CVaR-Constrained Heterogeneous-Agent Proximal Policy Optimization for Robust Multi-Echelon Inventory Control under Stochastic Demand and Lead Times

Youssef Ben Amor[1], Ilhem Slama[1], Zied Jemai[1,2], Evren Sahin[1]
[1]CentraleSupélec, Laboratoire Genie Industriel, Université Paris-Saclay, 3 rue Joliot-Curie, Gif-sur-Yvette, Ile-de-France, France
[2]LR-OASIS, National Engineering School of Tunis, University of Tunis El Manar, Tunis, Tunisia
Emails: {youssef.benamor; ilhem.slama; zied.jemai; evren.sahin}@centralesupelec.fr

***Abstract*—In this paper, we propose a CVaR-constrained Centralized-Training-Decentralized-Execution (CTDE) method: C-HAPPO, which is a constrained variant of Heterogeneous-Agent Proximal Policy Optimization (HAPPO). We study a three-echelon serial supply chain with Poisson demand, and stochastic lead times. We find a sharp median–tail trade-off: a CTDE baseline (HAPPO) achieves excellent typical performance (low median cost) by reducing the median cost of Robust BaseStock, yet exhibits catastrophic tail outcomes inflating mean cost and risk metrics. C-HAPPO mitigates tail failures while remaining competitive on cost and service, reducing the mean cost and the variance cost of Robust BaseStock by 4% and by 60%, respectively, and providing a practical path to risk-aware CTDE control under demand and lead-time uncertainties.**

***Index Terms*—multi-echelon inventory, supply chain, MARL, CTDE, HAPPO, CVaR, constrained RL, stochastic lead time, stochastic demand, tail risk**

I. Introduction

Multi-echelon inventory control features delayed feedback, partial observability due to uncertainty and amplification (bullwhip) [1]–[4]. With stochastic lead times and stochastic demand, the effective uncertainty window expands and credit assignment becomes harder [3]. Deep reinforcement learning (DRL) can learn non-linear policies directly from simulation [5], [6], and has shown promise in multi-echelon inventory settings [7], [8]. In decentralized supply chains, multi-agent RL (MARL) and CTDE stabilize learning while preserving decentralized execution [9], [10].

Despite these advances, many studies emphasize expected performance, whereas operations may be governed by tail events: rare backlog cascades and extreme bullwhip realizations. A key practical difficulty is that CTDE policies can look excellent on typical trajectories (e.g., median episode cost) yet remain brittle in the tail under lead-time uncertainty, where unlucky delay realizations align with demand surges, which highlights the value of risk-aware approaches [2]. We characterize and address the median–tail trade-off induced by CTDE policies under stochastic lead times and stochastic demand. We show that a coordinated multi-agent policy can achieve excellent typical performance while remaining fragile to rare but catastrophic delay realizations, and we propose a constrained-learning mechanism that directly targets these tail failures without sacrificing coordination benefits. CVaR has already been used to enhance risk management in many supply chain fields such as food industry as well as pharmaceutical supply chains. For these reasons, we use CVaR to control tail risks in our model. In short, we provide a multi-echelon MARL to frame demand and lead-time uncertainties as a distributional risk-control problem rather than a pure performance-optimization task. Using CTDE multi-echelon inventory under stochastic demand and stochastic lead times in our benchmark, we obserce 4% and 60% reductions in the mean cost and variance cost, respectively, compared to Robust BaseStock.

The main contributions of this paper are: (i) We expose the median–tail trade-off of CTDE inventory control under stochastic lead times. (ii) We introduce C-HAPPO, a CVaR-constrained CTDE method using a primal–dual update that surpasses a robust base-stock reference. (iii) We provide controlled experiments on a 3-echelon serial environment (Poisson demand, stochastic lead time), analyzing mean/median/variance and tail diagnostics.

A literature review that led us to such a research gap is presented in the next section. Section 3 details the problem setting and its assumptions. Furthermore, we introduce the methods conducted in section 4, and then the experimental protocol in section 5. The results and comparison between the RL variants and the baseline are specified in the penultimate section. Finally, we conclude our study and suggest perspectives for tackling the problem in further work.

II. Related Work

Multi-echelon inventory control provides structural benchmarks and motivates order-up-to (base-stock) policies for coupled echelons [11], [12]. In practice, one of the most difficult challenges in optimizing multi-echelon inventory comes from the interaction of uncertainty sources. Stochastic demand and

stochastic lead times can together amplify variability [4], [13]. [7] provides a benchmark study dealing with these uncertainties in several inventory settings (including multi-echelon variants) that shows that Deep Reinforcement Learning (DRL) can approach strong heuristic policies, but that reported gains are strongly conditioned on the modeled assumptions (e.g., lost sales vs. backorders, sourcing options, topology) and on the baseline chosen for comparison . A central reference for our setting is [1], which studies multi-echelon decentralized inventory management using heterogeneous-agent proximal policy optimization (HAPPO) under demand uncertainty. In the serial system considered there, the dynamics follows periodic review with backlog and holding costs, and each echelon observes local inventory and backlog together with a fixed-length vector of in-transit orders, under a fixed lead time assumption [1]. The study compares PPO with several HAPPO variants that differ by training-time information (limited neighbor information, full observation during training, and a fully centralized control-tower variant), showing that increased training-time information improves both cost and bullwhip mitigation. It also documents faster convergence as more information is shared during training (e.g., for $M = 3$ echelons: about 260 episodes for centralized HAPPO, 299 for full-observation HAPPO, and 311 for limited-observation HAPPO). The cost formulation in [1] includes fixed holding and backlog costs without taking into account an ordering cost, while other studies add an order-smoothing term to discourage excessive order variability [14]. The smoothing term is a principled way to internalize the dynamic and coordination-related costs of unstable replenishment decisions. It can improve ordering stability and, in some settings, inventory and service performance as well [15], [16]. Other works have also only stated the demand uncertainty, such as [17] where multi-echelon network experiments trained with PPO report cost reductions relative to heuristic baselines across different topologies, with gains -16.4% on a linear chain, for example. Several works emphasize that conclusions can change once the uncertainty regime includes lead time randomness in addition to demand uncertainty. [18] experiments larger distribution systems (factory–DCs–retailers), where actor–critic methods (A2C, PPO, DDPG, TD3) are compared to static inventory rules (e.g., (s, Q)-type baselines) under stochastic demand with additional lead-time uncertainty, and dynamic DRL policies improve average cost across tested configurations while robustness differs across algorithms. [3] shows a benchmark-style Multi Agent Reinforcement Learning (MARL) analysis which further makes this point under explicit demand and lead-time uncertainty by comparing multi-agent PPO variants between independent learners and centralized training decentralized execution (CTDE) learners and showing that HAPPO can remain competitive. In particular, under increasing stochastic lead time uncertainty, the distributed model-based baseline degrades sharply (e.g., mean episode reward falls by 75% from 416.8 to 102 as the lead time randomness parameter increases), whereas MARL methods exhibit a smaller relative degradation (roughly 42– 46% across the tested PPO variants), highlighting the sensitivity of baselines that assume deterministic lead times [3]. It is worth noting that these learned policies may strategically protect downstream customer-facing stages and shift part of the inventory and shortage upstream under asymmetric downstream shortage penalties [19].

Beyond expected-cost optimization, risk-sensitive RL targets low-probability, high-severity scenarios. [2] is a key reference where distributional RL for multi-echelon inventory demonstrates that optimizing a CVaR objective can improve tail protection relative to expectation-only training, while also improving sample efficiency compared to PPO and outperforming shrinking-horizon mixed-integer formulations. Quantitatively, the same work reports an expected performance improvement of about 11% over PPO for comparable training budgets, and shows that CVaR-optimized learning yields a better tail metric (e.g., CVaR improvement reported as 7.2% in their validation) [2]. More generally, CVaR objectives and constrained/safe RL provide principled tools to control tail outcomes through risk measures and primal–dual or trust-region mechanisms [20], [21]. More broadly, recent MARL work in supply chains highlights the need for coordinated yet stable decision-making under uncertainty and delays, reinforcing the importance of risk-aware control formulations [2], [8].

To conclude, existing studies cover CTDE coordination, demand uncertainty, or lead-time uncertainty separately, but rarely combine them with explicit tail-risk control. In parallel, CVaR-based learning has rarely been instantiated as a lightweight constraint inside a calibrated CTDE multi-echelon benchmark that aligns with the serial setting of [1] while adding explicit lead-time randomness. This work addresses this gap by studying CVaR-constrained CTDE inventory control under stochastic demand and stochastic lead times. Then, we adopt a three-echelon serial system closely aligned with [1], but with both stochastic demand and stochastic lead times explicitly modeled [3], and we integrate a CVaR-based constraint [2] into a CTDE baseline to mitigate catastrophic backlog cascades while preserving coordination benefits to practically close this gap. We explain our problem setting in the next section.

III. Problem Setting

In this section, we introduce our environment and then explain our constraints and assumptions.

A. Supply Chain Environment

We consider a periodic-review serial supply chain with $M = 3$ decision-making echelons between a downstream customer and an upstream supplier, as illustrated in Fig. 1. Echelons are indexed from downstream to upstream: demand D_t is realized at echelon 1, and each echelon $i \in \{1, \dots, M\}$ places a replenishment order to echelon $i + 1$. Following the standard Beer-Game-type abstraction, material flows are delayed and backorders are allowed, so unmet demand accumulates as backlog and is served when inventory becomes

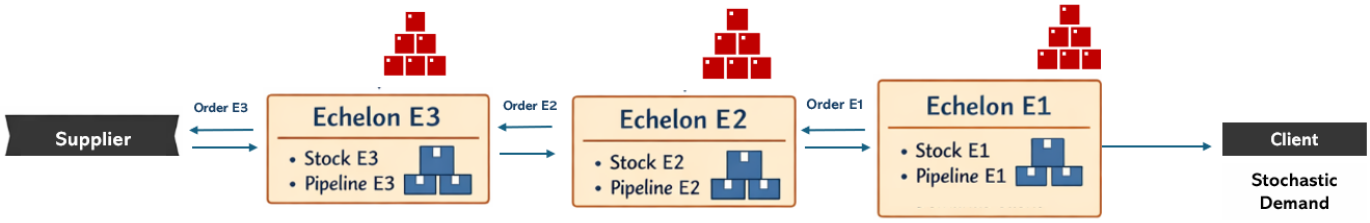

Fig. 1: Overview of a Three-echelon serial supply chain

TABLE I: Main notation used in the model

Symbol	Description
D_t	Customer demand observed at period t.
O_t^i	Replenishment order by echelon i at period t.
L_t^i	Realized lead time associated with an order of echelon i.
A_t^i	Quantity arriving at echelon i at period t.
$\tilde{D}_t^i$	Effective demand by echelon i at t, with previous backlog.
S_t^i	Shipment made by echelon i at period t.
I_t^i	On-hand inventory level of echelon i at the end of period t.
B_t^i	Backlog level of echelon i at the end of period t.
h_i	Unit holding cost coefficient at echelon i.
b_i	Unit backlog penalty coefficient at echelon i.
$k_{\Delta,i}$	Order-smoothing cost coefficient at echelon i.

available. Table I summarizes the main notation used to describe the serial multi-echelon inventory model. We assume an infinite-capacity upstream supplier. To preserve comparability across methods and uncertainty settings, we adopt the same observation structure as in [1]: at each period t, echelon i observes its local on-hand inventory I_t^i and backlog B_t^i, together with a fixed-length vector summarizing orders currently in transit that are expected to arrive within the next periods. This constant-dimensional representation captures the essential delayed-receipt information required for decision-making under replenishment lags, while avoiding variable-length histories that complicate learning and evaluation.

In contrast to [1], which assumes a fixed lead time, we model lead times as stochastic in order to reflect delivery-time variability and to study its interaction with demand variability in a controlled setting. Operationally, each placed order is associated with a random delivery delay, while the in-transit vector provides a finite-memory summary of forthcoming receipts, keeping the observation interface unchanged across realizations. Each echelon selects an order quantity $O_t^i \in \{0, \ldots, O_{\max}\}$, which represents practical ordering limits. Finally, downstream echelons have a single upstream source so their effective replenishment is constrained by upstream availability.

B. Flow conservation, backlog propagation, and costs

To account for backlogged demand, we define the effective demand $\widetilde{D}_t^i$ faced by echelon i at period t as the new requirement plus the previous backlog :

$$\widetilde{D}_t^1 = D_t + B_{t-1}^1, \tag{1}$$

$$\widetilde{D}_t^i = O_t^{i-1} + B_{t-1}^i, \qquad i > 1. \tag{2}$$

Shipments and state variables then follow flow conservation:

$$S_t^i = \min\big(I_{t-1}^i + A_t^i,\ \widetilde{D}_t^i\big), \tag{3}$$

$$I_t^i = I_{t-1}^i + A_t^i - S_t^i, \tag{4}$$

$$B_t^i = \widetilde{D}_t^i - S_t^i. \tag{5}$$

The total per-period cost includes, for each echelon $i \in \{1, 2, 3\}$, an inventory holding term $h_i I_t^i$ and a backlog penalty $b_i B_t^i$, where I_t^i denotes on-hand inventory, B_t^i the backlog, and h_i and b_i their respective unit cost coefficients. We also include an order-smoothing term per echelon i : $\mathrm{k}_{\Delta,i}\,|O_{i,t} - O_{i,t-1}|, where \mathrm{O}_t^i$ is the replenishment order placed at period t and $k_{\Delta,i}$ controls the strength of smoothing. As a result, the total cost at period t is

$$C_t = \sum_{i=1}^{3} \Big(h_i I_t^i + b_i B_t^i + k_{\Delta,i} |O_t^i - O_{t-1}^i| \Big). \tag{6}$$

In the next section, we specify the methods that we conduct.

IV. Methods

In this section, we present the three methods conducted in the experiments. We use a BaseStock (robust) that targets a worst-case protection period $L_{\max} = 5$, i.e., it hedges against the longest lead time realizations. This yields a strong and interpretable reference under stochastic lead times, avoiding an artificially weak "mean-LT" comparator.

We also consider HAPPO; this CTDE baseline uses decentralized recurrent actors (GRU) and a centralized critic trained on the global concatenation of observations. Actors execute locally, while the critic improves credit assignment under delayed arrivals and coupled dynamics. We then introduce C-HAPPO, a constrained variant of HAPPO: instead of optimizing only expected performance, it enforces a CVaR tail constraint to limit rare catastrophic episodes [2], [20]. Let the episode total cost be

$$J(\pi) = \sum_{t=1}^{T} C_t. \tag{7}$$

We constrain the tail at level $\beta = 0.90$ using a Robust BaseStock limit

$$\mathrm{CVaR}_\beta\big(J(\pi)\big) \le \mathcal{C}_\beta^{\mathrm{BS}}, \qquad \beta = 0.90, \tag{8}$$

where $\mathcal{C}_\beta^{\mathrm{BS}}$ is estimated from robust BaseStock episodes and kept fixed during training.

We estimate CVaR_β from a rolling buffer of episode costs $\{J_k\}$ (empirical average of the worst $(1-\beta)$ fraction) and optimize the Lagrangian relaxation

$$\min_{\pi} \ \mathbb{E}[J(\pi)] + \lambda \left[\widehat{\mathrm{CVaR}}_\beta\big(J(\pi)\big) - \mathcal{C}_\beta^{\mathrm{BS}}\right]^+, \qquad \lambda \geq 0. \quad (9)$$

The multiplier λ is updated with a smoothed primal–dual rule using an Exponential Moving Average (EMA), so that penalties increase only when the tail constraint is violated. In practice, we keep the HAPPO training structure and apply an episode-level penalty through a reward shift:

$$r_t^i \leftarrow r_t^i - \lambda\, \bar{v}, \quad (10)$$

which discourages trajectories that contribute to the upper tail while preserving the original clipped HAPPO updates. So we conduct HAPPO and C-HAPPO, and compare them to Robust BaseStock. The data-generation process is specified in the next section.

V. Experimental Protocol

First of all, we define on a horizon of $T = 700$ periods, the customer demand in our system as Poisson:

$$D_t \sim \mathrm{Poisson}(\lambda), \quad \lambda = 40. \quad (11)$$

Lead times are stochastic and independent:

$$L_t^i \sim \mathrm{Unif}\{1, 2, 3, 4, 5\}. \quad (12)$$

We set the vector summarizing orders currently in transit $L_{\mathrm{pipe}} = 5$ periods, and we set the ordering limit at $O_{\max} = 80$.

Training is performed for 300 episodes (i.e., $300 \times 700 = 210{,}000$ time steps). We then evaluate each trained policy on $N_{\mathrm{eval}} = 15$ different independent traces. In our implementation, the dual update in C-HAPPO uses an EMA smoothing factor $\rho = 0.90$, a dual stepsize $\eta_\lambda = 10^{-9}$ to stabilize training under noisy tail estimates. The effective dual stepsize depends on the scale of the constraint violation; rescaling the violation (e.g., dividing by BaseStock Cost) makes η_λ directly comparable across works. We use a damping factor $\delta = 0.02$, and clipping $\lambda_{\max} = 50$; the smooth violation uses a softplus parameter $k = 10^{-6}$.

We measure the total cost within three forms: mean cost, median cost, and variance cost as we report the sample variance across episodes.

The On-Time service level captures the average proportion of period-t demand that is fulfilled without delay once the backlog inherited from period $t-1$ has been cleared, while the horizon service level evaluates the proportion of cumulative demand satisfied over the full planning horizon. We define the customer On-Time service $\mathrm{SL}^{\mathrm{OT}}$ and the Horizon service SL^{H} metrics, where N^+ denotes the number of periods with strictly positive demand :

$$N^+ \doteq \sum_{t=1}^{T} \mathbf{1}_{\{D_t^1 > 0\}}, \quad (13)$$

$$\mathrm{SL}^{\mathrm{OT}} \doteq \frac{1}{N^+} \sum_{t\,:\,D_t^1>0} \frac{\min\Big(D_t^1,\, \max\big(S_t^1 - B_{t-1}^1, 0\big)\Big)}{D_t^1}, \quad (14)$$

$$\mathrm{SL}^{\mathrm{H}} \doteq \frac{\sum_{t=1}^{T} S_t^1}{\sum_{t=1}^{T} D_t^1}. \quad (15)$$

All time-series KPIs are computed within each evaluation episode over the T periods, and the reported values are then aggregated across evaluation episodes (e.g., median or mean). In the next section, we expose the results of KPIs for each method such as Service-level metrics, the average stock and the number of stockouts. These metrics help characterize the behavior of the compared methods.

VI. Results

In this section, we provide numerical results and related discussions. We first present the KPIs listed in the previous section, then we specify the tail diagnostic.

A. Key Performance Indicators

Table II summarizes KPIs for each method: the (mean, median, var) cost, the two Service-level metrics, the average stock and the average number of stockouts and Table III shows the results (average stock and stockouts) analyzed echelon by echelon.

Table II shows that the cost can be evaluated differently between mean, median, and var. It can have different results. To compare between the MARL methods (HAPPO and C-HAPPO) and Robust BaseStock, we evaluate Deltas of Method m (HAPPO, C-HAPPO) vs. BaseStock (%) defined by:

$$\Delta^{(\%)}(m) \ = \ 100 \times \frac{K_{\mathrm{BS}} - K_m}{K_{\mathrm{BS}}} \quad (16)$$

HAPPO achieves an excellent median cost (45% reduction vs. BaseStock, which is large), but exhibits extreme tail episodes that inflate mean and especially variance (the variance is dominated by rare outliers). C-HAPPO remains competitive on mean/median cost by reducing mean BaseStock cost by nearly 4% and significantly reducing variance relative to BaseStock by nearly 60%, and remains consistent with suppressing catastrophic runs.

Table II also shows that BaseStock and C-HAPPO have better On-Time service than HAPPO by 5% and 4%, and a similar gap for horizon service. Service-level metrics suggest a robustness trade-off: C-HAPPO may sacrifice some On-Time service relative to BaseStock to avoid tail disasters, while maintaining near-perfect horizon service.

For the average stock, we note that C-HAPPO has the highest average inventory levels and HAPPO the lowest, which is predictable since risk-aware methods tend to maintain a security margin for their stocks, whereas performance-based

TABLE II: Key KPIs per method

Method	Cost(mean)	Cost(median)	Cost(var)	On-time(mean)	Horizon(mean)	Avg stock(sum)	Stockouts(mean)
BaseStock	5.273e+6	5.174e+6	4.652e+11	0.9535	0.9998	69.1913	46.3333
HAPPO	1.155e+8	2.829e+6	1.9e+17	0.9040	0.9599	64.6551	103.8
C-HAPPO	5.065e+6	5.099e+6	1.893e+11	0.9389	1.0	81.8007	68.7333

TABLE III: C-HAPPO vs BaseStock : Inventory levels and stockouts per echelon

Method	Avg Stock E1	Stockouts E1	Avg Stock E2	Stockouts E2	Avg Stock E3	Stockouts E3
BaseStock	70.13	8.64	73.88	22.42	63.99	17.27
C-HAPPO	62.09	6.32	78.12	14.10	110.20	22.45

methods, such as HAPPO, tend to reduce their stock levels to reduce holding costs. Regarding stockouts, we observe that HAPPO experiences more than twice as many BaseStock stockouts due to catastrophic scenarios. C-HAPPO keeps an intermediate level between BaseStock and HAPPO, which confirms its combination of its performance-oriented nature being a HAPPO variant and moderate risk awareness by using moderate CVaR parameters. We observe that Average stock and stockouts for C-HAPPO are higher than BaseStock, which may initially appear to be a contradiction. But, examining the echelon-level results as shown in Table III, we notice that C-HAPPO reduces both inventory and stockouts at the customer-facing echelon 1, where holding and backlog penalties are the highest. Echelon 2 appears to play an intermediate buffering role, while echelon 3 absorbs the largest part of the extra inventory burden, where the associated cost impact is lower. Therefore, C-HAPPO should not be interpreted as uniformly improving all physical indicators at every stage; instead, it protects the downstream stage that matters most for service and penalty exposure, which makes it a shifting of inventory to a less cost-sensitive echelon. This interpretation is consistent with the idea that the best-performing HAPPO policies arise when each agent optimizes a combination of local and system-wide costs rather than a purely global objective, which induces different echelon behaviors, and with cooperative MARL that learn downstream prioritization policies under inventory scarcity.

B. Tail diagnostics

We illustrate the tail diagnostics using the distributions of episode total cost with median and $\text{CVaR}_{0.90}$ markers for each method used. Figures 2–4 provide complementary insights into the structure of the total-cost distribution under stochastic lead times. These visual diagnostics highlight the fundamental median–tail trade-off in CTDE-based inventory control.

Meanwhile BaseStock distribution in Fig. 2 is relatively compact and stable, with moderate dispersion around the median. This behavior reflects its conservative design: by targeting a worst-case protection period, the policy absorbs lead-time variability through higher inventory buffers. As a result, tail costs remain bounded and extreme events are limited. This confirms its role as a robust but conservative baseline that prioritizes stability over cost optimality. The HAPPO distribution in Fig. 3 reveals a strong asymmetry:

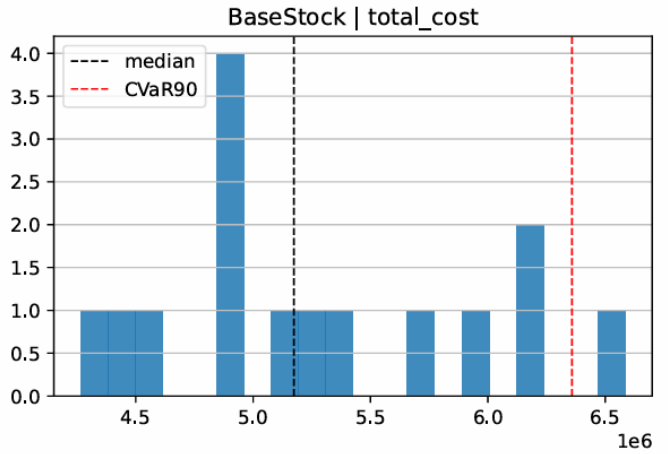

Fig. 2: Distribution of Robust BaseStock episode total cost

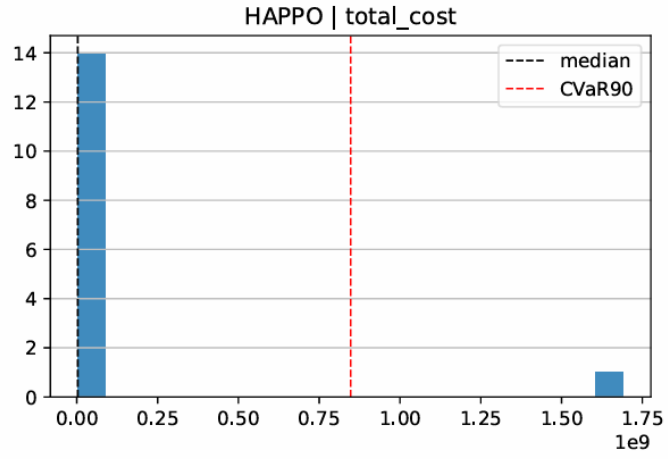

Fig. 3: Distribution of HAPPO episode total cost

most episodes cluster at very low cost levels, indicating highly efficient coordination in typical regimes. However, a small number of extreme outliers appear at very high cost levels, producing a heavy-tailed distribution. This explains the gap between the excellent median performance and the very large CVaR and worst-case values. Operationally, these rare episodes correspond to backlog cascades triggered by unfavorable alignments of long lead times and demand spikes.

Fig. 4 shows that the C-HAPPO distribution is noticeably more compact than HAPPO. While the median cost is slightly higher than the best HAPPO realizations, the upper tail is substantially reduced. The absence of extreme outliers indicates that the CVaR-based constraint effectively suppresses catastrophic backlog cascades. In other words, C-HAPPO trades a small loss in aggressiveness for a significant gain in

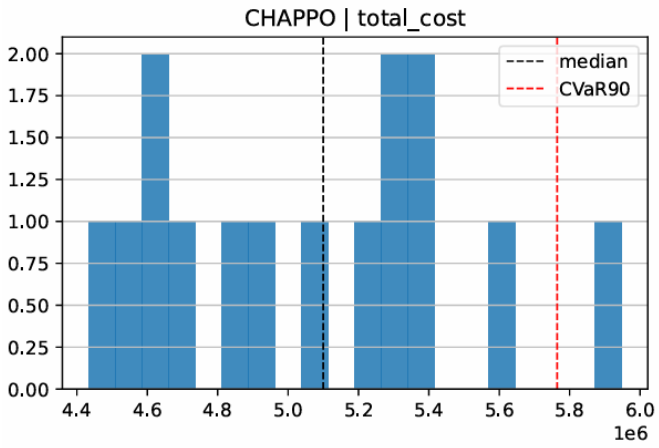

Fig. 4: Distribution of C-HAPPO episode total cost

robustness.

HAPPO achieves the lowest median cost, confirming its strong coordination capability in typical trajectories, but exhibits very large CVaR and worst-case values, indicating vulnerability to rare but severe disruptions. This confirms that explicitly constraining the tail during training effectively stabilizes CTDE policies under stochastic lead times.

Taken together, the figures highlight that evaluating only expected or median performance can be misleading in supply chain settings with stochastic delays. The main operational risk lies in rare extreme episodes rather than typical trajectories. This behaviour of HAPPO is probably due to the intensity of the order-smoothing term which can delay corrective reactions during shock+delay alignments. By explicitly penalizing tail events, C-HAPPO shifts the policy toward safer regimes, resulting in a more favorable cost–risk trade-off. Thus, C-HAPPO adds a training signal that is silent on typical episodes but activates on the upper tail. To prevent tail violations, the policy keeps slightly higher buffers or reacts more conservatively, which makes it less profitable in typical episodes.

VII. Conclusion & Perspectives

This paper studies a controlled serial topology under two sources of uncertainty and introduces C-HAPPO as a risk-aware variant of HAPPO to mitigate catastrophic tails. In fact, stochastic lead times along with stochastic demand induce a strong median–tail trade-off in CTDE multi-echelon control: a CTDE baseline (HAPPO) may look excellent on typical trajectories yet remain brittle in rare backlog cascades that dominate mean cost and risk. Empirically, C-HAPPO remains competitive on mean cost (4% lower than Robust BaseStock) and service, supporting risk-aware CTDE as a practical direction for robust multi-echelon inventory control. It achieves a 60% reduction in cost variance relative to Robust BaseStock.

Next steps include broader demand regimes (seasonality, non-stationarity), alternative lead-time distributions, and larger networks. A systematic sensitivity analysis of RL hyperparameters is particularly important for ensuring stable behavior across regimes. Future work should also compare C-HAPPO with other inventory-control policies and benchmarks.

References

[1] X. Liu, M. Hu, Y. Peng, and Y. Yang, "Multi-agent deep reinforcement learning for multi-echelon inventory management," *Production and Operations Management*, vol. 34, no. 7, pp. 1836–1856, 2024.

[2] G. Wu, M. A. de Carvalho Servia, and M. Mowbray, "Distributional reinforcement learning for inventory management in multi-echelon supply chains," *Digital Chemical Engineering*, vol. 6, p. 100073, 2023.

[3] M. Mousa, D. van de Berg, N. Kotecha, E. A. del Rio Chanona, and M. Mowbray, "An analysis of multi-agent reinforcement learning for decentralized inventory control systems," *Computers & Chemical Engineering*, p. 108783, 2024.

[4] X. Wang and S. M. Disney, "The bullwhip effect: Progress, trends and directions," *European Journal of Operational Research*, vol. 250, no. 3, pp. 691–701, 2016.

[5] A. Oroojlooyjadid, L. V. Snyder, and M. Takáč, "Applying deep learning to the newsvendor problem," *IISE Transactions*, vol. 52, no. 4, pp. 444–463, 2020.

[6] Y. Wang *et al.*, "Deep reinforcement learning for inventory control: A roadmap," *European Journal of Operational Research*, vol. 298, no. 2, pp. 401–412, 2022.

[7] J. Gijsbrechts, R. N. Boute, J. A. Van Mieghem, and D. J. Zhang, "Can deep reinforcement learning improve inventory management? performance on lost sales, dual-sourcing, and multi-echelon problems," *Manufacturing & Service Operations Management*, vol. 24, no. 3, pp. 1349–1368, 2022.

[8] A. Oroojlooyjadid, M. Nazari, L. V. Snyder, and M. Takáč, "A deep q-network for the beer game: Deep reinforcement learning for inventory optimization," *Manufacturing & Service Operations Management*, vol. 24, no. 1, pp. 285–304, 2022.

[9] L. Busoniu, R. Babuška, B. De Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*, 1st ed. Boca Raton, FL: CRC Press, 2010.

[10] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 6379–6390.

[11] A. Clark and H. Scarf, "Optimal policies for a multi-echelon inventory problem," *Management Science*, vol. 6, no. 4, pp. 475–490, 1960.

[12] P. H. Zipkin, *Foundations of Inventory Management*. McGraw-Hill/Irwin, 2000.

[13] K. Muthuraman, S. Seshadri, and Q. Wu, "Inventory management with stochastic lead times," *Mathematics of Operations Research*, vol. 40, no. 2, pp. 302–327, 2015.

[14] S. Cannella and E. Ciancimino, "On the bullwhip avoidance phase: supply chain collaboration and order smoothing," *International Journal of Production Research*, vol. 48, no. 22, pp. 6739–6776, 2010.

[15] F. Costantino, G. Di Gravio, A. Shaban, and M. Tronci, "Smoothing inventory decision rules in seasonal supply chains," *Expert Systems with Applications*, vol. 44, pp. 304–319, 2016.

[16] R. N. Boute, S. M. Disney, J. Gijsbrechts, and J. A. Van Mieghem, "Dual sourcing and smoothing under nonstationary demand time series: Reshoring with speedfactories," *Management Science*, vol. 68, no. 2, pp. 1039–1057, 2022.

[17] K. Geevers, L. van Hezewijk, and M. R. K. Mes, "Multi-echelon inventory optimization using deep reinforcement learning," *Central European Journal of Operations Research*, vol. 32, pp. 653–683, 2024.

[18] P. Hammler, N. Riesterer, and T. Braun, "Fully dynamic reorder policies with deep reinforcement learning for multi-echelon inventory management," *Informatik Spektrum*, vol. 46, pp. 240–251, 2023.

[19] M. Khirwar, K. S. Gurumoorthy, A. A. Jain, and S. Manchenahally, "Cooperative multi-agent reinforcement learning for inventory management," *CoRR*, vol. abs/2304.08769, 2023.

[20] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Policy gradient for coherent risk measures," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 1468–1476.

[21] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 42, pp. 1437–1480, 2015.