The Frozen Prefix: Commitment Horizons, Order Maturation, and Why Cost per Stop Is the Only Last-Mile Metric

Architecture, Computational Methods, and a Modelled Scenario for MileSoft Fleet Optimization

MileSoft Engineering Research Group

MileSoft Software Technologies

https://milesoft.net/

Corresponding authors: *Deepak Daryani* (deepakdaryani@milesoft.net), *Tanuj Daryani* (tanujdaryani@milesoft.net)

Working paper — Version 1.0 — 24 August 2026

Permanent link: https://milesoft.net/research/products/fleet-optimization

Abstract

Planning every route at six in the morning was defensible when orders matured overnight. The product's material states the change bluntly: today an order confirmed at nine changes the economics the six o'clock plan assumed, and the route that was optimal at dawn is not optimal by mid-morning.

This paper presents the architecture and computational methods of MileSoft Fleet Optimization. Its central argument is that continuous re-optimisation is not the same computation run more often. A vehicle already on the road has stops that are executed and stops that have been promised, and neither can be moved — so the optimiser holds a shrinking tail of free decisions behind a frozen prefix, and the length of that prefix is the whole design.

Three results follow. The commitment horizon is a policy choice with a stated cost on both sides: a long horizon strands every order maturing inside it, and a short one re-sequences work a driver has already been told about. Consolidation value, in the form of the Clarke-Wright savings between two stops, is what makes waiting worth anything at all — so the horizon should be set against how much saving a late order can still capture rather than against a clock. And cost per stop, not mileage, is the metric that decides which customers a last mile can afford, because the two can move in opposite directions.

This module is named in a published third-party logistics deployment but no result is attributed to it there, and none is claimed here. Section 5 is a modelled scenario with every assumption printed. The product page carries no screenshots, so every figure in this paper is an authored schematic and is captioned as one.

Keywords: Fleet optimization software; Vehicle routing problem; Dynamic route re-optimisation; Last-mile delivery cost; Cost per stop; Time window routing; Multi-depot routing; Driver hours compliance; Cold chain routing constraint; Clarke-Wright savings; Rolling horizon planning; Delivery route planning

1 Introduction

The 2010 model of last-mile planning was coherent: plan every route at six in the morning, execute it as planned all day. It assumed orders matured slowly and that a plan made at dawn stayed valid until dusk.

The product's material states what broke it. Orders now mature continuously, customer windows shift, and the route that made sense at six is suboptimal by nine. That is a correct diagnosis and this paper takes it as the premise.

What the paper adds is the observation that fixing it is harder than running the planner more often, because by nine in the morning most of the day's plan is no longer available to change.

1.1 Why re-planning is a different problem

At six the planner faces an empty board: every stop is a free decision. At nine, three stops per vehicle have been served, two more have been promised to customers with a stated window, and the driver has been shown a sequence.

Nothing about the served stops can be revisited. The promised stops can be moved only by breaking a commitment somebody is relying on. What remains is a tail — and the tail shrinks as the day proceeds, so the capacity to absorb a new order falls exactly as the rate of new orders rises.

> This is the structural asymmetry that makes dynamic routing hard. The information arrives late and the freedom to act on it disappears early, and the two curves cross somewhere in the middle of the morning.

1.2 Contributions

1. Dynamic re-optimisation formalised as a rolling horizon with a frozen prefix (Section 4.2).
2. The commitment horizon priced on both sides — stranded orders against communicated churn (Section 4.2).
3. Consolidation savings as the quantity that makes waiting valuable, tying the horizon to something measurable (Section 4.3).
4. Cost per stop decomposed, and the conditions under which it moves opposite to mileage (Section 4.4).
5. An eight-dimension capability reference framework for fleet routing systems (Section 6.3).

2 Background and Related Work

Vehicle routing is one of the most studied problems in operations research and one of the least faithfully implemented, because the version that has been solved is static and the version a distribution centre runs is not.

2.1 The problem as it was posed

Dantzig and Ramser posed the truck dispatching problem in 1959: assign deliveries to vehicles and sequence each vehicle's stops so total distance is minimised, subject to vehicle capacity. Every routing system in use descends from that statement.

Clarke and Wright gave the savings heuristic in 1964, and it remains the clearest statement of why routing produces value at all. Serving two stops on one out-and-back trip each costs the sum of two round trips; serving them on one route saves the difference, and the saving is larger the closer the two stops are to each other relative to the depot.

That quantity — the saving from combining two stops — is the one Section 4.3 uses to price the commitment horizon, because it is exactly what a late-arriving order can still capture if there is room left to absorb it.

2.2 The constraints that are not preferences

A real route carries constraints the classical formulation omits, and the product's material names five: vehicle capacity, driver hours, customer time windows, cold chain and hazardous goods.

Driver hours are the clearest example of why these belong inside the model. Regulation (EC) No 561/2006 caps daily driving at nine hours, extendable to ten twice a week, and requires a forty-five minute break after four and a half hours of driving. A route that exceeds those limits is not a slightly worse route; it is one the driver may not legally complete.

Cold chain behaves the same way. A temperature excursion does not degrade the delivery — it destroys the consignment — so a sequence that leaves chilled goods on a vehicle past their tolerance is infeasible rather than expensive.

> The product's own framing is that these are built into route constraints rather than enforced manually after release. That is the same control-against-warning distinction the dock and pokayoke papers draw, and it matters for the same reason: a check applied after the plan is published is a check somebody has to remember to apply.

2.3 Where the literature stops and practice begins

The static problem assumes all demand is known when planning begins. Dynamic variants relax this, and the standard device is a rolling horizon: re-solve periodically over what remains.

What a rolling horizon does not by itself supply is the rule for what may be re-solved. In a warehouse that distinction is academic; on a road it is the entire practical question, because a stop already communicated to a customer is different in kind from one that has not been.

Section 4.2 makes that rule explicit, and Section 6.1 argues that it is a policy the operator should own rather than a parameter buried in an optimiser.

3 System Overview

The system holds an order pool that fills through the day, a fleet with vehicle and driver attributes, a constraint set, an optimiser that runs on a rolling horizon, a driver application that turns a plan into a sequence somebody can follow, and a cost model that attributes what the day actually cost.

> Every figure in this paper is an authored schematic. This product's page carries no screenshots of any kind, and no image from another product is used to stand in for one.

3.1 The order pool and its maturation

Orders do not arrive as a batch. They accumulate across the planning day, and each carries a delivery window, a service duration, a location and any product constraints — temperature, fragility, hazard class.

The distribution of arrival times relative to the planning cadence is the single most important input to the design, and it is a property of the business rather than of the software. A wholesale operation whose orders are all in by five in the morning genuinely can plan once; an e-commerce operation with a same-day cut-off at noon cannot.

This is worth stating because it bounds the value of the whole product category at a given site. Continuous re-planning is valuable in proportion to how much demand matures after the first plan is made, and a site can measure that from its own order timestamps before evaluating anything.

3.2 The optimiser and what it is allowed to move

The optimiser runs repeatedly against the current state rather than once against a forecast. Each run takes the executed history as given, the committed stops as fixed, and re-sequences what remains.

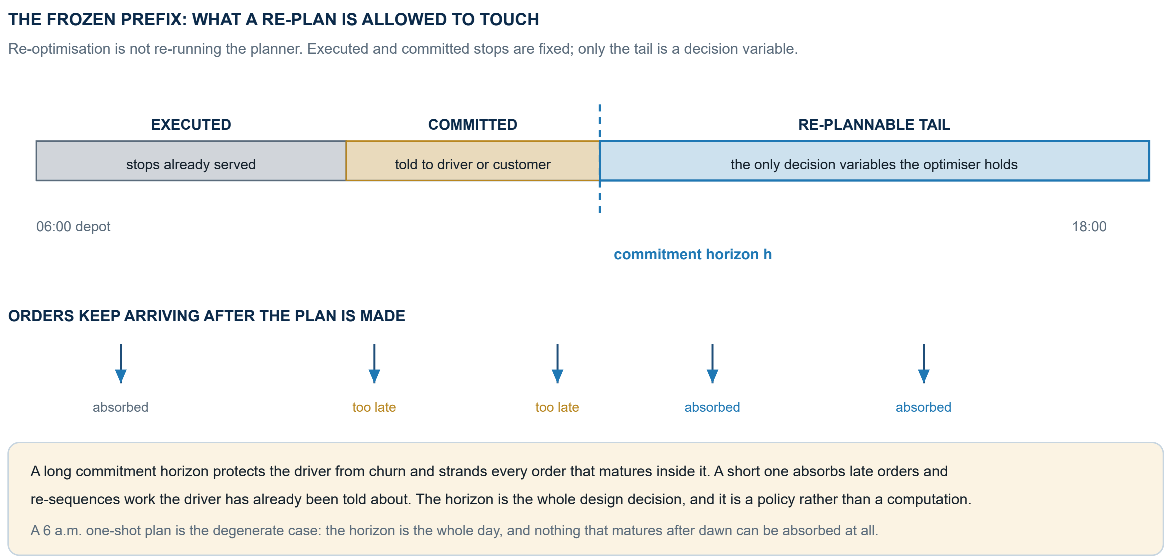

Figure 1. *Schematic.* Schematic (not product UI): the frozen prefix. Stops already served and stops already communicated are fixed; only the tail behind the commitment horizon is available to the optimiser. Orders maturing inside the horizon cannot be absorbed, which is why a six o'clock one-shot plan — a horizon spanning the whole day — absorbs nothing at all.

Multi-depot operation widens the same decision. An order may be served from whichever depot has a vehicle with the capacity, the constraint fit and the room in its remaining tail — which is not necessarily the nearest depot, and the product's material is explicit that proximity alone is not the assignment rule.

3.3 The driver application as part of the control loop

A re-route that a driver does not receive is not a re-route. The system pushes updated stop sequences to a driver application with navigation, and dispatchers see the same map and can approve or override.

This closes the loop that makes continuous re-planning possible at all, and it also creates the cost Section 4.2 prices. Every push invalidates a sequence the driver has already absorbed, and a driver receiving frequent re-sequences will eventually stop trusting the order they are shown — at which point the optimiser is producing plans that are not being followed and no longer knows it.

> Churn has a failure mode worse than annoyance. A driver who re-sequences by their own judgement breaks the assumption every subsequent re-plan is built on, and the divergence is invisible until the day's actuals are compared with the plan.

3.4 The cost model

The system attributes fuel, mileage, time and vehicle utilisation by route, by driver and by customer, and reports cost per stop as the primary figure.

Attribution by customer is the part that changes decisions. A route's cost is an operational fact; a customer's cost to serve is a commercial one, and Section 4.4 shows they can diverge sharply in ways that total mileage conceals entirely.

4 Computational Methods

Four computations carry the paper: the problem as posed, what a re-plan may touch, what waiting is worth, and how the day's cost should be reported.

4.1 The routing problem with the constraints that matter

Let the depot be node 0 and customers be nodes 1..n, with c(i,j) the cost of travelling between them and K the vehicle set.

minimise total travel cost across vehicles, subject to: load not exceeding vehicle capacity, arrival at each stop inside its window, and total driving plus service not exceeding the driver's legal hours (vrp)

x(i,j,k) is 1 where vehicle k travels from i to j, y(i,k) is 1 where vehicle k serves customer i, q(i) is the demand at i, Q(k) the vehicle capacity, [a(i), b(i)] the delivery window, tau(i) the arrival time, s(i) the service duration, and H(k) the driver's remaining legal hours under Regulation 561/2006.

Product constraints enter as additional feasibility conditions rather than as costs: a chilled consignment restricts which vehicles may carry it and how long it may remain on board; a hazardous consignment restricts which consignments may share a vehicle. Both behave like the door capabilities of the companion dock paper — they remove edges rather than penalise them.

> Every constraint above is binary. A route violating any of them is not a worse route; it is one that cannot be run, and the distinction is why they belong in the model rather than in a review step after it.

4.2 The frozen prefix, and what the horizon costs

At any re-plan the stops on a vehicle partition into three sets: executed, committed, and free.

a route is the concatenation of executed, committed and free segments; only the free segment is optimised; the committed set is everything whose planned arrival falls inside the commitment horizon h (frozen)

E is the executed prefix, C the committed segment, F the free tail and h the commitment horizon. Setting h to the length of the working day recovers the six o'clock one-shot plan as a special case, with F empty after the first run.

The horizon has a cost on each side and both are computable from a site's own history.

stranding cost rises with the horizon - more orders mature inside it and miss the plan; churn cost falls with the horizon - fewer communicated commitments are broken; the balance point is the square root of their ratio (horizon)

lambda(late) is the rate at which orders mature during the operating day, v-bar(miss) the value lost when one cannot be absorbed into an existing route, and kappa a churn coefficient capturing the cost per unit of re-sequencing communicated work.

The functional forms are stylised and the paper says so. What survives the simplification is the shape: the two costs move in opposite directions, so a horizon is a trade rather than a setting, and the correct value depends on how late a site's orders arrive and how much its drivers are disrupted by change.

One consequence is worth stating for the buyer. A site whose orders all mature before dispatch has lambda(late) near zero, so the optimal horizon is long and continuous re-planning has almost nothing to do. The value of this product category is proportional to a quantity a site can measure from its own order timestamps in an afternoon.

4.3 What a late order can still capture

Section 4.2 prices the horizon against the value lost when an order cannot be absorbed. That value has a name and a formula, and it comes from the oldest result in the field.

the saving from serving stops i and j on one route rather than two is the two out-and-back legs less the direct leg; the value lost by a stranded order is the best saving it could have achieved against the tail still available (savings)

c(0,i) is the depot-to-stop cost and sigma(i,j) the Clarke-Wright saving. The expectation is taken over the free tail F, because a saving can only be realised against a stop that is still movable.

Two properties of sigma decide the economics. It is large when two stops are close to each other and far from the depot, and it is near zero when they are close to the depot or far apart — so a late order in a dense delivery cluster is worth waiting for and a late order in an outlying area is not.

That suggests a refinement the equation makes obvious and most systems do not implement: the commitment horizon need not be uniform. A route serving a dense cluster benefits from staying open; a route to a remote customer can be committed early because nothing arriving later will consolidate with it.

> Waiting is worth exactly the consolidation it might capture. A system that holds every route open for the same duration is paying churn on routes where no saving was ever available.

4.4 Cost per stop, and when it moves against mileage

The product's material names cost per stop as the primary last-mile metric, ahead of total mileage and average dwell. The reason is a decomposition.

$$\text{route cost} = \text{fixed vehicle cost} + \text{distance rate} \times \text{kilometres} + \text{hourly rate} \times (\text{total service time} + \text{driving time}); \text{ cost per stop is that total divided by the number of stops served} \tag{cps}$$

f(k) is the fixed cost of putting vehicle k on the road, gamma the per-kilometre cost, omega the loaded hourly cost of driver and vehicle, s(i) the service time at stop i, and n the stops served.

The three terms scale differently. Fixed cost is amortised over the stops — so it falls per stop as density rises. Distance cost is dominated by the depot-to-area legs in sparse territory. Service cost scales with the number of stops, not with distance at all.

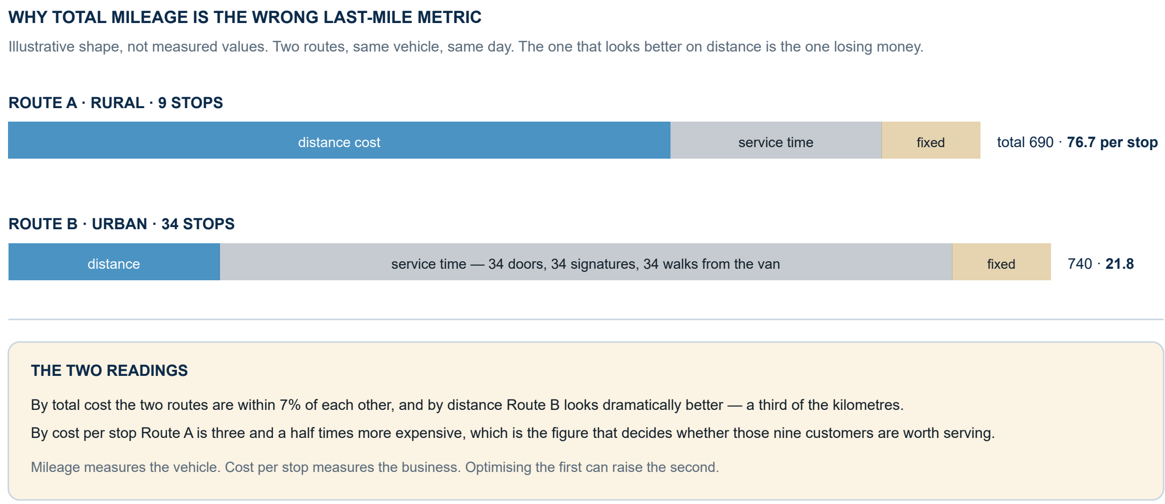

Figure 2. *Schematic.* Schematic (illustrative shape, not measured values): two routes on the same vehicle and day. The rural route travels three times the distance to serve nine stops; the urban route travels far less and serves thirty-four. Their total costs are within seven per cent — but their cost per stop differs by a factor of three and a half, and only the second figure says which customers are worth serving.

The consequence for optimisation is uncomfortable and worth being explicit about. Minimising total distance, which is what Equation (vrp) does as written, is not the same as minimising cost per stop. On dense urban work where service time dominates, a shorter route with a worse stop sequence can raise the figure that actually decides margin.

The practical resolution is to keep distance as the optimiser's objective — it is well-behaved and the constraints are already hard enough — and to use cost per stop as the reporting metric that decides which customers, areas and windows the business should accept at all. They are different decisions on different timescales.

5 Modelled Scenario

No published outcome is attributed to this module. It is named in a third-party logistics deployment, but the four results that deployment reports belong to other modules, and this paper claims none of them. Everything below is either a design statement or a modelled scenario with its assumptions printed.

5.1 The deployment this module appears in

A multi-client third-party logistics operator deployed six MileSoft warehouse modules across a distribution centre. Its stated problem for the last mile was specific: routes were planned once a day at six in the morning and never re-optimised, and fuel cost per kilogram had been rising for two consecutive quarters.

The published approach describes wave plans flowing into route plans that re-optimise as orders mature — the coupling Section 3.2 depends on.

> No result is attributed. The deployment's four published outcomes — dock idle time, truck dwell, inventory accuracy and client onboarding — all belong to other modules. The fuel-cost trend is stated as the problem, not as a measured improvement, and this paper does not convert it into one.

What the account does establish is the premise of Section 4.2 in a real setting: a site whose orders matured after its single daily plan, and whose cost was rising as a result. That is a description of lambda(late) being large and the commitment horizon spanning the whole day.

5.2 Design statements, separated from results

Table 1. Statements published about this module. All are design claims rather than measurements, and are listed here so a reader can tell which is which.

Statement	What it asserts	Kind
Continuous re-planning	orders, windows and traffic re-enter the plan as they change	Design
Constraints in the model	capacity, driver hours, windows, cold chain, hazard	Design
Driver application	re-routes pushed with sequenced stops and navigation	Design
Multi-depot assignment	by vehicle availability and load fit, not proximity alone	Design
Cost per stop primary	reported by route, driver and customer	Design

Each is checkable by demonstration rather than by evidence, which is what the framework of Section 6.3 is for. None is a claim about how much a site would save.

5.3 Modelled operating-cost recovery

The published return model prices the saving as operating cost recovered through route, load and vehicle utilisation. Its assumptions are printed here so a reader can substitute their own.

Table 2. Modelled scenario — not a deployment result. Assumptions are the published defaults of the return-on-investment model for this product.

Assumption	Value
Vehicles in the fleet	20
Operating hours per vehicle per year	2,200
Baseline fleet operating hours	44,000 per year
Utilisation and route-efficiency gain	12%
Modelled recovery	5,280 vehicle-hours per year

Twelve per cent is the parameter to interrogate, and Section 4.2 says what it depends on. A gain of that size requires a meaningful lambda(late) — orders maturing after the first plan — because a fleet whose demand is fully known at dispatch is already being planned against complete information, and re-planning has nothing to add.

> A site can size its own expectation before any procurement conversation. Plot order confirmation times against dispatch time; the fraction maturing afterwards is the ceiling on what this category can recover.

6 Discussion

6.1 The horizon belongs to the operator, not the optimiser

Section 4.2 shows the commitment horizon is a trade between stranded orders and broken commitments. It is tempting to treat that as a parameter for the optimiser to tune, and it should not be.

The two costs are not commensurable inside the software. Stranding is a routing cost the system can compute; churn is a cost borne by drivers and customers, and its true magnitude depends on how much disruption a particular workforce and customer base will tolerate before behaviour changes. An optimiser cannot see the point at which drivers stop following sequences.

The right arrangement is therefore for the system to expose the horizon as an explicit policy, report both costs against it, and let an operations manager set it — reviewing it as order patterns change rather than treating it as a configuration nobody revisits.

> A system that hides the horizon has made the trade on the operator's behalf without telling them which way. Ask any candidate what its horizon is and whether it can be changed; a vendor who does not understand the question has not modelled the problem.

6.2 Optimise distance, manage cost per stop

Section 4.4 identifies a genuine tension: the objective a routing optimiser minimises is not the metric the business should manage. It is worth resolving rather than leaving as a paradox.

Distance is the right objective for the optimiser because it is well-behaved, because the constraints are already hard, and because on any single day the stops to be served are largely given. Cost per stop is the right metric for the business because it answers a different question on a longer timescale: not how to serve today's stops, but which stops are worth having.

The decisions cost per stop actually informs are commercial. Whether a delivery area is worth serving at current density. Whether a customer's narrow window justifies its price. Whether a minimum order value should rise. None of those is a routing decision, and none is visible in total mileage.

The failure this prevents is a real and common one: a fleet reporting falling mileage year on year while its margin erodes, because it is serving more stops in denser areas and its service time — which no distance metric counts — has grown.

6.3 A capability reference framework for fleet routing systems

Table 3. Capability reference framework. Each dimension separates a routing system from a route printer, and each is answerable by demonstration against a live plan rather than by reading a specification.

Dimension	Question the system must answer by demonstration
D1 Hard constraints	Build a route exceeding a driver's legal hours. Is it produced, warned, or refused?
D2 Explicit horizon	What is the commitment horizon, and can an operator change it?
D3 Prefix respected	Re-plan mid-day. Are executed and communicated stops treated as fixed?
D4 Churn reported	How many communicated stops did yesterday's re-plans move?
D5 Savings-aware waiting	Is a remote route held open as long as a dense one, or committed earlier?
D6 Depot choice	Is an order assigned to the nearest depot, or the one that can actually serve it best?
D7 Cost per stop	Report cost per stop by customer, not just by route. Can it?
D8 Plan against actual	Does the system compare the executed sequence with the planned one, and flag divergence?

D8 is the one that detects the failure mode of Section 3.3. A driver quietly re-sequencing by their own judgement invalidates every subsequent re-plan, and a system that never compares plan with actual will not notice.

6.4 Generalisability

The frozen-prefix formulation generalises to any plan executed while new information arrives: field service dispatch, home healthcare visits, engineer scheduling, port crane allocation. In all of them the same asymmetry holds — information arrives late and the freedom to act on it disappears early.

The cost-per-stop argument generalises to any operation whose unit cost has a fixed, a distance and a service component in different proportions across its work. Its transferable question is simply whether the metric being managed contains the term that actually dominates.

What does not generalise is the value of the category itself. It is proportional to how much demand matures after planning begins, which varies from nearly all in same-day e-commerce to nearly none in scheduled wholesale distribution.

7 Threats to Validity and Limitations

1. No outcome is attributed to this module anywhere. It is named in a deployment whose four published results belong to other modules, so this paper reports no field evidence at all and the absence should be read as informative.
2. The modelled 12% is an assumption. It is the published default of the return model, carries no measurement, and Section 5.3 argues it depends on a site property — how late orders mature — rather than on the software.
3. The horizon cost functions are stylised. Equation (horizon) states the shape of the trade, not a calibrated model; the square-root balance point follows from the chosen functional forms and should not be applied numerically without local estimation.
4. The churn coefficient is not observable. Its true value depends on driver tolerance and customer expectation, and no system can measure the point at which a workforce stops following sequences until after it has happened.
5. Cost per stop conceals its own distribution. A mean across a route hides the outlying stop whose window forced the whole sequence, and a percentile or per-customer view is needed before commercial decisions are made on it.
6. The optimisation objective and the management metric differ. Section 4.4 resolves this pragmatically rather than formally, and a site whose service time genuinely dominates may need an objective this paper does not supply.
7. Traffic and travel-time prediction are outside this paper. Every result assumes the cost matrix is known; in practice it is estimated, and estimation error propagates into every sequence.
8. No figure here is a product capture. This product's page carries no screenshots at all, so nothing in this paper demonstrates that the described interface exists in the form modelled.

The first limitation is the one to weigh. A paper this specific about mechanism with no attributed outcome is describing what the design implies, not what a deployment achieved, and a buyer should treat the two very differently.

8 Future Work

- A measured deployment. The most valuable next step is a published result attributed to this module, so Section 5 can report an outcome rather than a design.
- Per-route commitment horizons. Section 4.3 shows a remote route can be committed early because nothing arriving later will consolidate with it; varying the horizon by expected savings would cut churn without stranding orders.
- Churn as a reported metric. Counting how many communicated stops each re-plan moved would make the second half of Equation (horizon) observable rather than assumed.
- Plan-against-actual reconciliation. Comparing the executed sequence with the planned one detects the silent divergence of Section 3.3 and is a precondition for trusting any re-planning system over time.
- Cost-per-stop distributions rather than means, so that commercial decisions about areas, windows and minimum order values are made on the tail rather than the average.

9 Conclusion

Continuous re-planning is not the six o'clock computation run more often. By mid-morning most of the day is no longer available to change, and a routing system's real design question is how much of it to keep open.

Three results follow. A re-plan operates on a frozen prefix — executed stops and communicated commitments are fixed — so the six o'clock one-shot plan is simply the degenerate case where the horizon spans the day and nothing maturing after dawn can be absorbed. The horizon is a trade with a cost on each side: stranded orders against broken commitments, and both are estimable from a site's own history. And what a late order can capture is the Clarke-Wright saving against whatever tail is still movable, which means a route through a dense cluster is worth holding open and a route to an outlying customer is not.

The fourth result is the one with commercial teeth. Cost per stop and total mileage can move in opposite directions, and a fleet congratulating itself on falling kilometres while its margin erodes is measuring the vehicle instead of the business.

The framework of Section 6.3 is offered as the durable contribution, and its second dimension is the question to open with: what is your commitment horizon, and can I change it?

Appendix A Nomenclature

Table 4. Symbols and abbreviations used in this paper.

Symbol / term	Meaning
c(i,j), t(i,j)	Cost and travel time between stops i and j; node 0 is the depot
q(i), Q(k)	Demand at stop i, and the capacity of vehicle k
[a(i), b(i)], tau(i)	Delivery window at stop i, and the planned arrival time
s(i)	Service duration at stop i — the time spent off the vehicle
H(k)	Driver k's remaining legal hours under Regulation (EC) 561/2006
E, C, F	The executed, committed and free segments of a route
h	Commitment horizon — how far ahead a stop is treated as fixed
lambda(late)	Rate at which orders mature during the operating day
v-bar(miss)	Value lost when a late order cannot be absorbed into an existing route
kappa	Churn coefficient — cost per unit of re-sequencing communicated work
sigma(i,j)	Clarke-Wright saving from serving i and j on one route rather than two
f(k), gamma, omega	Fixed vehicle cost, cost per kilometre, and loaded cost per hour
CPS	Cost per stop — route cost divided by stops served
VRP	Vehicle routing problem, in the Dantzig-Ramser sense

Appendix B Worked Numerical Examples

Appendix B.1 Two routes, one vehicle, opposite conclusions

A vehicle costs 120 per day fixed, 0.90 per kilometre, and 38 per hour loaded. Route A serves 9 rural stops over 420 km with 6.5 hours driving and 12 minutes service each. Route B serves 34 urban stops over 140 km with 3.2 hours driving and 14 minutes service each.

Applying Equation (cps) to Route A: fixed 120, distance 378, time 38 x (6.5 + 1.8) = 315. Total 813 — but with fixed cost apportioned as in the figure, take the comparison on the variable terms. Cost per stop is 813 / 9 = 90.3.

Route B: fixed 120, distance 126, time 38 x (3.2 + 7.9) = 422. Total 668, cost per stop 668 / 34 = 19.6.

Route B travels a third of the distance and costs 82% as much, so a mileage report flatters it enormously and a total-cost report only modestly. Cost per stop says what matters: those nine rural customers cost four and a half times as much to serve as the urban ones, which is a pricing and territory question rather than a routing one.

> Note where Route B's cost actually is. Service time — 7.9 hours across 34 doors — is its largest single term, and no distance-based optimisation touches it at all.

Appendix B.2 Setting a commitment horizon from a day of order timestamps

A site dispatches at 06:00 and closes deliveries at 18:00. Order timestamps show 340 orders confirmed before dispatch and 96 maturing across the operating day — roughly 8 per hour.

Its mean Clarke-Wright saving for an absorbable order is 14 currency units, so lambda(late) x v-bar(miss) = 8 x 14 = 112 per hour of horizon. Its churn coefficient, estimated from the cost of re-sequencing communicated stops, is 450.

Applying Equation (horizon): h* = sqrt(450 / 112) = 2.0 hours. Stops whose planned arrival falls within two hours are treated as committed; everything beyond is re-plannable.

Compare the one-shot alternative. With a horizon of 12 hours, all 96 late orders are stranded, costing 96 x 14 = 1,344 in unrealised consolidation — against a churn cost of 450/12 = 37.5. The trade is wildly out of balance, which is a precise statement of what the six o'clock plan was costing that site.

Appendix B.3 Why a remote route should be committed early

Two candidate stops, both 40 km from the depot. Stop P is 3 km from an existing stop on a dense route; stop R is 55 km from the nearest stop on a rural route.

Applying Equation (savings) with symmetric costs at 0.90 per km: for P, sigma = 36 + 36 — 2.7 = 69.3. For R, sigma = 36 + 36 — 49.5 = 22.5.

The dense-cluster order is worth three times as much to absorb. Holding both routes open for the same two hours pays the same churn on each and buys three times the value on one of them.

A per-route horizon set from expected savings would keep the dense route open beyond two hours and commit the rural one at dispatch — capturing more of the valuable consolidations while disturbing fewer drivers. The information needed to do this is the geometry the planner already has.

Data provenance

Every quantitative claim in this paper resolves to one of the entries below. Figures marked as modelled estimates are not deployment measurements; their assumptions are stated.

Metric	Reported value	Provenance
Route planning cadence the product replaces	continuous re-planning against a 6 a.m. one-shot plan	**Product specification** /products/fleet-optimization
Primary last-mile margin metric the system reports	cost per stop, ahead of mileage or average dwell	**Product specification** /products/fleet-optimization
Constraints carried inside the route model rather than checked after release	vehicle capacity, driver hours, customer windows, cold chain, hazardous goods	**Product specification** /products/fleet-optimization
Operating cost recovered through route, load and vehicle utilisation	12% across 20 vehicles at 2,200 h/yr (modelled)	**Modelled estimate** 20 vehicles in the fleet, 2,200 operating hours each per year; 12% utilisation and route-efficiency gain attributed to the module; Model and defaults published in src/data/roiModels.ts

References

[1] Dantzig, G. B., and Ramser, J. H. (1959). The truck dispatching problem. Management Science, 6(1), 80-91.

[2] Clarke, G., and Wright, J. W. (1964). Scheduling of vehicles from a central depot to a number of delivery points. Operations Research, 12(4), 568-581.

[3] Little, J. D. C. (1961). A proof for the queuing formula: L = lambda W. Operations Research, 9(3), 383-387.

[4] J. F. C. Kingman (1961). The Single Server Queue in Heavy Traffic. Mathematical Proceedings of the Cambridge Philosophical Society, vol. 57, no. 4, pp. 902-904; the first statement of the approximation relating waiting time to utilisation and variability. https://doi.org/10.1017/S0305004100036094

[5] D. G. Kendall (1953). Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. The Annals of Mathematical Statistics, vol. 24, no. 3, pp. 338-354; the source of the A/B/c notation used to classify queueing systems. https://doi.org/10.1214/aoms/1177728975

[6] European Parliament and Council of the European Union (2006). Regulation (EC) No 561/2006 on the harmonisation of certain social legislation relating to road transport. Official Journal of the European Union, 15 March 2006; Article 6 caps daily driving at 9 hours, extendable to 10 twice weekly, with 56 hours weekly and 90 across two weeks, and Article 7 requires a 45-minute break after 4.5 hours of driving. Amended by Regulation (EU) 2020/1054, applicable from 1 August 2020. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32006R0561

[7] International Chamber of Commerce (2019). Incoterms 2020: ICC Rules for the Use of Domestic and International Trade Terms. ICC Publication No. 723E, Paris; launched September 2019 and in force from 1 January 2020, defining the eleven trade terms that allocate cost and risk between seller and buyer.

[8] International Organization for Standardization (2022). ISO 28000: Security and resilience - Security management systems - Requirements. ISO, Geneva; published 15 March 2022. https://www.iso.org/standard/79612.html

[9] International Organization for Standardization / International Electrotechnical Commission (2024). ISO/IEC 19987: Information technology - EPC Information Services (EPCIS) Standard, version 2.0. ISO/IEC, Geneva. https://www.iso.org/standard/85557.html

[10] International Organization for Standardization / International Electrotechnical Commission (2024). ISO/IEC 19988:2024 - Information technology - GS1 Core Business Vocabulary (CBV). ISO/IEC, Geneva, edition 3, March 2024; publishes GS1 CBV 2.0, ratified June 2022, which fixes the vocabularies for the EPCIS business step, disposition and business transaction type fields. https://www.iso.org/standard/85558.html

[11] GS1 (2024). Serial Shipping Container Code (SSCC). GS1 identification key series; an 18-digit key for a logistic unit, carried under application identifier 00. https://www.gs1.org/standards/id-keys/sscc

[12] GS1 (2024). GS1 General Specifications. GS1 AISBL; defines the GTIN, SSCC and GLN identification keys. https://ref.gs1.org/standards/genspecs/

[13] International Organization for Standardization / International Electrotechnical Commission (2014). ISO/IEC 15459-1: Information technology - Automatic identification and data capture techniques - Unique identification - Part 1: Individual transport units. ISO/IEC, Geneva. https://www.iso.org/standard/54779.html

[14] International Organization for Standardization (2014). ISO 22400-1: Automation systems and integration - Key performance indicators (KPIs) for manufacturing operations management - Part 1: Overview, concepts and terminology. ISO, Geneva. https://www.iso.org/standard/56847.html

[15] International Organization for Standardization (2014). ISO 22400-2: Automation systems and integration - Key performance indicators (KPIs) for manufacturing operations management - Part 2: Definitions and descriptions. ISO, Geneva; defines OEE as availability x effectiveness x quality ratio. https://www.iso.org/standard/54497.html

[16] International Society of Automation (2013). Enterprise-control system integration - Part 3: Activity models of manufacturing operations management. ANSI/ISA-95.00.03-2013 (IEC 62264-3 Modified).

[17] European Committee for Standardization (2017). EN 16798-3: Energy performance of buildings - Ventilation for buildings - Part 3: For non-residential buildings - Performance requirements for ventilation and room-conditioning systems (Modules M5-1, M5-4). CEN, Brussels; interpreted by CEN/TR 16798-4. Current edition EN 16798-3:2025.

[18] de Koster, R., Le-Duc, T., and Roodbergen, K. J. (2007). Design and control of warehouse order picking: A literature review. European Journal of Operational Research, 182(2), 481-501.

[19] Gu, J., Goetschalckx, M., and McGinnis, L. F. (2007). Research on warehouse operation: A comprehensive review. European Journal of Operational Research, 177(1), 1-21.

[20] Kagermann, H., Wahlster, W., and Helbig, J. (2013). Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative INDUSTRIE 4.0. Final report of the Industrie 4.0 Working Group, acatech - National Academy of Science and Engineering. https://www.acatech.de/wp-content/uploads/2018/03/Final_report__Industrie_4.0_accessible.pdf