Integrated Master Planning and Demand Fulfillment in Multi-Echelon Supply Chains: Partial vs. All-or-Nothing Fulfillment

Felix P. Broesamle[1,*], Hannah Bakker[1], and Stefan Nickel[1]

Abstract

Modern supply chains differ in lead times, bottlenecks, and demand volatility. Advanced Planning Systems conventionally decouple Master Planning (MP) and Demand Fulfillment (DF) hierarchically, which improves tractability but ignores critical interdependencies between material availability, capacity utilization, and delivery promising. We propose an integrated MP-DF model, formulate it as a Minimum-Cost Multicommodity Flow (MCMCF) on a time-expanded graph, and compare partial and all-or-nothing fulfillment policies in a dynamic rolling-horizon simulation. For single-commodity orders with priority-based penalties, partial fulfillment achieves service levels and delivery punctuality indistinguishable from the all-or-nothing policy while reducing runtime by more than an order of magnitude.

Keywords: Supply Chain Optimization, Production Planning, Integrated Planning, Rolling Horizon, Multicommodity Flows

1 Introduction

Many modern multi-echelon supply chains share a geographically dispersed topology yet differ fundamentally in their operational requirements. Fast-moving consumer goods typically feature short lead times, perishable inventory, and stable demand. In contrast, capital-intensive high-tech industries, such as semiconductor manufacturing, pharmaceuticals, and aerospace, face capital-intensive bottlenecks, long lead times, volatile demand, and additional constraints like re-entrant process flows, which are common in semiconductor manufacturing [7].

Advanced Planning Systems [8] commonly manage this complexity via hierarchical decomposition. Master Planning (MP) optimizes mid-term capacity allocation and inter-facility material flows, thereby constraining short-term Demand Fulfillment (DF). DF then applies rule-based Available-to-Promise (ATP) [6] logic to confirm, delay, or reject customer orders. Although computationally tractable and widely adopted, this architecture neglects the interdependencies [5] between material availability, capacity utilization, and order promise logic, which may lead to suboptimal capacity allocation, delivery lateness, and unnecessary rejection of high-priority orders.

[*]Corresponding author.

[1]Discrete Optimization and Logistics, Institute for Operations Research, Karlsruhe Institute of Technology, Kaiserstraße 89, Karlsruhe, Germany. Email: {felix.broesamle, hannah.bakker, stefan.nickel}@kit.edu

Date: August 29, 2026.

An initial integrated MP-DF model considers partial demand fulfillment [3]. Yet practical order promising often follows an all-or-nothing logic: an order is fully confirmed, delayed, or rejected. This raises the question whether enforcing all-or-nothing fulfillment is computationally justified, or whether partial fulfillment is sufficient. We address this by extending the graph-based formulation with all-or-nothing decisions and comparing both policies in a dynamic rolling-horizon setting.

The remainder of this paper is organized as follows. Section 2 formalizes the integrated MP-DF model and its graph-based formulation. Section 3 establishes the rolling-horizon simulation logic. Section 4 reports computational results.

2 Integrated Master Planning and Demand Fulfillment

We formulate a deterministic mathematical model that couples a multi-echelon production network with alternative order-fulfillment paradigms. We present a graph-based Minimum-Cost Multicommodity Flow (MCMCF) formulation of the problem.

The supply chain operates over a discrete time horizon $\mathcal{T} = \{0, \dots, T\}$ and production echelons $\ell \in \mathcal{L} = \{1, \dots, L\}$, where the final echelon (L) represents the customer-facing distribution centers (DCs). Facilities $i \in I_\ell$ have capacities $u_{i,\ell}$ consumed at a rate of $\alpha^k_{i,\ell}$ per unit of commodity $k \in K$, with deterministic lead time $\tau^k_{i,\ell}$. Transportation is allowed from any facility at echelon ℓ to any at $\ell + 1$ and we assume that each transport takes one time-period. Inventory is exclusively held post-production.

The demand portfolio $\mathcal{D} = \mathcal{O} \cup \mathcal{F}$ comprises firm customer orders $\mathcal{O}$ and aggregated forecasts $\mathcal{F}$. Each demand element $d \in \mathcal{D}$ specifies a commodity k_d, a volume q_d, an assigned DC $i_d \in I_L$, a target delivery date π_d, a priority class p_d, and economic penalties for shortfall (γ^{short}_d) and lateness (γ^L_o). Firm orders possess a flexible fulfillment window $\mathcal{W}_o = \{e_o, \dots, l_o\}$, whereas forecasts are temporally fixed, $\mathcal{W}_f = \{\pi_f\}$.

Graph-based Minimum-Cost Multicommodity Flow Formulation
We map the physical network and integrated model onto a tailored time-expanded graph $G = (V, E)$ following the approach presented in [3]. The node set $V = V^{\text{MP}} \cup V^{\text{DF}} \cup \{s\}$ partitions into master planning nodes $v^t_{i,\ell,0}$ (production starts) and $v^t_{i,\ell,1}$ (post-production inventory), time-independent demand nodes v_d, and a global super-source s.

The directed edge set $E = E^{\text{inv}} \cup E^{\text{prod}} \cup E^{\text{aux}} \cup E^{\text{trans}} \cup E^{\text{DF}}$ captures physical material processing. Inventory arcs E^{inv} propagate flow forward in time from $v^t_{i,\ell,1}$ to $v^{t+1}_{i,\ell,1}$ at cost $h^k_{i,\ell,1}$. Production arcs E^{prod} model manufacturing processes via $(v^t_{i,\ell,0}, v^{t+\tau^k_{i,\ell}}_{i,\ell,1})$. Commodity-dependent capacities $u^k_e \in \{0, \infty\}$ enforce that each commodity k follows its prescribed route. Auxiliary arcs E^{aux} connect the super-source s to each first-echelon production start nodes $v^t_{i,1,0}$. Transportation arcs E^{trans} connect post-production inventory nodes to the subsequent echelon's production start nodes with a one-period transit: $(v^t_{i,\ell,1}, v^{t+1}_{j,\ell+1,0})$. Demand Fulfillment arcs E^{DF} $(v^t_{i_d,L,1}, v_d), t \in \mathcal{W}_d$, bridge final DCs to specific demand nodes. On-time fulfillment arcs incur zero penalty, while late arcs carry a time-weighted cost coefficient $c^{k_o}_e = \gamma^L_o \cdot \max(0, t - \pi_o)$. Unfulfilled demand is handled by rejection arcs from s to v_d at cost $c^{k_d}_e = \gamma^{\text{short}}_d$. Unless stated otherwise, arc costs are zero and arc capacities are set to a sufficiently large constant (e.g., the total demand volume). Arcs whose head would lie outside the time horizon $\mathcal{T}$ are omitted during construction, thereby naturally enforcing temporal feasibility.

For partial fulfillment (Approach 1), let $x_e^k \geq 0$ be the flow of commodity k on edge $e \in E$. For each $k \in K$, node demands are $b_s^k = \sum_{d \in \mathcal{D}: k_d = k} q_d$, $b_{v_d}^k = -q_d$ if $k = k_d$, and $b_v^k = 0$ otherwise. The problem is a variant of the MCMCF:

$$\min \quad \sum_{k \in K} \sum_{e \in E} c_e^k \, x_e^k \tag{1a}$$

$$\text{s.t.} \quad \sum_{e \in \delta^{\text{out}}(v)} x_e^k - \sum_{e \in \delta^{\text{in}}(v)} x_e^k = b_v^k, \qquad \forall\, v \in V,\ k \in K, \tag{1b}$$

$$\sum_{k \in K} \sum_{e \in E_{i,\ell}^t} \alpha_{i,\ell}^k \, x_e^k \leq u_{i,\ell}, \qquad \forall\, \ell \in \mathcal{L},\ i \in I_\ell,\ t \in \mathcal{T}, \tag{1c}$$

$$0 \leq x_e^k \leq u_e^k, \qquad \forall\, e \in E,\ k \in K, \tag{1d}$$

where $\delta^{\text{in}}(v)$ and $\delta^{\text{out}}(v)$ denote incoming and outgoing edges of node v, respectively, and $E_{i,\ell}^t \subseteq E^{\text{prod}}$ the production arcs originating at $v_{i,\ell,0}^t$. The shared capacity constraints (1c) couple the commodities and differ from a standard MCMCF [1, 2]. The graph-based formulation exposes the combinatorial structure of the problem.

For the all-or-nothing policy (Approach 2), we introduce a binary variable $y_e \in \{0,1\}$ for every incoming arc of demand node v_d, i.e., the fulfillment arcs from final DCs and the rejection arc from the super-source. Letting E_d^{DF} denote this arc set, we impose the additional constraints

$$x_e^k = q_d^k y_e, \quad \forall d \in \mathcal{D},\ e \in E_d^{\text{DF}},\ k = k_d, \quad \text{and} \quad \sum_{e \in E_d^{\text{DF}}} y_e \quad = 1, \quad \forall d \in \mathcal{D}, \tag{2}$$

which yields a Mixed-Integer Multicommodity Flow (MIMCF) problem. We observe the following property for the above constructed graph $G = (V, E)$.

Proposition 2.1. The time-expanded graph $G = (V, E)$ is a directed acyclic graph (DAG).

Proof. Assign each node a rank: $r(s) = -1$, $r(v_{i,\ell,s}^t) = t$, $r(v_d) = T + 1$. Every arc strictly increases this rank: E^{inv} (t to $t+1$), E^{prod} (t to $t + \tau$, $\tau \geq 1$), E^{aux} (s to $v_{i,1,0}^t$ with $t \geq 0$), E^{trans} (t to $t+1$), E^{DF} (DC node with $t \leq T$ to v_d for fulfillment and s to v_d for rejection). Thus r provides a topological ordering. Therefore, G is a DAG. □

3 Rolling Horizon Logic and State Evolution

We embed the integrated MP-DF model in a discrete-time rolling-horizon simulation with global planning horizon $\mathcal{T}$, forecast-horizon T, and roll step Δ (typically $\Delta = 1$). Each snapshot P_θ optimizes over the next T periods (forecast-horizon); afterwards the state is shifted by Δ. All time indices are relative to the current epoch. Let $x_e^{*,k}$ denote the optimal flow of commodity k on arc $e \in E$ (and y_e^* the binary selection in Approach 2). The transition from P_θ to $P_{\theta+1}$ is governed by three mechanisms: physical state evolution, order management, and demand evolution.

Physical State Injection. Physical flows that bridge the execution window $\mathcal{T}_\Delta = \{0, \ldots, \Delta - 1\}$ become immutable initial conditions for the next epoch. For each master-planning node $v \in V^{\text{MP}}$ we define the induced supply

$$a_v^k \ = \sum_{e=(u,v) \in E^{\text{MP}} \,:\, t_u < \Delta,\, t_v \geq \Delta} x_e^{*,k}, \tag{3}$$

where t_u and t_v are the time indices of the tail and head nodes. In $P_{\theta+1}$, all temporal coordinates are shifted by $-\Delta$, and the induced supplies a_v^k are injected as initial inventory or deterministic work-in-progress (WIP).

Future Commitments. If an order o has a target delivery date inside the execution window ($\pi_o < \Delta$) but the snapshot schedules its fulfillment from a future period $t \geq \Delta$, that shipment is locked as a *Future Commitment (FC)*. Formally, for each distribution center $i \in I_L$ and future period $t \geq \Delta$, $FC_{k,i,t-\Delta}^{\theta+1} = \sum_{o \in \mathcal{O}:\, i_o=i,\, k_o=k, \pi_o<\Delta} x^{*,k}_{(v^t_{i,L,1}, v_o)}$. In $P_{\theta+1}$, these commitments are enforced as deterministic outflows from the corresponding DC inventory. The associated orders are administratively classified as fulfilled and removed from the active demand pool.

Demand Evolution and Dynamic Order Management. Only orders with $\Delta \leq \pi_o, l_o \leq T$ are considered. To evaluate their fulfillment status within the current simulation period, we compute the shortage ratio ρ_o^{short} (Approach 1: $x^{*,k}_{(s,v_o)}/q_o^k$; Approach 2: $y^*_{(s,v_o)}$) and the late-flow ratio $\rho_o^{late} = \frac{1}{q_o^k} \sum_{t>\pi_o} x^{*,k}_{(v^t_{i_o,L,1}, v_o)}$. Active orders are rejected, penalized, and removed from the simulation if $\rho_o^{short} > 0.5$; otherwise, if the order is predominantly late $\rho_o^{late} > 0.7$, its promised date π_o is shifted to the latest period t with positive flow and the earliest arrival date is updated ($e_o \leftarrow \pi_o$), truncating the fulfillment window.

After disposition, all time-dependent parameters (π_o, e_o, l_o) of surviving orders are shifted by $-\Delta$ to align with the new epoch's origin.

4 Computational Experiments

We evaluate the impact of partial versus all-or-nothing fulfillment on solution quality within the rolling-horizon framework. All experiments were performed using `Python 3.13.5` on a workstation equipped with 32 GB RAM and an Intel Core Ultra 7 255 U processor. Optimization models were implemented via the `gurobipy 13.0.2` interface to the commercial solver `Gurobi 13.0.2`, with `Threads` and `Presolve` restricted to 1. LP models (partial fulfillment) are solved with the dual simplex method (`Method`=1); MIPs (all-or-nothing fulfillment) use `MIPGap` $= 10^{-3}$.

Setup. We consider four-echelon networks with $|I_\ell| \in \{5, 10\}$ facilities per echelon and $|K| \in \{3, 5, 7, 10\}$ commodities. Demand streams are generated by the MMFE model [4] with mean volumes of 100–1000 per DC, commodity, and period, disaggregated into firm orders with five priority classes ($p = 1$ lowest, $p = 5$ highest) and fulfillment windows with $e_o = \pi_o, l_o \sim \pi_o + \mathcal{U}\{1, 4\}$. Penalties are $\gamma_o^L = 5p$, $\gamma_d^{\text{short}} = 250p$, and $\gamma_f^{\text{short}} = 240$; holding costs are drawn from $[0.1, 1.0]$. Capacities are calibrated so that 30–85% of the demand volume can be fulfilled per snapshot, mimicking a highly-congested environment. The simulation runs for 36 epochs: a 10-epoch warm-up, solved with the all-or-nothing MIP to initialize the system, followed by 26 evaluation epochs where the policy under test (partial or all-or-nothing) is applied. Each snapshot optimizes over a forecast horizon of $T = 26$ periods with roll step $\Delta = 1$. Cycle times $\tau_{i,\ell}^k$ are sampled from $\{1, \ldots, 4\}$ ($\ell = 1$), $\{1, 2\}$ ($\ell = 2, 3$), and set to 1 ($\ell = 4$); capacity consumption coefficients $\alpha_{i,\ell}^k$ are drawn from $[1.0, 2.0]$ for $\ell = 1$, and set to 1.0 otherwise. All instances were generated using `seed` 2001.

Metrics. We define the Priority-Weighted Order Fulfillment Rate (PWOFR) and the Priority-Weighted Average Lateness (PWAL), by:

$$PWOFR = \frac{\sum_{o\in\mathcal{O}} p_o\, \mathbf{1}_{\{\phi_o=1\}}(o)}{\sum_{o\in\mathcal{O}} p_o}, \qquad PWAL = \frac{\sum_{o\in\mathcal{O}:\phi_o=1} p_o(t_o - \pi_o)}{\sum_{o\in\mathcal{O}:\phi_o=1} p_o},$$

where ϕ_o is the fill rate of order o, and t_o is the absolute period in which it is delivered (latest positive-flow period for partial, and the single period with $y_e = 1$ for all-or-nothing). PWOFR counts only fully fulfilled orders weighted by priority, PWAL measures the average delay of those orders, also priority-weighted. FR denotes the total volume fill rate over all demands.

Results. Tables 1 and 2 report the outcomes for $|I_\ell| = 5$ and 10. Across all instances, PWOFR differs by at most 0.01 and PWAL by at most 0.05 between partial and all-or-nothing fulfillment policies, while the all-or-nothing policy requires $\geq 10\times$ more runtime. Because shortage penalties dominate lateness penalties and both scale with priority, the model primarily favors serving high-priority orders. Fractional allocations therefore remain small or do not change the resulting order-level classification. Altogether, for single-commodity orders, partial fulfillment provides the same operational performance as all-or-nothing fulfillment at a fraction of the computational cost.

Table 1: Priority-weighted service quality and runtime for $|I_\ell| = 5$.

K	Policy	Ful.	Part.	Rej.	FR	PWOFR	PWAL	Runtime (s)
3	Partial	1883	23	7591	0.157	0.273	0.917	11.4
3	All-or-nothing	1900	0	7605	0.163	0.276	0.882	143.2
5	Partial	3703	75	11374	0.283	0.346	1.169	21.6
5	All-or-nothing	3678	0	11476	0.285	0.343	1.171	285.7
7	Partial	4938	125	16702	0.232	0.345	1.012	50.5
7	All-or-nothing	4929	0	16808	0.231	0.344	1.024	462.8
10	Partial	4704	209	26215	0.147	0.241	0.910	45.5
10	All-or-nothing	4587	0	26541	0.150	0.235	0.955	469.6

Table 2: Priority-weighted service quality and runtime for $|I_\ell| = 10$.

K	Policy	Ful.	Part.	Rej.	FR	PWOFR	PWAL	Runtime (s)
3	Partial	9018	112	9203	0.478	0.673	0.925	51.3
3	All-or-nothing	9104	0	9205	0.487	0.681	0.928	526.5
5	Partial	18316	119	10554	0.631	0.778	1.009	104.8
5	All-or-nothing	18369	0	10612	0.634	0.778	1.004	3458.1
7	Partial	29891	121	12030	0.687	0.825	0.567	431.3
7	All-or-nothing	29769	0	12237	0.687	0.823	0.569	5470.5
10	Partial	26712	479	31847	0.456	0.627	0.969	438.6
10	All-or-nothing	26555	0	32511	0.455	0.624	0.983	5506.2

We remark that the graph-based MCMCF again showed the runtime advantage over an equivalent LP formulation as already observed in [3].

5 Conclusions

We introduced an integrated MP and DF model, formulated as a MCMCF on a time-expanded graph, and embedded it in a rolling-horizon simulation. The computational results indicate that, for single-commodity orders, partial fulfillment achieves service levels nearly indistinguishable from the all-or-nothing policy, while reducing runtime by more than an order of magnitude. This suggests that, under such penalty structures, enforcing all-or-nothing fulfillment adds computational complexity without materially changing service levels. Future work will exploit the exposed network structure with specialized algorithms and extend the framework to stochastic demand and multi-commodity orders.

References

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Upper Saddle River, NJ, 1993.

[2] A. A. Assad. Multicommodity network flows—a survey. *Networks*, 8(1):37–91, 1978. `doi:10.1002/net.3230080107`.

[3] H. Bakker, F. P. Broesamle, F. Fiand, J. Hasselbring, M. Hess, and S. Nickel. A graph-based rolling-horizon approach for integrated master planning and demand fulfillment in semiconductor manufacturing. In *Proceedings of the 2026 Winter Simulation Conference*, 2026. To appear.

[4] D. C. Heath and P. L. Jackson. Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. *IIE Transactions*, 26(3):17–30, 1994. `doi:10.1080/07408179408966604`.

[5] R. Herding and L. Mönch. A rolling horizon planning approach for short-term demand supply matching. *Central European Journal of Operations Research*, 32(3):865–896, Sep 2024. `doi:10.1007/s10100-023-00891-z`.

[6] H. Meyr, M. Wagner, and J. Rohde. *Structure of Advanced Planning Systems*, pages 99–106. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. `doi:10.1007/978-3-642-55309-7_5`.

[7] L. Mönch, R. Uzsoy, and J. W. Fowler. A survey of semiconductor supply chain models part iii: master planning, production planning, and demand fulfilment. *International Journal of Production Research*, 56(13):4565–4584, 2018. `doi:10.1080/00207543.2017.1401234`.

[8] H. Stadtler, B. Fleischmann, M. Grunow, H. Meyr, and C. Sürie. *Advanced Planning in Supply Chains: Illustrating the Concepts Using an SAP® APO Case Study*. Springer, 2012.