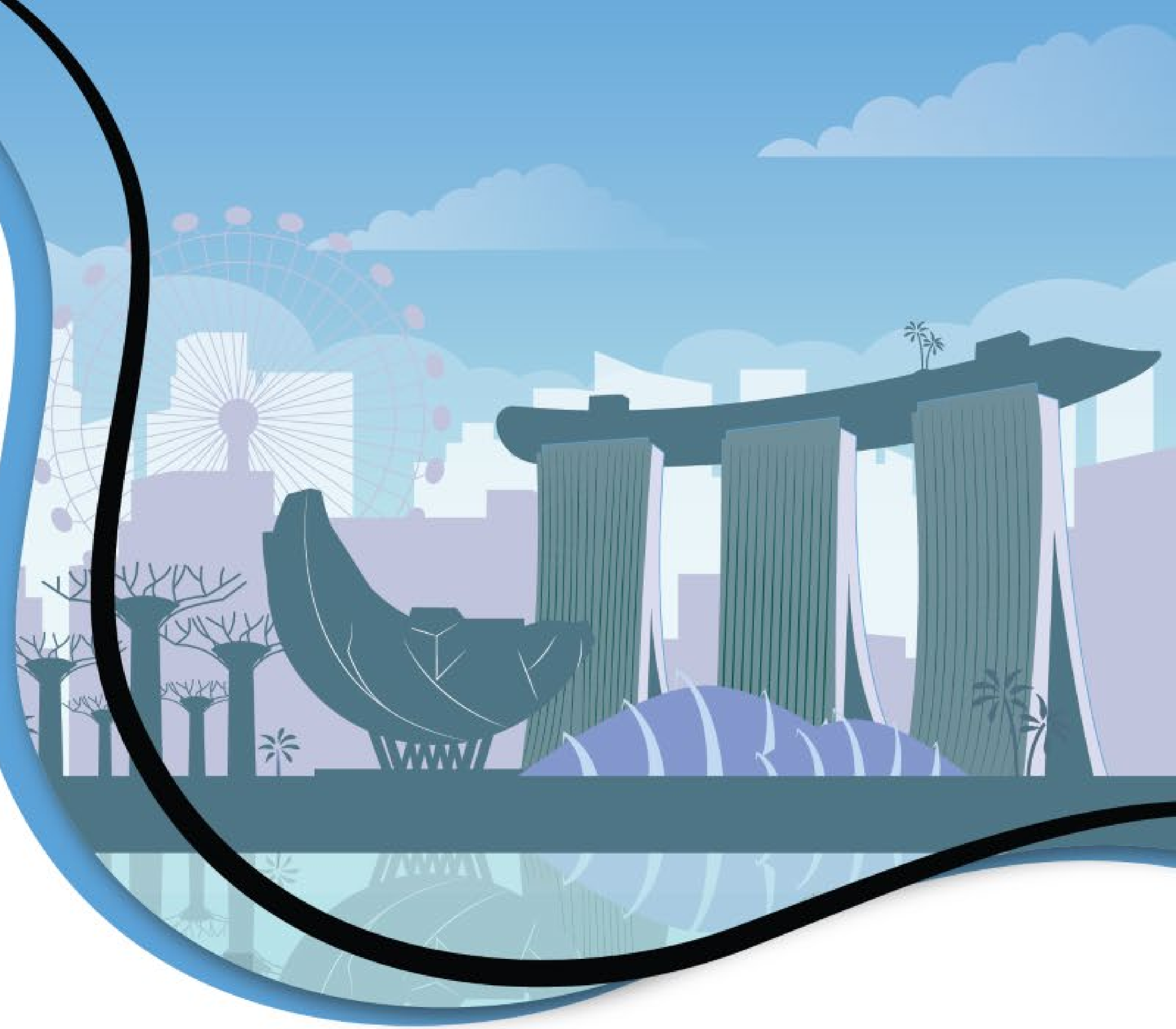
IEEE IEEM VIRTUAL 2021



13 - 16 December 2021

2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)

www.ieem.org

Organizers: IEEE

IEEE Singapore Section
IEEE TEMS Singapore Chapter
IEEE TEMS Hong Kong Chapter

IEEE IEEM VIRTUAL 2021
2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)
13 - 16 December 2021 www.ieem.org

Control of Shared Production Buffers: A Reinforcement Learning Approach

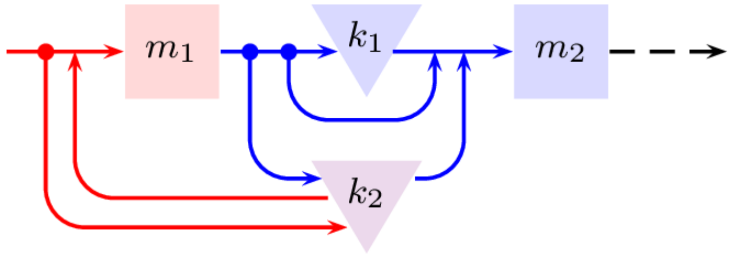
Nora Krippendorff, Christoph Schwindt
Clausthal University of Technology

Organizers: IEEE
IEEE Singapore Section
IEEE TEMS Singapore Chapter
IEEE TEMS Hong Kong Chapter

Problem statement: flow line and shared buffer

- Asynchronous flow line for mixed-model production
- Machines $m = m_1, \ldots, m_M \in \mathcal{M}$ decoupled by buffer slots
- Machine blocked if no slot available for processed item
- Assume one shared central buffer with slots $k \in \mathcal{K}$
- Transfer actions $a \in \mathcal{A}$ among machines and slots
  - supply released part to machine m_1 or slot k
  - pass part from machine m to machine m' or slot k
  - retrieve part from slot k for machine m or slot k'
  - transfer finished part from machine m_M to stock
- Actions a incur transfer costs $c(a)$, finished parts yield contribution margins $v(a)$, payments $r(a) = v(a) - c(a)$

- Model complex configurations using prohibitive costs $c(a)$



$$\mathbf{C} = \begin{array}{c} \\ k_1 \\ k_2 \\ m_1 \\ m_2 \end{array} \begin{array}{c} \begin{array}{cccc} k_1 & k_2 & m_1 & m_2 \end{array} \\ \begin{pmatrix} - & \infty & \infty & 0 \\ \infty & - & 0 & 0 \\ 0 & 0 & - & 0 \\ - & - & - & - \end{pmatrix} \end{array}$$

Problem statement: stochastic process and objective

- Parts released to line form renewal process with rate λ
- Processing times on machines m independent random variables S_m with $\mathbb{E}(S_m) = \mu_m^{-1}$
- Transportation times supposed to be negligible
- System status i encoded as tuple $i = ((i_k)_{k\in\mathcal{K}}, (i_m)_{m\in\mathcal{M}})$ with $i_k \in \{0\} \cup \mathcal{M}$ and $i_m \in \{0, 1, 2\}$
- Status i left upon release or completion event $e \in \mathcal{E} = \{e_0, e_1, e_2, \dots, e_M\}$
- Deterministic event-driven buffer control policy π selects $a \in \mathcal{A}$ at occurrence of $e \in \mathcal{E}$
- Stochastic evolution of system $(t^n, i^n, e^n, a^n)_{n\in\mathbb{N}_0}$ represents random sample path induced by policy π with $i^{n+1} = \sigma(i^n, e^n, a^n)$ and $a^n = \pi(t^0, i^0, e^0, a^0, \dots, t^n, i^n, e^n)$
- Buffer control problem: determine policy π maximizing $npv(\boldsymbol{a}) = \mathbb{E}\left(\sum_{n=0}^{\infty} r(a^n) \cdot e^{-\alpha t^{n+1}}\right)$ with discounting rate $\alpha > 0$

Continuous-time Markov decision problem

- Assume exponentially distributed interrelease and processing times of parts
- Proposition: Buffer control problem can be interpreted as infinite-horizon finite-space discounted continuous-time Markov decision problem $(\mathcal{S}, \mathcal{A}, q, r, P^0, \alpha)$ with
  - state space $\mathcal{S}$ with states $s = (i, e)$, event $e \in \mathcal{E}(i)$ terminating stay in system status i
  - action space $\mathcal{A}$ of transfer actions a
  - transition rates $q_{ss'}^a$ from states s to states s' when a is selected upon entering s
    - $q_{e_0} = \lambda$, $q_{e_m} = \mu_m$, $q_s = q_{i,e} = \sum_{f \in \mathcal{E}(i)} q_f$, $q_{ss'}^a = q_s \cdot \frac{q_{e'}}{q_{s'}}$ with $s' = (i', e')$, $i' = \sigma(i, e, a)$
  - rewards $r(s, a, s') = r(a)$ received upon entering s' when a has been taken in s
  - initial probability distribution P^0 over $\mathcal{S}$ with $P^0(s) = 1$ if $s = (0, e_0)$ and $P^0(s) = 0$, else
  - continuous-time discounting rate α
- Theorem (Puterman 1994): $(\mathcal{S}, \mathcal{A}, q, r, P^0, \alpha)$ admits an optimal stationary policy $\pi: \mathcal{S} \to \mathcal{A}$

Discrete-time Markov decision problem

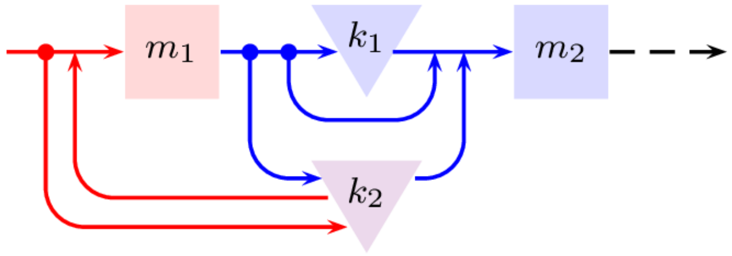
- Theorem (McMahon 2008): Continuous-time Markov decision problem $(\mathcal{S}, \mathcal{A}, q, r, P^0, \alpha)$ can be transformed into discrete-time Markov decision problem $(\mathcal{S}, \mathcal{A}, p, \bar{r}, P^0, \bar{\gamma})$
  - Scale transition rates $q_{ss'}^a$ by factor $1/\nu$ with $\nu = \lambda + \sum_{m \in \mathcal{M}} \mu_m$ → rates $\bar{q}_s = \frac{q_s}{\nu} \leq 1$
  - Set one-step transition probabilities $p_{ss'}^a = \bar{q}_{ss'}^a$ from s to $s' = (i', e')$ with $i' = \sigma(s, a)$
  - Put self-loop probability $p_{ss}^a = 1 - \sum_{s' \neq s} p_{ss'}^a = 1 - \bar{q}_s \geq 0$
  - Scale rewards $r(a)$ by factor $\frac{q_s + \alpha}{\nu + \alpha}$ → rewards $\bar{r}(s, a)$
  - Replace discount factor $\gamma = e^{-\alpha}$ by discount factor $\bar{\gamma} = \frac{\nu}{\nu + \alpha}$
- Discrete-time Markov decision problem amenable to traditional reinforcement learning

Q-learning for optimal buffer control

- Simulation-based learning method for problems of type $(\mathcal{S}, \mathcal{A}, p, \bar{r}, P^0, \bar{\gamma})$
- Learn state-action value function $Q(s,a) = \bar{r}(s,a) + \bar{\gamma} \sum_{s' \in \mathcal{S}} p^a_{ss'} \cdot \max_{a' \in \mathcal{A}(s')} Q(s', a')$
- Q-learning iteratively approaches Q-values by estimators $\hat{Q}(s,a)$
  1. initialize: put $\hat{Q}(s,a) \coloneqq 0$ for $s \in \mathcal{S}, a \in \mathcal{A}(s)$ and $s \coloneqq s^0$
  2. select action $a \in \mathcal{A}(s)$ based on probabilities $\mathbb{P}(a \mid s)$ depending on values $\hat{Q}(s,a)$
  3. randomly draw next state s' with one-step probabilities $p^a_{ss'}$
  4. update $\hat{Q}(s,a) \coloneqq (1-\eta) \cdot \hat{Q}(s,a) + \eta \cdot \left(\bar{r}(s,a) + \bar{\gamma} \cdot \max_{a' \in \mathcal{A}(s')} \hat{Q}(s', a')\right)$ with dynamic learning rates $\eta = \eta(s,a)$, put $s \coloneqq s'$, and return to step 2
- Determine policy π: for $s \in \mathcal{S}$ choose $\pi(s) \in \arg\max_{a \in \mathcal{A}(s)} \hat{Q}(s,a)$, put $npv \coloneqq \hat{Q}(s^0, \pi(s^0))$

Experimental validation: problem instance

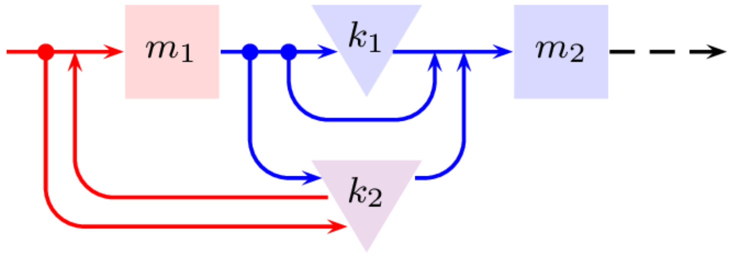
- Toy instance with two machines and two buffer slots
- Action space $\mathcal{A}$: 6 atomic, 5 compound transfer actions a
- State space $\mathcal{S}$: 25 states s, four of which require decision between two alternative actions
  - Upon completion of a part on m_1, store the part in k_1 or in k_2 if m_2 is occupied and both slots are available?
  - Upon completion of a part on m_2, retrieve a part from k_1 or from k_2 if both slots are occupied with parts for m_2?
- Optimal policy π^*: store parts for m_2 in dedicated slot k_1 whenever possible and retrieve parts for m_2 preferably from shared slot k_2

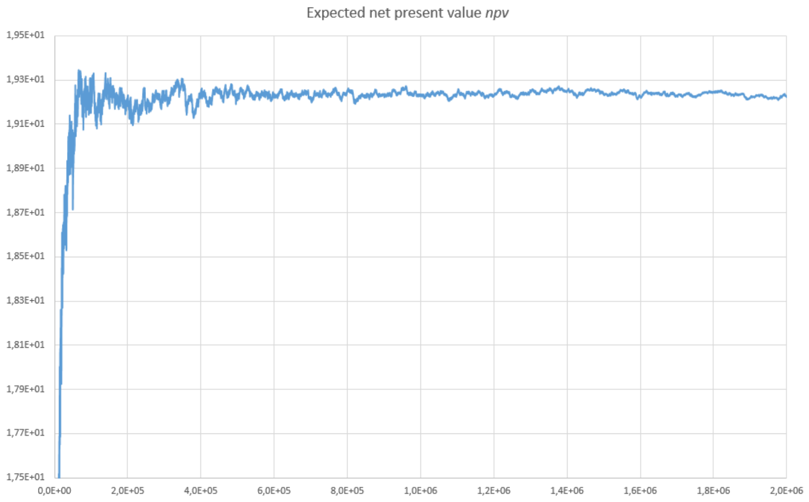


- Machine m_1 can access k_2, parts for machine m_2 can be stored in k_1, k_2
- Transfer costs $c(a) = 0$
- Unit contribution margin $v = 1$
- Arrival and processing rates $\lambda = \mu_1 = \mu_2 = 1$
- Discount rate $\alpha = 0.\overline{03}$

Experimental validation: results

- $\mathcal{A}$ and $\mathcal{S}$ rather small, toy instance challenging though
- No transfer cost: Q-learning only guided by the weak effects of storage and retrieval decisions on production rate, so high accuracy of estimators $\hat{Q}(s, a)$ needed
- In 20 replications of experiment, optimal policy found within 2,000,000 iterations
- Speed of convergence largely depends on parameters controlling learning rate η and probabilities $\mathbb{P}(a \mid s)$
- Exact values $Q(s, a)$ obtained by solving Bellman's optimality equation with linear programming
- Mean relative error of estimators $\hat{Q}(s, a)$: 0.135 %





Conclusions

- Buffer control for stochastic flow lines as continuous-time Markov decision problem
- Transformation to discrete-time Markov decision problem
- Q-learning reliably provides optimal stationary buffer control policy for small instance
- Research avenues:
  - More advanced methods: deep reinforcement learning representing function $Q(s, a)$ as neural network, approximate dynamic programming, e. g., linear programming with approximate value functions
  - More general models: inventory holding cost via permanence rewards, more general production systems including convergent flows, semi-Markov production processes, or time durations following phase-type distributions

Thank you for your attention

TU Clausthal

Nora Krippendorff
Christoph Schwindt
Operations Management Group
Clausthal University of Technology

operations@tu-clausthal.de
www.wiwi.tu-clausthal.de/produktion