informs
https://pubsonline.informs.org/journal/ijaa

INFORMS JOURNAL ON APPLIED ANALYTICS
Vol. 54, No. 1, January–February 2024, pp. 5–19
ISSN 2644-0865 (print), ISSN 2644-0873 (online)

THE FRANZ EDELMAN AWARD
Achievement in Operations Research

Optimizing Walmart's Supply Chain from Strategy to Execution

Prakhar Mehrotra,[a] Mingang Fu,[a] Jing Huang,[a,*] Sai Rajesh Mahabhashyam,[b] Minghui Liu,[a] Ming (Arthur) Yang,[b] Xiaojie Wang,[a] Joseph Hendricks,[b] Ranjith Moola,[a] Daniel Morland,[b] Kim Krozier,[b] Tiantian Nie,[a] Ou Sun,[a] Fereydoun Adbesh,[a] Ti Zhang,[a] Monika Shrivastav,[a] Jiefeng Xu,[b] Sudarshan Rajan,[b] Michael Turner,[b] Samuel Tucker,[b] Megan D. Jones,[b] Fei Xiao,[a] Ankush Bhargava,[a] Deepak Deshpande,[a] Shwetal Mokashi,[a] Travis Johnson,[b] Chandramouli Raman,[b] Megan Ferguson,[b] Mike Keller,[b] Scott Donahue,[b] Rajiv Bhutta,[a] Mohan Akella,[b] Parvez Musani,[a] Srinivasan Venkatesan,[a] David Guggina,[b] John Furner[b]

[a] Walmart Global Tech, Sunnyvale, California 94086; [b] Walmart, Bentonville, Arkansas 72716
*Corresponding author

Contact: prakhar.mehrotra@walmart.com (PrM); mingang.fu@walmart.com (MiF); jing.huang@walmart.com, https://orcid.org/0009-0000-0201-1975 (JiH); sairajesh.mahabhashy@walmart.com (SRM); minghui.liu@walmart.com (ML); arthur.yang0@walmart.com (M(A)Y); xiaojie.wang@walmart.com (XW); Joseph.hendricks@walmart.com (JoH); ranjith.moola@walmart.com (RM); daniel.morland@walmart.com (DM); kim.krozier@walmart.com (KK); tiantian.nie@walmart.com (TN); ou.sun@walmart.com (OS); fereydoun.adbesh@walmart.com (FA); ti.zhang@walmart.com (TZ); monika.shrivastav@walmart.com (MS); jiefeng.xu@walmart.com (JX); sudarshan.rajan@walmart.com (SR); michael.turner@walmart.com (MT); samuel.tucker@walmart.com (ST); megan.d.jones@walmart.com (MDJ); fei.xiao0@walmart.com (FX); abhargava01@walmart.com (AB); deepak.deshpande@walmart.com (DD); shwetal.mokashi@walmart.com (SM); travis.johnson@walmart.com (TJ); craman@walmart.com (CR); megan.ferguson@walmart.com (MeF); mike.keller@walmart.com (MK); scott.donahue@walmart.com (SD); rajiv.bhutta@walmart.com (RB); makella@walmart.com (MA); parvez.musani@walmart.com (PaM); srinivasan.venkatesan0@walmart.com (SV); david.guggina@walmart.com (DG); furner.john@walmart.com (JF)

Accepted: November 27, 2023

https://doi.org/10.1287/inte.2023.0093

Abstract. Walmart takes a holistic approach to its supply chain, which integrates strategic and operational decisions consisting of three components: (1) network planning and transformation, which recommends the long-term network with step-by-step recommendations on how to achieve the end state; (2) a routing and loading system (Load Planner), which determines how to efficiently move products across the network; and (3) a simulation platform, which combines strategic and executional decision engines to enable a holistic decision-making process. Walmart built a set of scalable and fast optimization decision engines and deployed them using underlying innovative algorithms and models. The company fully adopted this next-generation optimization capability throughout its entire grocery supply chain in the United States, and approved optimization model-based network design and transformation plans for long-term investments involving billions of dollars. As a result of this efficient routing and loading executional system, Walmart prevented 98.6 million pounds of CO_2 emissions and saved \$91.5 million by eliminating 108,000 truck routes covering 33 million miles in fiscal year 2023 (FY23). Moreover, this optimization-empowered decision framework is evolving and transforming Walmart's supply chain while keeping its Every-Day-Low-Price (EDLP) promise to its customers.

Keywords: Edelman Award • supply chain optimization • network design • simulation • truck routing and loading • mixed-integer programming • metaheuristics

Walmart Supply Chain

Walmart operates a complex system of supply chain facilities in the United States. This includes 117 distribution centers (DCs), 26 fulfillment centers, 3 sortation centers, 96 transportation offices, around 600 Sam's Clubs, and more than 4,700 stores, which cover 90% of the U.S population within a 10-mile radius. This massive supply chain network serves as the backbone for providing customers with a seamless and convenient shopping experience.

One of the critical foundations of Walmart's supply chain operations is the applications and systems that support supply chain management across the decision tiers from strategy to execution. The work we describe in this paper is the most recent step in the evolution of these applications and systems.

Need for Next-Generation Optimization Capabilities

Walmart is a leading retailer whose customers can choose to shop in stores, pick up orders at curbside, or have their orders delivered from our stores to their homes. The transition into an omnichannel retailer brought significant changes in the demand patterns that its brick-and-mortar stores faced. Furthermore, the macroeconomic changes caused by COVID-19 have brought greater ambiguity to consumer demand. In an era of high inflation, people rely on Walmart and its Every-Day-Low-Price (EDLP) promise. Meanwhile, the recent advancements in warehouse automation technologies, especially in the cold chain space, have unlocked enormous potential to dramatically improve productivity. Therefore, strategically transforming the supply chain network, on a large scale and in a short timeframe, was essential. The fast-changing demand patterns with high ambiguity created and continue to create unprecedented challenges for us as we seek to build resilience into our supply chain.

At the execution level, Walmart has a rich history of adopting industry-leading supply chain optimization technologies to continuously drive operational costs down. It has reached a point where any incremental cost reduction on standalone systems becomes challenging. For example, in 2021, for the delivery of dry groceries, using our previously deployed multiple-stop loading system, our trucks were nearly fully utilized in terms of weight and space capacities, providing little opportunity for further utilization increases. A breakthrough in optimization technologies was required to drive continuous improvement.

A Multiple-Tier Decision Framework with Advanced Optimization Capabilities

To tackle the challenges we describe, we identified opportunities in an end-to-end optimization framework that ranges from network strategy to execution optimization technologies. We see a continuous improvement cycle between network strategy and execution optimization. A supply chain network designed and planned for higher efficiency could unlock the greater potential of the optimization applications, while faster optimization applications can enable more simulation runs, which allow the evaluation of more scenarios, and thus improve the probability that executive decision makers will adopt the recommendations.

Walmart follows a four-step hierarchical decision process for designing, planning, and executing its supply chain (Figure 1). It starts with strategic network design decisions such as investing in building a new DC or retrofitting an existing DC by looking ahead a few years. Next, it determines facility alignment (e.g., which stores should be served by which DC) and capacity planning at each facility on a yearly or quarterly basis. It then makes execution decisions, such as how to efficiently pick, load, and route items, prior to moving products. Finally, as items are physically moved through the supply chain, it makes dynamic decisions to help its associates adjust to the plan in real time.

Achieving continuous improvement in this multiple-tier decision process is difficult for two primary reasons: (1) considering the most granular level of operating costs at the strategic planning level makes the problem intractable and (2) optimization applications need to run much faster than they do currently if they are to evaluate more scenarios. Therefore, we decided to build our next generation of optimization capabilities with advanced decision engines plugged into each critical decision step and with a tightly coupled information feedback flow (Figure 1). At the strategic and tactical levels, a set of scalable optimization models outlines the future-state network and a transformation roadmap with detailed year-by-year planning to move from the current state to a future state. The outputs of these models also enable resource alignment and planning across the supply chain. At the execution level, the state-of-the-art optimization engines enable optimized operations in a dynamic fashion. More importantly, optimization engines enable a simulation capability to evaluate the impact of various scenarios, while making strategic and tactical decisions. Building these optimization capabilities was not trivial; we faced numerous business and technology challenges.

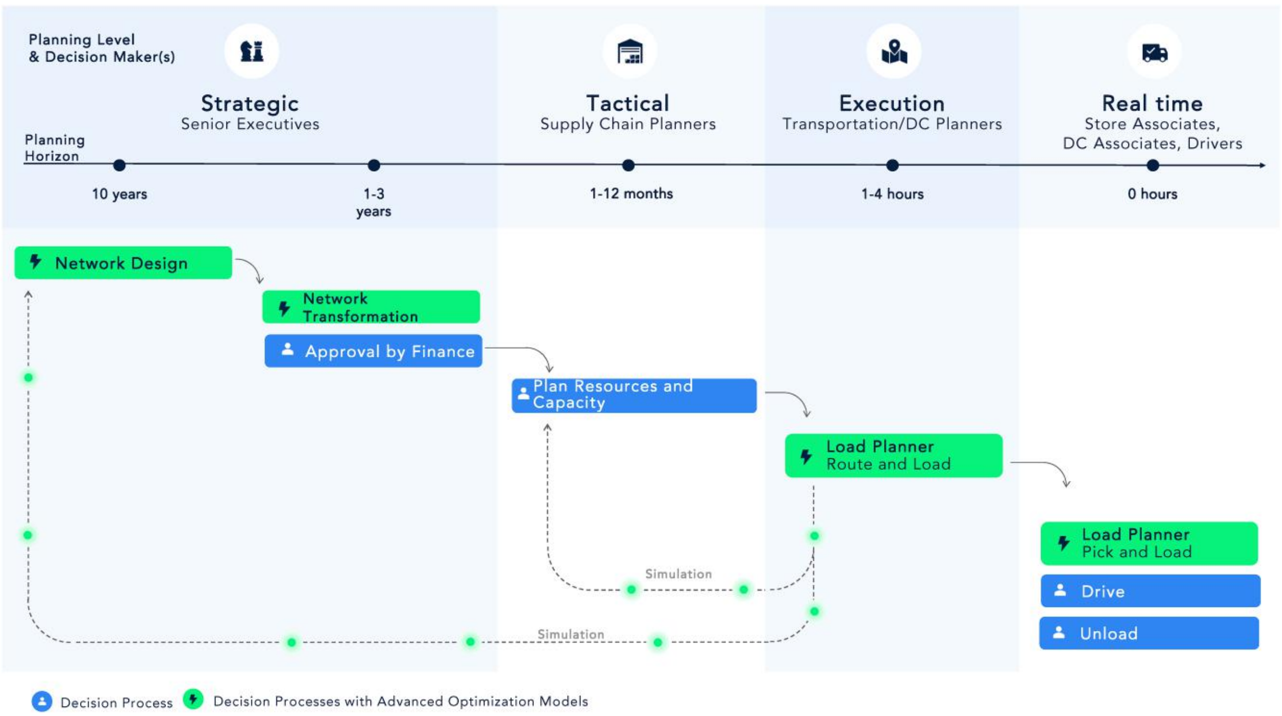
Next, we will discuss in further detail our next-generation optimization capabilities across the supply chain using grocery DC (GDC)-to-store outbound operations as an example. A GDC specializes in storing and distributing dry food and perishable commodities (e.g., fresh produce and meats). With the recent hypergrowth in online channels, the grocery segment of merchandising became the strategic growth engine for Walmart. Given shorter life cycles and cold chain compliance necessary for grocery commodities, especially perishable items, faster turnover and special handling are required in all stages of inventory movement, including storage, transportation, and stocking. Compared with general merchandising, the complexities of moving grocery products, which range from network design to daily routing and loading execution, are significantly greater, creating more opportunities for improving both cost efficiency and the experiences of customers and associates. With minor modifications the models we discuss in this paper are applicable to general merchandising products.

Network Design and Transformation Planning Solution

Introduction

Designing and transforming supply chain networks are the first two steps in Walmart's integrated decision process. The transformation process of each GDC usually takes several years from the design stage to the completion of all construction. It is therefore crucial to make long-term strategic decisions such as retrofitting existing GDCs with next-generation fulfillment capabilities or building greenfield sites. We define two sequential decision steps: (1) define the future grocery network and (2) generate a transformation roadmap.

Figure 1. (Color online) Walmart's Integrated Decision Processes Empowered by Advanced Optimization Capabilities



Note. Load Planner is a system that plans and optimizes Walmart's truck routes and loads.

The objective of designing the future grocery network is to minimize the total capital and operational costs of the network. The costs considered include (1) fixed costs of running each GDC; (2) volume-based variable cost, which is dependent on the type (e.g., automation level) of each GDC; and (3) transportation costs between GDCs and stores. The problem is constrained by practical and business considerations. The output of the process includes (1) the locations of the GDCs in the future grocery network, (2) the type and size of the GDC at each location, and (3) the stores that are aligned to each GDC.

The process of retrofitting an existing manual GDC or building a new GDC involves land purchase (if applicable), site permitting, building construction, installation of automation equipment, and equipment commissioning. Given the defined future network, the transformation plan we develop must be feasible (i.e., customer service cannot be interrupted during the transformation) and the cost must be minimal.

Challenges and Complexities

Walmart's GDC network supports around 600 Sam's Clubs and more than 4,700 stores. The supply chain provides unparalleled access to everyday essentials for our customers. Given the structure and size of the network, there are two main contributors to the complexity of the problem: (1) the interdependent constraints based on various business considerations and assumptions and (2) the computational difficulty caused by the extremely large scale of the problem size.

Business Process Complexity. Each type of GDC conversion (e.g., retrofit) must be laid out with transformation assumptions about moving from the current state to the future state. This breakdown provides a view of how costs, capacity, and capital outlay will occur over time. For new greenfield and freezer assets, the breakdown includes time for real estate and construction activities, material handling equipment (MHE) installation, and building ramp plans. For retrofits, these assumptions include how some legacy capacity will need to be removed to provide sufficient space to allow the retrofit work to begin. Assumptions differ based on the size of the GDC.

Computational Complexity. We can theoretically build a greenfield GDC at any location within the continental United States. Even if we discretize the map by zip code, there are over 30,000 locations from which to choose. Each location would also have approximately 31–78 type/size options. Moreover, each GDC could be aligned with any of the 600 Sam's Clubs and 4,700 Walmart stores. Given that this is a mixed-integer programming (MIP) problem, solving it in a reasonable time using commercial optimization solvers is impossible.

A multiple-year transformation plan results in a multiple-period problem. That is, for each GDC, the transformation status in the current period is dependent on its previous status. Meanwhile, the GDC-to-store alignment in each period is dependent on the availability of the GDCs. Furthermore, because considering the GDC-to-store alignment in each period is required for feasibility and optimality, the problem scale increases further.

Solution Approach

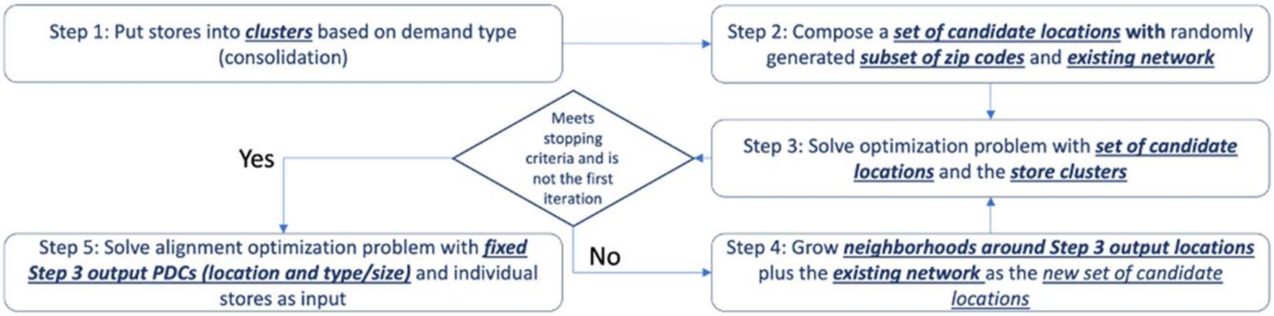
We develop tailored models and algorithms to effectively address challenges and solve the problems. We first construct a MIP model, which considers demand, node capabilities, automation options, operating costs, capital, and outbound transportation costs to recommend an optimal future-state network. We develop an iterative metaheuristic (Figure 2) to solve the MIP problem in a reasonable amount of time.

In Step 1, we first cluster demand (items in cases ordered by stores) based on its type (e.g., freezer items, fresh items) to reduce the size of the input variables. We generate a subset of zip codes randomly together with the existing GDC locations as the candidate locations of the new greenfield GDCs in Step 2. In Step 3, we solve a network optimization model using the data generated from the previous two steps. If the result does not meet the stopping criteria or if it is the first iteration, we move to Step 4 to determine an updated candidate location set for the greenfield GDCs and then loop back to Step 3. Otherwise, we move to Step 5, in which we use the resulting GDCs as a fixed network and calculate the optimal store assignment to GDC. The stopping criteria can be a comparison between the resulting GDCs of the last two consecutive iterations. For example, if the two resulting networks (GDC locations and types/sizes) are sufficiently similar, where distances of GDC locations and sizes are within a tolerance factor defined based on business understanding, the stopping criterion is met.

Once the optimal future-state network has been defined, we need to generate the optimal network transformation roadmap. We categorize the model constraints of this problem into two groups: (1) a scheduling problem, which describes the dependent transformation stages (e.g., building construction, installation of automation equipment, and equipment commissioning) of each GDC among different periods, and (2) assignment problems of each period, which model the detailed GDC-to-store alignment. These two groups of constraints are interconnected by shared decision variables on the transformation process of each GDC and the objective function on the total cost and capital investment. Appendix A shows the detailed formulation for the scheduling problem. The formulation for the assignment problem is similar to the model in network design; however, it adds a time dimension to all the variables and parameters. The result of this model is a detailed transformation schedule (e.g., when to start the transformation and how each GDC evolves over time) for each GDC involved and the corresponding optimal store alignment during each period. Because the scale of the problem is largely dependent on the number of periods considered, we hierarchically solve it with larger increments of periods at the beginning and break down each period into smaller increments later.

Note that we perform multiple-scenario runs with deterministic models for a more comprehensive presentation of the results under different assumptions, to account for stochastics of assumed parameters. We highlight two key innovations in network design and transformation planning in our solution. The first innovation is our algorithm for evaluating candidate locations for the future grocery network. The concept of doing neighborhood generation of the candidate GDC locations was motivated by the local search method and improving the resulting location in each iteration. In our proposed algorithm, only a smaller problem with a subset of all candidate locations is needed for consideration each time. Given that this MIP problem is NP-hard and exponentially increases in scale as more candidate locations are considered, this reduction in the problem scale significantly improves the solution efficiency. The second innovation is our network transformation modeling and algorithm. We developed a novel approach for the transformation optimization model. Walmart had done no previous work on this,

Figure 2. (Color online) Steps for Defining the Future Grocery Network Within the Solution Algorithm



Note. A perishable DC (PDC) is a temperature-controlled segment of a GDC, which is powered by freezers.

nor did it exist in any commercial software suite that we explored.

Load Planner: Routing and Loading Solution

Introduction

Following the four-step decision process illustrated in Figure 1, once facility alignment and capacity at each facility are determined, execution planning determines how to pick, load, and deliver the products before they are physically moved. Walmart has always been at the forefront of leveraging optimization technologies to make the best operational decisions. Over the past decades, its operational efficiency has improved continuously, as evidenced by the near-full utilization of trucks for grocery deliveries of dry goods while meeting store demand. Given the rapid volume growth and evolution of Walmart's supply chain, we decided to build the next-generation outbound routing and loading optimization system, Load Planner, with a goal of simplifying, automating, and optimizing supply chain outbound execution processes.

Load Planner runs each day to create routing plans in which each route specifies a sequence of stops, including detailed time information about all the activities involved in the trip; examples include driving, breaks, and waiting and layover times. To ensure the feasibility of a route, each route generated at this stage is guaranteed to have a feasible truck-load design that complies with all truck-loading constraints. The transportation command center associates review and make any necessary changes with some exceptions (e.g., certain roads are not operable on that day) with the help of functionality provided by Load Planner. Next, Load Planner generates the best arrangement of pallets, configuration of compartments, and temperature settings in trucks. DC associates review the load plans and make additional revisions, if necessary. Finally, DC associates start to pick and load products by following the plans. Typically, associates only need to follow the load plan and load pallets in sequence. However, in unforeseen situations, such as when pallets planned to be loaded at the front of the trailer are not ready or are canceled, or an associate loads a pallet into an undesignated floor spot, Load Planner can help associates dynamically adjust the load plan to ensure all loading compliances are met.

Figure 3 outlines detailed decision steps for current grocery outbound executions, with Load Planner's backbone decision modules to support each decision step. Please note that the planning of picking trips (sequence of picking cases to build pallets in DCs) is based on simple rules; therefore, Load Planner does not address it.

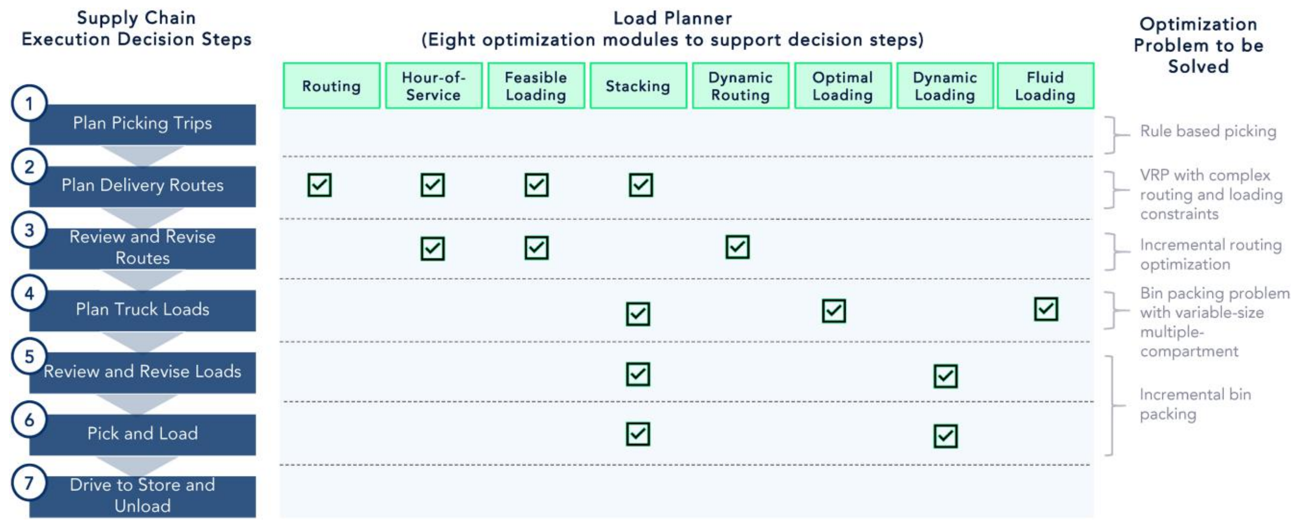
At the core of Load Planner, eight major optimization modules collectively empower each execution decision.

1. *Routing* generates optimal multiple-stop routes that deliver pallets to stores within a store's receiving time window.

2. *Hour-of-Service* (HOS) checks quickly (i.e., within milliseconds) if a route sequence complies with delivery time windows and U.S. Department of Transportation (DOT) HOS regulations. If compliant, the module determines when a driver should take a break and layover on the route and estimates the total transit time.

3. *Feasible Loading* checks quickly (i.e., within milliseconds) if the given pallets and their delivery sequence can yield a feasible load plan that complies with all loading constraints.

Figure 3. (Color online) Decision Process Flow Including Eight Optimization Modules That Support Each Step and Optimization Problems That Are Solved During Each Step



4. *Stacking* generates plans to stack one pallet on top of others to save floor space (stacking). We refer to such stacked pallets as a stack.

5. *Dynamic Routing* replans the stop sequence for one route given a changed set of pallets or validates a predefined stop sequence.

6. *Optimal Loading* generates an optimal load plan, including the orientation and position for each pallet, compartment configurations, and temperature settings, to ensure all loading regulations are met.

7. *Dynamic Loading* provides an optimal plan for loading a set of pallets onto a truck that has been partially loaded.

8. *Fluid Loading* identifies a group of stacks where the positions of the stacks are interchangeable within the block.

Challenges and Complexities

With recent advancements in computing technologies and handheld devices, we believed we could build a faster and better system (Load Planner) to support both execution planning and dynamic executions. Yet, we were facing three major challenges from both business and algorithm perspectives.

Handling Grocery Products. Grocery products, especially perishable commodities, require strict adherence to cold chain compliance throughout the entire storage, handling, and transportation processes. They require the use of temperature-controlled trailers to maintain the proper temperature range. In the United States, a typical 53-foot temperature-controlled trailer (e.g., a tri-temp trailer) features two bulkheads that can divide the interior into up to three compartments, each of which can be set to a different temperature. Figure 4 shows a top view of pallet arrangements in a reefer trailer, which is commonly used for transporting perishable commodities.

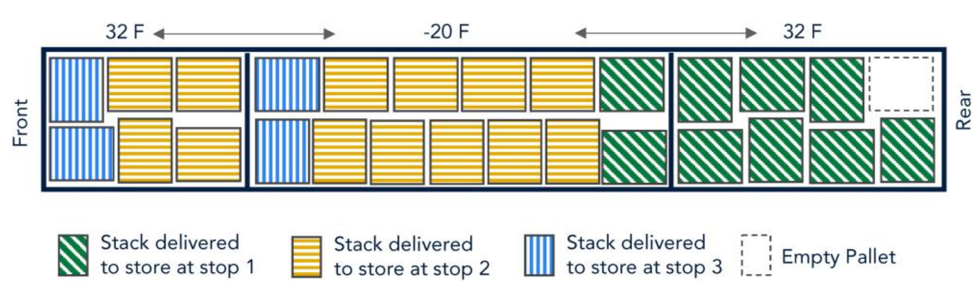
A typical load plan includes the placement of pallets, bulkhead positions, temperature settings for each compartment, and (if needed) the empty pallet (i.e., a stack of pallets) at the rear of the trailer for stability purposes. The task of optimally arranging pallets within a trailer is a complex bin packing problem, and multiple variable-size compartments with different temperature settings make it more difficult to solve. In addition, a set of cold chain requirements makes the problem more challenging. For example, we prohibit the delivery of freezer pallets on the tail of a trailer and require that the temperature difference between adjacent compartments be kept within a specific range. The combination of additional constraints and decision points makes this variable-size, multiple-compartment loading problem more difficult.

Establishing Intelligence and Transparency. One of the major motivations in building a new generation of routing and loading systems is to improve the associates' experience; thus, the users must have a high degree of trust in the new system. We believe there are two ways to establish this trust: (1) provide operationally friendly plans, and (2) make the decision process transparent and interactive.

To provide a more operationally friendly plan, we must tackle many challenging problems, such as minimizing reloads. A reload occurs when store associates must first unload pallets to be delivered to other stores prior to unloading their own store's pallets and must later load the other stores' pallets back onto the truck. Figure 5 illustrates an example of the reload of pallets, where pallets delivered to a store must be unloaded to unblock the other two pallets in another compartment. For perishable products, even when we force a first-in-last-out (FILO) sequence for the pallets in each compartment, we cannot eliminate the necessity for reloads using existing technologies. However, by using advanced algorithms, we can dramatically reduce the incidence of reloads. Doing so increases the complexity of the loading problem and transforms it into a multiple-objective problem.

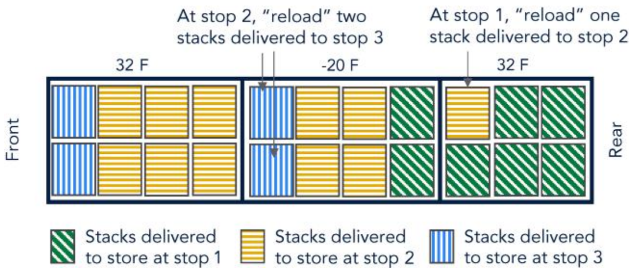
Making the decision-making process transparent is also a challenging task that requires many supporting functions. For example, in our routing optimization, we allow users to edit solutions by adding, removing, or reoptimizing routes. To support this, we needed to develop fast-running algorithms in the background to

Figure 4. (Color online) Stack Arrangements for a Temperature-Controlled Trailer



Notes. The bulkheads are constrained to move only within a predefined range, as we illustrate with arrows to accommodate the position of air conditioning fans. Each square represents a stack; each pattern indicates a different destination.

Figure 5. (Color online) Load Map Demonstrating Reloading of Stacks at Store Stops



Notes. In this example, when the trailer arrives at stop 1, we must unload one stack to be delivered to stop 2 to clear the access to the middle compartment to continue to unload two stacks for stop 1. After the completion of the unloading, this stack needs to be reloaded back into the truck. Similarly, when arriving at stop 2, two stacks to be delivered to stop 3 in the middle compartment need to be unloaded and reloaded back into the truck.

validate the feasibility of a route or incrementally improve existing routes. Developing each of these algorithms was a nontrivial task.

Building Fast, Flexible, and Scalable Algorithms. The vehicle routing problem (VRP) was initially introduced in Dantzig and Ramser (1959) and later in Clarke and Wright (1964) to optimize route plans from a central depot to geographically scattered customers using a fleet of trucks with varying capacities. VRP and its variant problems have grown ever more popular in the past decades. We refer the reader to Laporte (2009) for solution methods and to Eksioglu et al. (2009) and Braekers et al. (2016) for problem categories and classifications. Two major loading problems, which the optimization community studies in conjunction with VRP, are the two-dimensional bin packing problem (2BPP) and the three-dimensional bin packing problem (3BPP). Martello and Vigo (1998) propose the exact solution and lower bounds for 2BPP, and Martello et al. (2000) propose the same for 3BPP. The integration of routing and loading problems is a challenging but promising research area because of its wide applications in real-world transportation. Iori and Martello (2010) and Pollaris et al. (2015) review recent works related to VRPs with loading constraints.

Despite the extensive research and study of both the vehicle routing and bin packing problems in academia and industry, we discovered that the conventional methodologies are inadequate for our specific problem settings. Our solutions must consider over 50 unique rules, which we summarize in Table 1. In addition to the traditional constraints of delivery time window and load capacity in terms of dimension, volume, and weight, these rules include additional considerations of transportation safety, operational efficiency, and user experience. Some constraints are specific to a DC, store, or equipment, further complicating the problem. For example, the routing and loading requirements for offshore stores (e.g., in Alaska, Hawaii, and Puerto Rico) differ from those of stores within the contiguous United States. Offshore routing requires the ability to ship dry and perishable pallets together in the same trailer with certain limitations. Our operations are continuously evolving, and new requirements are constantly arising; therefore, it is crucial to have a solution framework that is adaptable to accommodate these ever-changing requirements.

Moreover, the vast scale of the outbound grocery network requires the design of scalable algorithms and solution architectures. Our solution empowers decision making for a network that delivers more than hundreds of thousands of pallets of items from 47 grocery DCs to over 600 Sam's Clubs and 4,700 Walmart stores, comprising multiple functionalities with varying computational time requirements. For example, Load Planner commits a maximum of 15 minutes to plan delivery routes for a DC, and one second to plan an efficient truck load or dynamically adjust a truck load. The last, but not the least, challenging aspect is that as the routing algorithm searches for the best routes, it constantly checks if the given route is feasible. This requires the HOS module and feasible loading module to provide a response within milliseconds.

Solution Approach

Because of the complexity of the integration of routing and loading problems, most problems are solved by heuristic methods. Based on the findings in Côté et al. (2017), although the integrated problem is solved heuristically, the solution is significantly better, both theoretically and empirically, than separate solutions. We developed a metaheuristic-based framework, which integrates a suite of algorithms including various neighborhood searches, heuristics, and MIP models. For each module, we selected the best algorithm and parameter settings based on learnings from extensive experimentation using historical data. The framework provides us with the flexibility of adding incremental features, as well as high computational efficiency, which has been a general challenge when solving NP-hard problems. In the following sections, we provide a more in-depth explanation of our solutions for each module.

Optimizing Routes While Considering Loads. In this section, we outline our solution approach for routing modules (Fu et al. 2020, Huang et al. 2020a, Liu et al. 2020). It is important to note that the loading problem is seamlessly integrated into this solution framework in two ways: first, load construction and sampling are performed in each local search iteration of the route improvement algorithm to determine if a given route

Table 1. Outbound and Loading Rules We Follow

<table>
<tr><th>Constraints</th><th>Pallet stacking</th><th>Loading</th><th>Routing</th></tr>
<tr><td>Classic constraints</td><td>• Stack cannot exceed the trailer height.</td><td>• Stacks' total weight and cubic volume cannot exceed the trailer weight limit.
• Stacks (i.e., width, length, height) must fit into the trailer.</td><td>• The arrival time at each delivery store must be within that store's delivery time window.
• Start time and end time of a route must be within a time range.</td></tr>
<tr><td>Safety constraints</td><td>• Pallets carrying certain high-risk types of products (e.g., chemicals) cannot be stacked.</td><td>• Weight distributed to front and rear axles must be within a prespecified range throughout the delivery trip.
• The difference between curbside and roadside weight must be within a threshold limit.
• The rear axle of certain trailer types can be adjusted within a range limit.
• Perishable items must be held in the refrigerated compartment within certain temperature settings.
• The gaps between two columns of pallets must be limited to a given threshold.
• Empty pallets and airbags must be used when applicable for stability.</td><td>• Routes should comply with DOT HOS regulations:
1. 11-hour driving limit
2. 14-hour limit
3. 30-minute driving break
• If delivery is handled by day cab driver, driver must return to the distribution center on the same day.</td></tr>
<tr><td>Distribution center operations constraints</td><td>• If there are empty spots in the trailer, pallets should not be stacked.</td><td>• Loading pattern should be loader friendly with limited stacking operations.</td><td>• Two pallets that are picked in the same picking trip must be delivered in the same truck to their destination stores.</td></tr>
<tr><td>Store operations constraints</td><td>• Pallets to be delivered to different stores cannot be stacked.
• Pallets carrying items with different temperature ranges (e.g., ice cream and bananas) cannot be stacked.</td><td>• Reloads are not allowed for delivery of dry grocery items.
• Reloads for delivery of perishable items must be minimized.
• Many specific loading rules apply to offshore loading (e.g., frozen compartment must be at the front of the truck).</td><td>• Route must honor the delivery position specified by the store if it exists.
• Store can make a request to not split delivery orders into multiple trucks.
• A drop-and-hook store must be the last stop and the trailer must be dropped at this store.
• Delivery routes must consider backhaul loads' origination location and time windows.</td></tr>
</table>

Note. Drop-and-hook is the trucking industry's term for when a driver drops a full container at a facility and hooks the tractor to a preloaded trailer at the same facility.

can yield a feasible load plan; second, utilizing load information helps in the design of customized algorithms and in achieving better solutions.

Our approach to the routing problem involves two steps: initial route construction and route improvement. Initial route construction can be modeled either naturally as a VRP or as an assignment problem by assigning pallets or groups of pallets to a set of route templates. The routes are then improved with a customized tabu search framework that includes neighborhood searches such as one-zero exchange, one-one exchange, two-opt, and large neighborhood search. One-zero exchange helps reduce the number of routes, whereas the latter three focus on minimizing travel distance. We tailored our implementation of tabu search by using techniques (e.g., approximate tabu list matching, dynamic tabu tenure, intelligent restart) to improve its performance. The routing module is enabled by three other modules—Feasible Loading, Stacking, and HOS.

For the Feasible Loading module, our algorithm consists of four crucial steps: (1) constructing stacks, (2) assigning stacks to compartments, (3) determining floor spot positions to meet dimensional requirements, and (4) assigning stacks to floor spots using heuristically generated templates. If a feasible assignment exists at the completion of these four steps, we obtain a set of feasible load plans. Meanwhile, the Feasible Loading module returns the loading feasibility status for each local search step in the routing algorithm. In the following paragraphs, we provide more details on these steps.

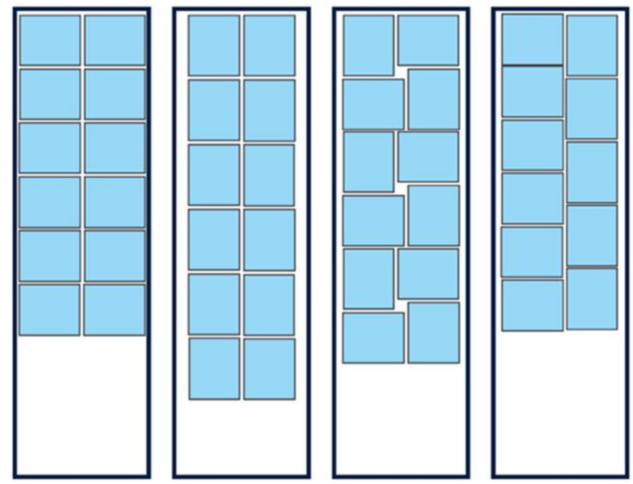
The first step of the Feasible Loading algorithm is accomplished through the Pallet Stacking Module as introduced in Figure 3. The greedy method is used to generate initial stacks. As mentioned in the earlier section, we benchmark various metaheuristics methods, such as simulated annealing and tabu search, and select the best method for each module. The simulated annealing algorithm is adopted in this module to iteratively improve the stacking solution with the objective of minimizing the number of stacks. The remaining steps start with generating a set of feasible stack-to-compartment assignments, each of which implicitly determines the position of bulkheads, compartment temperature, and the number of reloading operations. We take a three-step approach: (1) assigning temperatures to each compartment; (2) grouping stacks by store and temperature and assigning them to compartments with compatible temperatures (note: at this point, constraints such as bulkhead position and trailer dimension may be violated); and (3) using a stack-shifting algorithm to adjust the assignments and make them feasible from a loading dimension perspective (loading dimension feasibility is formulated in Appendix B). The assignment with a lower number of reloading operations will be given a higher priority for the remaining load-plan steps.

For each feasible stack-to-compartment assignment, we identify floor spot positions in each compartment with a loading pattern matching algorithm. These floor spots serve as virtual placeholders for stacks with fixed positions and orientations. The algorithm utilizes a set of carefully designed loading patterns (see Figure 6) to reduce the number of possible floor spot configurations. This improves the efficiency of the algorithm and enhances its usability by DC associates.

Once floor spots are determined, stacks are assigned to them to ensure compliance with constraints on the load axle weight limit. Assignments are based on templates generated by a heuristic method and are then filtered to produce feasible load solutions following a validation against the axle weight limit constraint using a mathematical formula. Appendix B shows the formulation of loading constraints.

In the HOS module (Huang et al. 2020a, 2021), our algorithm starts by defining a set of hierarchical driver states and abstracting HOS regulations, enabling adaptability to ever-evolving rules. It then determines initial time stamps for events such as driving, service, breaks, wait time, and layover time, and follows with an iterative step that adjusts both time and events to drive toward feasibility. Finally, it applies a wait-time reduction algorithm to reduce the total duration while ensuring compliance with all rules.

Figure 6. (Color online) Four Mainstream Loading Patterns



Notes. In the lengthwise pattern, all stacks are loaded into a trailer from their longer side (i.e., 48 inches). In the widthwise pattern, all stacks are loaded into the trailer from their shorter side (i.e., 40 inches). In the pinwheeling pattern, four stacks create a block by alternating the orientation of stacks in each row and column. That is, the first stack in a row has a different orientation than the next stack in the same row. Additionally, the orientation of stacks differs from that of the next stack in the same column. A hybrid pattern is a special loading pattern in which all stacks in a column have the same orientation but differ from the other columns. That is, if all the stacks in column 1 are lengthwise, the orientation of all stacks in column 2 is widthwise.

Planning an Efficient Load with and Without Loaded Pallets. In this subsection, we describe the methodologies we followed in three loading modules: Optimal Loading (Huang et al. 2020b), Dynamic Loading (Liu et al. 2021), and Fluid Loading (Sun et al. 2021).

The optimal loading solution builds on the same process as the feasible loading construction algorithm; however, it includes an additional step to improve the loading using a simulated annealing framework. It is designed to improve weight balance through local search steps such as swapping stacks. At each iteration, stacks are selected based on a heuristic method, which ensures that all loading constraints, except the axle weight limit, are satisfied after the swap.

The Dynamic Loading module creates an incremental load plan to optimize the load weight balance by determining the placement of remaining stacks, while keeping the already loaded portion intact and adhering to all loading constraints. It adds complexity because the loading process does not start with the assumption of an empty trailer, and stacks to be loaded must accommodate the existing load, which can be unpredictable. We designed a gap-filling algorithm, like the game Tetris. In this algorithm, we designed a set of loading units, and their utilization is intelligently determined for the loading situation based on the trailer configuration, partial loading gap, and the number of stacks left to be loaded. Figure 7 shows an example illustrating that different loading units are used to fill the gap and the choice of the loading unit is based on the number of remaining stacks.

Traditionally, loaders must follow a loading plan from the front to the tail of a truck to place each stack in its position. This loading process is rigid and often causes congestion at the loading dock if stacks placed in the front position are not ready or picked from the chamber. On the contrary, the fluid loading process identifies loading blocks in which the position of each stack is interchangeable. Figure 8 illustrates an example of a fluid loading result denoted by blocks highlighted using dotted lines. An iterative heuristic approach is proposed with the objective of maximizing the number of stacks within each loading block (Sun et al. 2021). The initial solution includes loading blocks based on compartment and delivery store stop combinations. Next, in each iteration, a new solution with new loading blocks is generated based on heuristics by performing three major operations: splitting a block, moving stacks to an adjacent block, or combining two adjacent blocks. The feasibility of the new loading blocks is checked in each iteration, and finally the best loading blocks are identified.

In summary, our solution framework allows us to leverage unique characteristics of problem structures to achieve fast computation. We use the Feasible Loading module as an example. The number of permutations of loading stacks onto a trailer increases exponentially by the maximum number of stacks a trailer can carry. A predesigned loading pattern (refer to Figure 6 for visualizations) is leveraged to narrow down the number of potential floor spot configurations for exploration. This allows us to efficiently identify feasible floor spot positions without exploring all possibilities. The fast computational time of each module improves the routing solution quality and also enables real-time interactions between a human reviewer and the system to build trust in the Load Planner, which plays a critical role throughout the development journey.

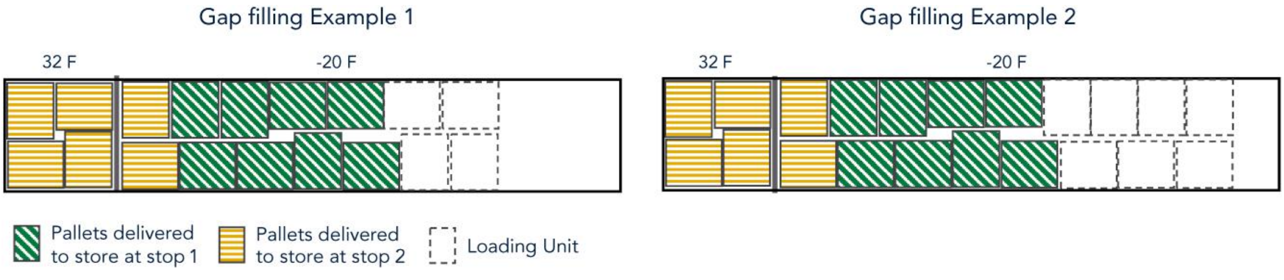
Simulation Platform: Integration of Strategy and Execution

As we illustrate in Figure 1, integrating these fast and powerful optimization models and algorithms for strategic and execution decisions through what-if scenarios benefits our supply chain planning process from an end-to-end perspective.

Network design and planning models frequently involve the analysis of multiple initiatives, including store and DC alignment, trailer configurations, and delivery frequencies. For example, we need to evaluate the scenario in which the temperature-controlled trailer that carries the commodities requires the same temperature. Alternatively, a thorough examination is essential to assess the feasibility of using temperature-controlled trailers for a wider variety of products while restricting deliveries to a single store per route. Additionally, it is crucial to evaluate the potential impact of adjusting the delivery frequency to a store, transitioning from a seven-days-a-week schedule to a five-days-a-week arrangement, or even contemplating a complete overhaul of the entire delivery network structure.

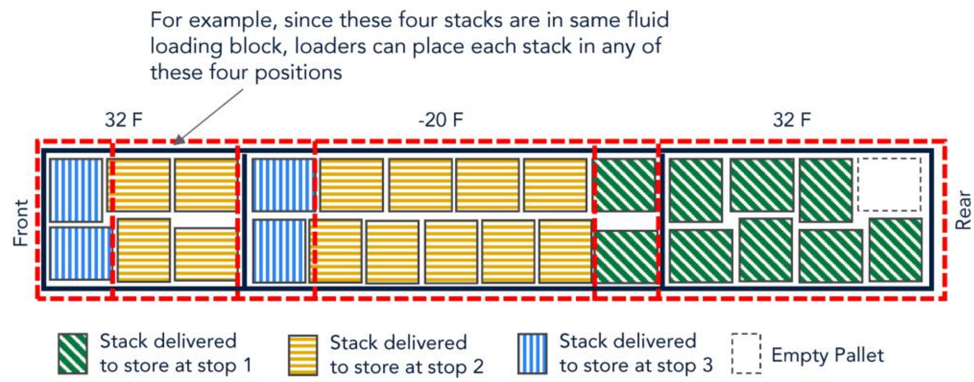
Therefore, we built a simulation platform with Load Planner's optimizers at its core to evaluate the impact of what-if scenarios. The fast computational time of the optimizer allows us to run multiple scenarios in parallel, enabling decision makers to choose a lower-cost network plan than they used previously. Additionally, the simulation platform allows users to configure their what-if scenarios and quickly review the results.

Figure 7. (Color online) Gap-Filling Methods, Which Depend on the Number of Remaining Stacks



Notes. In this diagram, we demonstrate the use of different loading units to optimize trailer utilization based on the number of remaining stacks. Example 1 showcases how the algorithm fills the gap when four stacks are to be loaded; example 2 illustrates the loading result with seven remaining stacks.

Figure 8. (Color online) Fluid Loading Result



Note. The loading blocks are highlighted using dashed lines.

Moreover, leveraging the outcomes of the simulation allows us to validate the assumptions that were previously formulated for strategic analysis of network design and transformation planning. Subsequently, the analysis can be rerun with revised inputs, ensuring an up-to-date and comprehensive evaluation. Establishing this feedback loop enables Walmart to make holistic end-to-end decisions.

This faster simulation iteration enables the strategy and planning teams to estimate realistic end-to-end impacts and therefore improve the speed and accuracy of strategic decisions. Meanwhile, the capacity to design and plan a better network creates greater opportunity for cost-effective executions. By integrating our strategic and execution decision engines into a unified, cohesive decision-making process, we have discovered remarkable synergies that surpass the cumulative potential of their individual components.

Implementation Journey

The journey to bring next-generation optimization capabilities into each supply chain decision step and shape the future of integrated decision making has been challenging. The road to implementation involved meticulous planning, experimentation with algorithms, collaboration among various groups, and unwavering dedication to align all components with our vision.

At a strategic level, historically, decisions such as network design and planning were supported by isolated optimization models, leading to separate designs for e-commerce and store networks. In 2020, the first modeling efforts were made to consolidate these networks under a unified strategy to enable end-to-end omnichannel fulfillment. A year later, the optimization models were developed and used to support dry and perishable grocery network design. As of fiscal year 2023 (FY23) (February 2022 to January 2023), Walmart fully adopted analytical and optimization models for omnichannel network design and planning.

At the execution level, it took approximately four years to turn a proof-of-concept, which combines a solver for a VRP with time windows and a container-loading solver, into a fully functional product, Load Planner, which is now used by Walmart's entire U.S. grocery network. We invested major efforts in building model capabilities to capture rich features and functionalities, enhancing solution algorithms to achieve cost advantages over existing solutions, and testing the applications in the field. Throughout this journey, we created a solution framework that keeps the optimization engines running fast with the flexibility to add incremental functionalities. We also developed innovative solutions to efficiently tackle both the unique problems that Walmart faced and the common problems within the grocery industry. After extensive testing and a successful pilot at one of our grocery DCs in 2020, Load Planner was gradually rolled out across our entire grocery network starting in April 2021. By July 2022, all 47 grocery DCs in the United States and the entire grocery outbound delivery network were equipped with this next-generation system providing faster, better, and more transparent solutions.

Moreover, we have not only advanced our optimization engines to support both strategic and operational decisions but also integrated them through what-if simulation runs. The simulation platform was developed during the time when Load Planner was being rolled out. Network design and planning models now use it to evaluate various options, including store-to-DC alignment, trailer configurations, and delivery frequencies.

Benefits

Impact on Costs, Sustainability, and Investment

As of FY23, Walmart fully adopted the proposed optimization solutions to make network design and transformation planning decisions. The direct effects of the network strategy will not materialize in the short term given our long-range planning horizon; therefore, we measure its benefits primarily based on business adoption. In FY23, the network strategy and transformation roadmap recommended based on the optimization solutions were reviewed and approved for implementation by senior management. One program received approval and funding for the construction of three new GDCs with multibillion-dollar capital investments.

In FY23, with the full network rollout of Load Planner, Walmart prevented 98.6 million pounds of CO_2 emissions and saved \$91.5 million by eliminating 108,000 truck routes covering 33 million miles. We calculated these savings monthly considering cases, cubic feet, and weight of products to normalize to a consistent calculation for determining miles and trailers saved. The combination of fewer miles driven using fewer trailers multiplied by cost per mile and cost per trailer amounted to the annualized cost savings. We calculated CO_2 emission prevention using average miles per gallon (MPG), as reported by Walmart, evaluating fuel consumed divided by odometer miles driven by all tractors in the Walmart fleet. Total miles saved divided by MPG determines gallons of fuel avoided. Using the International Carbon Bank & Exchange conversion rate of 10,180 grams CO_2 per gallon of diesel fuel, we derived the CO_2 emissions that Walmart has prevented.

Impact on Our Associates and Customers

Our network design and transformation planning solution helps Walmart make long-term investment decisions in our next-generation DCs either through retrofitting a current facility or through building a new facility. Newer automated buildings eliminate labor-intensive tasks such as lifting and carrying heavy cases to build pallets; they also provide better data accuracy and reduce handling errors, which improve store associates' visibility to incoming delivery pallets. As a result of this improved visibility, store associates can better plan their stocking-to-shelf tasks, thus improving item availability to customers.

The Load Planner solution offers a variety of features that enhance the experience of Walmart associates and customers. For example, the fluid loading feature provides DC associates with the flexibility to load a set of pallets in any sequence preferred. It improves loading experiences for associates because they do not have to follow a strict loading sequence. Moreover, it improves associate productivity in situations in which a pallet is not ready for loading but another pallet behind it is on the outbound dock and ready for loading.

Trailer loading algorithms are also helpful. Delivering perishable commodities to stores using multitemperature trailers often requires accessing multiple compartments. This often leads to the need to reload pallets for the next stops. Our trailer loading algorithms are designed to minimize the need for reloading, thus reducing the number of pallets store associates need to reload and improving their productivity.

In addition, the implementation of Load Planner has resulted in a consistent improvement in on-time store deliveries across the entire network. This translates to additional thousands of truckloads arriving at stores on schedule during holiday seasons. Improved on-time deliveries reduce the need for store associates to work overtime and improve item in-stock rates.

The simulation platform that connects strategy and execution enables rapid network changes in the event of a pandemic or a natural disaster. On September 24, 2022, a hurricane struck Florida and cut off the roads to many of our stores. Walmart had to close 252 facilities including DCs and stores, resulting in the disruption of food and water replenishment. The network strategy team quickly collaborated with operations to develop a new alignment of DCs and stores in the impacted areas. The simulation platform then was used to evaluate the feasibility of the plan and select the best plan using Load Planner to reroute deliveries of essential supplies from other hubs to stores in the impacted area. The process took hours rather than days, and helped Walmart quickly reopen the closed facilities.

Transferability

Although the models and algorithms presented in this paper were developed for Walmart's grocery outbound distribution network, they could be adopted by other sectors within or outside of Walmart. The network design and transformation planning optimization models are adaptable to make strategic recommendations for other retailers and supply chain operators. The integrated routing and loading solution offered by Load Planner has proven to be flexible during the GDC rollout. With a few additions to its existing functionalities, it can be extended to other Walmart networks, such as the reverse logistics network (e.g., from stores to DCs).

Dynamic multiple-stop routing of variable pallet sizes with multiple variable-sized temperature chambers is a generic problem encountered in various industries. The solution algorithms we developed in Load Planner can be used separately to tackle different problems or to support grocery first-, middle-, and last-mile operations.

The DOT HOS module we built into a VRP solver is generally applicable to any long-distance commercial truck routing problem. It can provide timing estimates for tasks such as driving, waiting, breaks, and layovers, or perform HOS and time window compliance checks of a single-stop or multiple-stop trip. These two functionalities make the resulting routes more trustworthy and executable.

For businesses that use single- or multiple-temperature trailers (with fixed or variable compartment sizes) to transport palletized commodities, our truck loading algorithms can provide well-developed solutions to various use case scenarios, including batch planning, incremental planning with partially loaded trailers, and other innovative planning solutions that include flexibilities for loaders to change loading sequences (i.e., the fluid loading feature).

For retailers that do not receive full-pallet quantities directly from suppliers, our algorithms can provide an opportunity to increase trailer utilization through improved optimization logic (i.e., volume of freight per delivery) while minimizing the cost of mileage, fuel, driver utilization, and wear and tear on equipment. Although case-loaded (i.e., floor-loaded) trailers may yield higher utilizations, the efficiency of palletized deliveries reduces loading and unloading times; thus, drivers spend more of their time driving and less of their time waiting.

Conclusion

We have described and displayed the work of hundreds of Walmart associates since 2019 and have shown how the collaboration among various groups was instrumental in helping these groups develop scientific approaches to support key decisions across multiple tiers of supply chain management. The groups faced many difficulties:

- The initiatives were carried out to shape one of the largest retail distribution operations in the world during a time of omnichannel transformation and extreme challenges faced by the global supply chain.
- The network was already highly efficient, with truck loads for dry commodities close to fully utilized; therefore, any incremental improvement became challenging.
- We needed to coordinate the efforts of an extremely diverse group of associates, including scientists, engineers, product managers, strategy officers, DC/store associates, and drivers who work on the ground. Building trust among all the stakeholders was fundamental, because the outputs of one group were necessary to the next group in the line of work.

We are proud of the breakthroughs we achieved throughout the development and implementation of our solutions in three major areas.

- We filled the gaps between theoretical studies and real-world applications by simultaneously solving routing and loading problems using all realistic complex operational constraints.
- We developed an exact method to solve a variable-size multiple-compartment container loading problem with complex real-life constraints within a millisecond.
- We were able to accurately run hundreds of scenarios with varying inputs in reasonable time in response to stakeholders' and management queries, thus resulting in a high adoption success rate.

Importantly, the algorithms/models we developed are used not only in Walmart's grocery outbound operations. We are exploring the use of Load Planner's routing and loading capabilities for other Walmart networks, including the general merchandise outbound network and the reverse logistics network. Moreover, these capabilities may be externalized to other organizations who operate similar DC-to-store distribution operations. The HOS module we built into a VRP solver is generally applicable to any long-distance commercial truck routing problem.

Finally, and most importantly, the solution saved millions of dollars, avoided millions of pounds of CO_2 emissions, improved our associates' experiences, and served our customers when and where they needed us most.

Appendix A. Transformation Roadmap: Scheduling Problem

In this appendix, we describe the mathematical formulation of the transformation scheduling problem, which describes the dependency between the GDCs at different periods. For illustration purposes, we provide two groups of formulations, which could cover all types of transformations with minor input adjustments.

- Group 1: transformations that can be completed over a single period.
- Group 2: transformations that require multiple periods for completion.

Sets and Indices

$T0$:= set of all periods (including the initial), $t \in T0 = \{0, 1, 2, \ldots, \tau_T\}$;

T := set of periods considered, $t \in T = \{1, 2, \ldots, \tau\}$;

TL := set of processes of the group 2 GDC transformation, $tl \in TL = \{1, 2, \ldots, \tau_{TL}\}$;

L_1 := set of group 1 GDC location identifiers (IDs);

L_2 := set of group 2 GDC location IDs;

Parameters

VendorCap_t := vendor capacity of the maximum number of group 2 constructions to start at time $t \in T$.

Binary Decision Variables

$y^1_{l,t}$:= binary indicator of whether the group 1 site at location l at time t has completed its transformation $l \in L$, $tl \in TL$, $t \in T0$;

$y^2_{l,tl,t}$:= binary indicator of whether the group 2 site at location l at time t is in the transformation process tl, $l \in L$, $tl \in TL$, $t \in T0$.

Constraints

1. Dependency between the periods (i.e., condition/construction continuity)

a. Group 1, $\forall l \in L_1$

$$y^1_{l,t-1} \geq y^1_{l,t}, \; t \in T$$

b. Group 2, $\forall l \in L_2$

$$\sum_{tl \in TL} y^2_{l,tl,t} \leq 1, \quad \forall t \in T0$$

$$\sum_{t\in T} y^2_{l,1,t} = 1 \text{ (remove if the transformation is in progress at the beginning)},$$

$$y^2_{l,tl,t} \le y^2_{l,tl+1,t+1}, \quad \forall tl \in TL\backslash\{\tau_{TL}\},\ t\in T0\backslash\{\tau_T\}$$
$$y^2_{l,\tau_{TL},t} \le y^2_{l,\tau_{TL},t+1}, \quad \forall t \in T0\backslash\{\tau_T\}$$

2. Vendor capacity (for group 2 only)

$$\sum_{\forall l\in L_2} y^2_{l,0,t} \le \text{VendorCap}_t, \quad \forall t\in T$$

3. Other constraints

a. Given schedule of certain sites by fixing the values of the binary decision variables;

b. Flow balance constraints on GDC-to-store alignment.

Appendix B. Perishable Load Plan Feasibility

In this appendix, we describe the formulation of feasibility conditions for perishable-commodity loading. This formulation assumes that stacks have been assigned to each compartment for a route with o stops with n_{kt} ($k\in\{1, 2, 3\}$, $t\in\{1,\dots,o\}$) stacks from the tth stop in the kth compartment. We define the following notation:

$m_k = \sum_{t=1}^{o} n_{kt}$:= number of stacks in kth compartment, $k\in\{1, 2, 3\}$;

$m = \sum_{k=1}^{3} m_k$:= total number of stacks in the trailer;

w_{ik} := weight of stack i in kth compartment, $i\in\{1,\dots,m_k\}$, $k\in\{1, 2, 3\}$;

a_{jk} := distance from jth floor spot's center point to the innermost position of the kth compartment, calculation varies for loading patterns, $j\in\{1,\dots,m_k\}$, $k\in\{1, 2, 3\}$; we show the formula for widthwise, lengthwise, and pinwheeling (defined in Figure 6 in the text) below:

$$a_{jk} = \begin{cases} \dfrac{v}{2} + \left\lfloor \dfrac{j-1}{2} \right\rfloor v & \text{Widthwise} \\ \dfrac{u}{2} + \left\lfloor \dfrac{j-1}{2} \right\rfloor u & \text{Lengthwise} \\ \begin{cases} \dfrac{u}{2} + \left\lfloor \dfrac{j-1}{2} \right\rfloor \dfrac{u+v}{2}, & j\%2 = 1 \\ \dfrac{v}{2} + \left\lfloor \dfrac{j-1}{2} \right\rfloor \dfrac{u+v}{2}, & j\%2 = 0 \end{cases} & \text{Pinwheeling} \end{cases}$$

$w^{(k)} = \sum_{i=1}^{m_k} w_{ik}$:= total weight of stacks in kth compartment, $k\in\{1, 2, 3\}$;

b^k_{lb}, b^k_{ub} := lower/upper bounds of bulkheads, $k\in\{1, 2\}$;

ℓ^k_S := length of stacks in kth compartment, $k\in\{1, 2, 3\}$;

$g = \sum_{k=1}^{3} w^{(k)}$:= total weight of stacks on a given route;

δ_1, δ_2 := locations of front and rear axles;

g_1, g_2 : maximum weight limits of front and rear axles;

$\mathcal{C}_k(\cdot)$: set of feasible floor spots for a given stack;

$x_{ijk}\in\{0, 1\}$: binary variable denoting whether jth stack is assigned to ith floor spot in kth compartment;

The perishable loading feasibility can be expressed in the constraints we show below. The first constraint is for validation of axle weight limit; the second and third constraints are assignment constraints to enforce one-to-one matching between stacks and floor spots. The final three are constraints on dimensions and bulkhead positions.

$$\delta_2 - \frac{g_1(\delta_2-\delta_1)}{g} - \ell_1 \le \sum_{k=1}^{3}\sum_{(i,j)\in\mathcal{S}} \frac{w_{ik}a_{jk}}{g} x_{ijk} \le \delta_1 + \frac{g_2(\delta_2-\delta_1)}{g} - \ell_1$$

$$\sum_{i\in\mathcal{C}_k(j)} x_{ijk} = 1, k\in\{1,2,3\}, j\in\{1,\dots,m_k\}$$

$$\sum_{j\in\mathcal{C}_k(i)} x_{ijk} = 1, k\in\{1,2,3\};\ i\in\{1,\dots,m_k\}$$

$$\ell_1 = \left(w^2 \max\{b^1_{lb}, \ell^1_s\} + w^3 \max\{b^1_{lb} + \ell^2_s, \ell^1_s + \ell^2_s, b^2_{lb}\}\right)/g$$

$$\ell^1_S \le b^1_{ub};\ \ell^3_S \le \ell_T - b^2_{lb};\ \ell^2_S \le \min\{b^2_{ub}, \ell_T - \ell^3_S\} - \max\{\ell^1_S, b^1_{ub}\}$$

$$w_T \ge 2u,\ u+v,\ 2v$$

References

Braekers K, Ramaekers K, Van Nieuwenhuyse I (2016) The vehicle routing problem: State of the art classification and review. *Comput. Indust. Engrg.* 99(September):300–313.

Clarke G, Wright JW (1964) Scheduling of vehicles from a central depot to a number of delivery points. *Oper. Res.* 12(4):568–581.

Côté JF, Guastaroba G, Speranza MG (2017) The value of integrating loading and routing. *Eur. J. Oper. Res.* 257(1):89–105.

Dantzig GB, Ramser JH (1959) The truck dispatching problem. *Management Sci.* 6(1):80–91.

Eksioglu B, Vural AV, Reisman A (2009) The vehicle routing problem: A taxonomic review. *Comput. Indust. Engrg.* 57(4):1472–1483.

Fu M, Nayak A, Vasantham M (2020) Automatic generation of load and route design. U.S. Patent No. 20200242543, filed January 30, 2020, issued July 30, 2020. United States Patent and Trademark Office, Alexandria, VA.

Huang J, Liu M, Fu M (2020a) Validation of routes in automatic route design. U.S. Patent No. 20200242554, filed January 30, 2020, issued July 30, 2020. United States Patent and Trademark Office, Alexandria, VA.

Huang J, Liu M, Fu M, Nayak A (2020b) Automatic generation of load design. U.S. Patent No. 20200242285, filed January 30, 2020, issued July 30, 2020. United States Patent and Trademark Office, Alexandria, VA.

Huang J, Sun O, Liu M, Fu M (2021) Flexible dock-out time. U.S. Patent No. 20210150475, filed January 31, 2021, issued May 20, 2021. United States Patent and Trademark Office, Alexandria, VA.

Iori M, Martello S (2010) Routing problems with loading constraints. *TOP* 18(1):4–27.

Laporte G (2009) Fifty years of vehicle routing. *Transportation Sci.* 43(4):408–416.

Liu M, Huang J, Fu M (2021) Automatic generation of incremental load design. U.S. Patent No. 20210150101, filed January 28, 2021, issued May 20, 2021. United States Patent and Trademark Office, Alexandria, VA.

Liu M, Huang J, Fu M, Nayak A (2020) Automatic generation of route design. U.S. Patent No. 20200242555, filed January 30, 2020, issued July 30, 2020. United States Patent and Trademark Office, Alexandria, VA.

Martello S, Vigo D (1998) Exact solution of the two-dimensional finite bin packing problem. *Management Sci.* 44(3):388–399.

Martello S, Pisinger D, Vigo D (2000) The three-dimensional bin packing problem. *Oper. Res.* 48(2):256–267.

Pollaris H, Braekers K, Caris A, Janssens GK, Limbourg S (2015) Vehicle routing problems with loading constraints: State-of-the-art and future directions. *OR Spectrum* 37(2):297–330.

Sun O, Liu M, Huang J, Fu M (2021) Automatic generation of flexible load design. U.S. Patent No. 20210150102, filed January 28, 2021, issued May 20, 2021. United States Patent and Trademark Office, Alexandria, VA.

Prakhar Mehrotra is a vice president, applied artificial intelligence (AI), at Walmart Global Tech. He leads the applied AI organization, a full stack AI group of 370+ data scientists, machine learning engineers, and data engineers, responsible for building algorithms across all areas of Walmart merchandising, supply chain, and operations for both stores and e-commerce. During his 10+ years of experience in the broader field of AI with specialization in machine learning, computer vision, and operations research, he successfully built AI teams and led AI strategy for Walmart US and Uber. He has an advanced engineer's degree in aeronautics from the California Institute of Technology and a double master's in aerospace and applied math from Ecole Polytechnique and Caltech. He did undergraduate work in mechanical engineering at the National Institute of Technology, India.

Mingang Fu is a senior director of applied AI, US Omni-Tech, at Walmart Global Tech. He leads a global team of 70+ operations research and machine learning professionals at Walmart. He and his team have been responsible for prototyping, implementing, and improving optimization systems in the areas of supply chain and store operations. He was also the main contributor to the Load Planner's prototyping model. In his 12 years of experience, he has worked in areas of supply chain, retail, e-commerce, transportation, and last mile. Prior to Walmart, he had scientific roles at Amazon. He holds a PhD in industrial and system engineering from the University of Washington in Seattle and a master's degree in engineering management from Duke University.

Jing Huang is a director of applied AI at Walmart Global Tech. She leads a full-stack AI team to define, develop, and refine core optimization engines for first-mile, middle-mile, last-mile delivery, and e-commerce fulfillment production systems. She was also the main contributor to the Load Planner's routing modules and its overall algorithmic architecture. She is deeply passionate about optimizing decision-making efficiencies via data, algorithms, and computing and has years of experience in launching end-to-end AI products across on-demand delivery, e-commerce supply chain, logistics, and transportation sectors. She holds a PhD in industrial engineering, specializing in operations research, from Ohio University.

Sai Rajesh Mahabhashyam is a senior director of data science in the supply chain strategy team at Walmart, leading long-term strategic transformative initiatives across Walmart's supply chain. His responsibilities include ideating and developing multi-year asset strategy and transformation models that are foundational to Walmart's end-to-end supply chain strategy. He has 19 years of experience in developing and embedding various optimization and machine learning modules within the supply chain and logistics software products. He holds master's and doctoral degrees in industrial engineering and operations research from Penn State University, an MBA from the Kelley School of Business (Indiana University), data science training from Harvard, and a bachelor's degree from the Indian Institute of Technology, Madras, India.

Minghui Liu is a principal data scientist, applied AI, at Walmart Global Tech. He has been leading scientific work in various supply chain domains, such as next generation e-commerce fulfillment platform, last-mile delivery, supply demand matching for on-demand delivery platform, truck loading optimization, and demand forecasting. He was also the main contributor to the Load Planner's loading modules. Before joining Walmart, he spent three years at the SAS Institute focusing on the design and development of revenue management and pricing optimization systems. He holds a PhD in statistics and operations research from the University of North Carolina at Chapel Hill and a bachelor's degree in mathematics from Peking University.

Ming (Arthur) Yang is a senior data science manager in the supply chain strategy team at Walmart. He serves as the lead data scientist of middle-mile perishable network strategy, concentrating on the development of optimization tools for designing perishable supply chain networks, guided by a detailed step-by-step transformation roadmap. He holds a PhD in industrial engineering from The Ohio State University, an MS in financial operational research from the University of Edinburgh, and a BS in applied mathematics from Xiamen University.

Xiaojie Wang is a staff data scientist, applied AI, at Walmart Global Tech. She led major feature developments for the Load Planner during the Load Planner network rollout. She continues to lead the design and implementation of optimization engines in transportation domain. Prior to joining Walmart, she worked at NielsenIQ as a senior data scientist, where she focused on pricing strategies. She holds a PhD in operations research from the University of Florida, a master's degree from the Georgia Institute of Technology, and a bachelor's degree in mathematics from Tsinghua University.