Delta\mathcal{T}_t - w_{ct}\Delta C_t - w_{setup}\Delta S_t - w_{sc}(t)\Delta\mathcal{L}_t\right)\right) \quad (4)$$

여기서 $\Delta\mathcal{T}_t$ 는 생산이 완료된 Lot 의 수 변화량, ΔC_t 는 평균 사이클 타임의 변화량, ΔS_t 는 step 에서 발생한 Setup time 의 총량, $\Delta\mathcal{L}_t$ 는 폐기(scrap)된 Lot 의 수이다.

Ⅳ. 성능 평가

본 연구에서는 제안하는 강화학습 기반 스케줄링 프레임워크의 성능을 검증하기 위해 현실적인 반도체 Fab 시뮬레이션 환경을 구축하였다. 시뮬레이터는 재진입 흐름, 배치 처리, 제품군 및 작업 순서에 따른 셋업, 그리고 확률적 처리 시간과 Poisson 기반 Lot 도착률 등 웨이퍼 제조 공정의 핵심 특성을 반영한다. Fab 은 총 100 대의 제조 장비로 구성되며, 5 가지 제품군이 각각 약 400~700 단계의 공정을 거치도록 설계하였다. 학습은 총 20,000 스텝의 환경 상호작용을 통해 수행되었으며, 각 에피소드는 60 일 동안의 Fab 운영을 시뮬레이션하였다.

Table I 의 주요 성과 비교 지표는 제안하는 GenAI-RL 방식의 우수성을 보여준다. GenAI-RL 은 기존 베이스라인 알고리즘 대비 유의미한 성능 향상을 달성했다: Selector-only RL 대비 8.3%의 Throughput 증가 (2.60 vs 2.40 lots/day), DQN 대비 13.3%의 Cycle Time 단축 (33.45 vs 38.56 days)과 39.4%의 Tardiness 개선 (38.13% vs 62.88%)을 달성했다.

특히 Utilization 은 79.07%의 수준으로 유지하면서도 모든 비교군 중 가장 낮은 Setup Time (1189.21 hrs/day)을 기록했는데 이는 불필요한 작업을 최소화하는 배칭(Batching) 및 의사결정이 이루어졌음을 시사한다. 또한 WIP Level 을 81.10 lot 으로, Queue Time 을 33.43 days 로 감소시킨 것은 탁월한 흐름 관리 능력을 입증하며, 병목 현상과 공정 지체 위험을 효과적으로 방지함을 보여준다. 이러한 결과는 제안하는 GenAI-RL 방식이 동적인 Fab 환경에 적응하여 다중 목표를 통합적으로 최적화하는 정교한 스케줄링 정책을 효과적으로 달성했음을 증명한다.

Ⅴ. 결론

본 연구에서는 생성형 AI 기반 강화학습을 통해 반도체 Fab 의 복잡한 공정 문제를 해결하고 생산 효율을 극대화하는 스케줄링 최적화 기법을 제안한다. 제안된 기법은 방대한 연속 행동 공간에서의 유효 후보 생성과 정밀한 가치 평가가 가능하기에, 기존 규칙 기반 및 단일 강화학습 방식과 비교했을 때 Setup time 을 획기적으로 줄이면서 Throughput 을 극대화하는 데 성공한다. 이는 고부하 반도체 공정에서 요구되는 유연한 대응력과 공정 안정성을 확보하는 데 있어 유용한 전략임을 시사한다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2025-00561377).

참 고 문 헌

[1] K. Arulkumaran, M. P. Deisenroth, M. Brundage and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26-38, Nov. 2017.

[2] N. C. Luong *et al.*, "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133-3174, May. 2019

[3] P. Dai, W. Yu, H. Wang and S. Baldi, "Distributed Actor-Critic Algorithms for Multiagent Reinforcement Learning Over Directed Graphs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 7210-7221, Oct. 2023.