



Beyond the Limits of Traditional Optimization

**How GPU Acceleration is Transforming
Supply Chain Planning**



Foreword

by Chris Doersen, JBF Consulting

Assumptions and truisms are prevalent in many aspects of the transportation industry, though one that rarely generates a discussion is the traditional idea of using Central Processing Units (CPUs) to solve routing problems. For decades it has been a given that more power and faster processors would speed up optimization and generate a better result. But this path reached its diminishing returns far longer ago than most of us have realized.

It has taken an entirely new, and in hindsight exceedingly logical, idea of shifting away from CPUs and using the latest breakthrough in Graphics Processing Units (GPUs) to transform our approach to transportation algorithms. The results are – and we state this without exaggeration – a massive leap forward in processing times and quality of results.

We are only scratching the surface of what these optimization speeds can unlock. This will go well beyond what can improved in a single fleet routing cycle, ultimately impacting the entire tactical process of daily and weekly routing, before then transforming what can be accomplished with strategic supply chain network optimization.

Executive Summary

Traditional CPU-based solvers have for a long time been the biggest bottleneck in optimizing supply chain and distribution networks. As global logistics operations are growing in complexity, the limitation of this approach has only gotten worse. In addition, traditional CPU-based solvers not only take long to solve, but they are also reaching limits as routing, scheduling, and compliance constraints become more critical to generating executable results and saving costs.

Chora, a GPU-powered optimization engine developed by Opsi Systems that is used in its TMS platform Tramm, represents a fundamental leap forward – delivering faster, more scalable, and higher-quality route optimization. This paper, co-authored with JBF Consulting, presents validated performance findings from both a comparative study between Tramm's original solver and results from two well-respected commercial routing and scheduling tools.

In this paper we will describe how traditional methods have reached a plateau in performance, while the burgeoning AI technology infrastructure has opened a new world of possibilities that were unavailable only a few years ago and that Chora can now leverage to profound effect.

The Industry Challenge

As supply chains expand, transportation planners face more complex optimization problems: multi-day horizons, time windows, dock scheduling, fatigue rules, dynamic capacity constraints, traffic, and more. Existing solvers built on traditional CPU architectures struggle to maintain both speed and solution quality at the scale demanded by modern supply chains. These systems, designed decades ago, are ill-equipped for the parallelized, data-intensive needs of today's logistics routing problems.

Why Traditional Solvers Plateau

Even leading optimization engines, like those underpinning current market-leading transportation management platforms, depend primarily on CPU-based architecture. As problem size and constraint diversity grow, computation time scales exponentially. In real-world conditions, this leads to long solve times, suboptimal fleet utilization, and reduced agility.

Chora's predecessor, Solo (built on Google OR-Tools), has been used successfully with numerous customers and is already considered to be a strong solver by industry standards. Despite that success, Solo is still limited by its CPU-based optimization approach, motivating a full architectural redesign by Opsi Systems based on the theoretical improvements using GPUs as the primary optimization tool.

The Chora Architecture

Chora is a GPU-powered vehicle routing solver engineered to address large-scale problems involving complex constraints. Built using NVIDIA's CUDA framework, it leverages thousands of GPU threads and advanced genetic algorithms to explore solution spaces faster and more thoroughly than CPU solvers. This GPU-first approach enables Chora to deliver high-quality solutions in seconds – solving problems that previously took minutes or hours.

Implications for the Logistics Industry

Chora's success heralds a structural change in how optimization can be performed. GPU-based computing scales linearly with problem size, offering consistent performance even in highly complex, multi-constraint environments. This technology empowers planners to re-optimize in near real-time, improving fleet utilization, reducing costs and emissions, and enabling dynamic responsiveness to operational variability.

Opsi Systems Internal Study: Tramm's Original Solver (Solo) Versus Chora

Opsi Systems conducted a study benchmarking Chora against Tramm's original solver, Solo. Four of the study scenarios are outlined below. Note that all the scenarios used the same data set that included:

- 1 Depot
- 34 Vehicles
- 417 Jobs and Sites

Scenario 1: Single Day Planning with Time Windows and Capacity Constraints

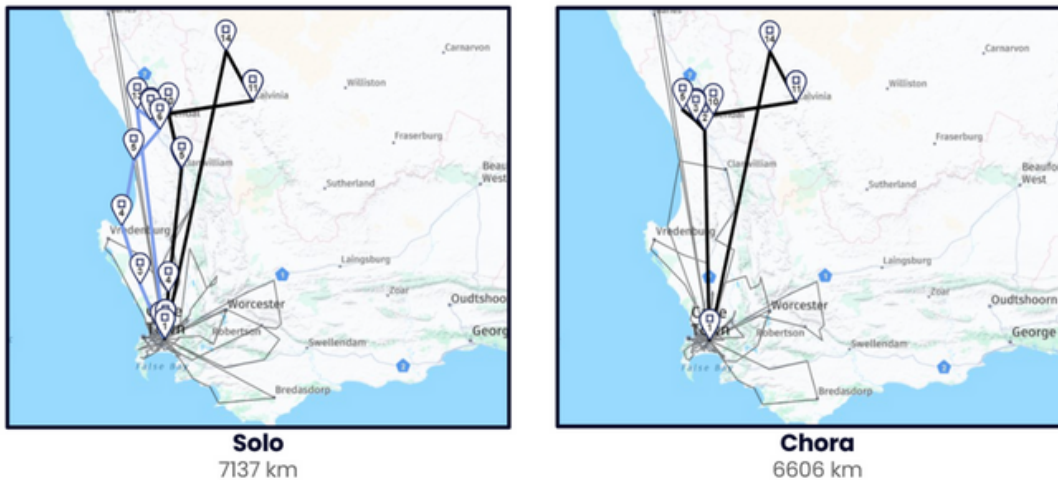
The first scenario was a simple case that looked at single day planning with time windows and capacity constraints.

All of the 417 jobs were planned by both Solo and Chora. Solo planned them all in 10 minutes while Chora managed to plan them all in 1 minute. However, Chora showed significant improvements in trip duration, distance travelled, vehicles used, and costs shown in the table below:

Solver	Runtime	Jobs Planned	Total Trip Duration	Total Trip Distance	Total Vehicles Used	Total Cost
Solo	10 min	417/417	430 hr 38 min	7137 km	27/34	\$77 385.70
Chora	1 min (-90%)	417/417 (0%)	405 hr 48 min (-6.1%)	6606 km (-7.4%)	26/34 (-3.7%)	\$74 586.10 (-3.6%)

To understand how Chora was able to achieve such significant improvements, we examined the routes provided by each solver. Screen shots of the two solutions are shown below.

Two major differences stand out. For the northern section of trips, you can see that Chora was able to consolidate two of the major long-distance trips into a single trip. For the south-eastern trips, Chora managed to consolidate three of the major long-distance trips into two trips.



Decreasing distance by 7.4% and saving a vehicle was an unexpected large improvement for such a simple scenario.

Scenario 2: Scenario 1 Constraints Plus Planning of High-Priority Jobs Over Low-Priority Jobs

The second scenario includes the constraints from Scenario 1 (single day planning with time windows and capacity constraints) but adds the prioritization of high-priority jobs over low-priority jobs while respecting working hours constraints.

The idea here is to simulate an environment where we do not have sufficient vehicles to deliver all jobs but maintain a rolling set of jobs on the floor prioritizing jobs that are reaching the end of their delivery window. This approach is often used to maximize vehicle utilization and to reduce the number of trips to distant locations.

In this scenario, the available vehicles were reduced from 34 to 12. The scenario included 150 high-priority jobs and 267 low-priority jobs.

Solo managed to plan 128 out of the 150 high-priority jobs and 177 out of the 417 total jobs. This was done in a solve time of 10 minutes. Chora planned all 150 of the 150 high-priority jobs and 201 of the 417 total jobs. This was done in a solve time of 1 minute. Both Solo and Chora used all 12 vehicles available. See the results in the table below:

Solver	Runtime	High-Priority Jobs Planned	Total Jobs Planned	Total Vehicles Used
Solo	10 min	128/150	177/417	12
Chora	1 min	150/150 (+17.2%)	201/417 (+11.9%)	12

Adding priorities is well known to be difficult with traditional solving methods¹; therefore we weren't surprised that Solo did not do well with this constraint. The fact that Chora beat Solo by 11% in the total number of jobs planned was an impressive level of improvement.

Scenario 3: Scenario 1 Constraints Plus Fatigue Rules and Time Windows Extended Over Two Days

The third scenario includes the constraints from Scenario 1 (single day planning with time windows and capacity constraints) but extends the delivery windows over a two-day period and includes European Union (EU) driver fatigue rules (similar to US fatigue rules).

In this scenario, a midday break of 30 minutes must be taken after 4.5 hours of working. In addition, an overnight break of 11 hours must be taken after 13 total hours of working or 9 hours of driving.

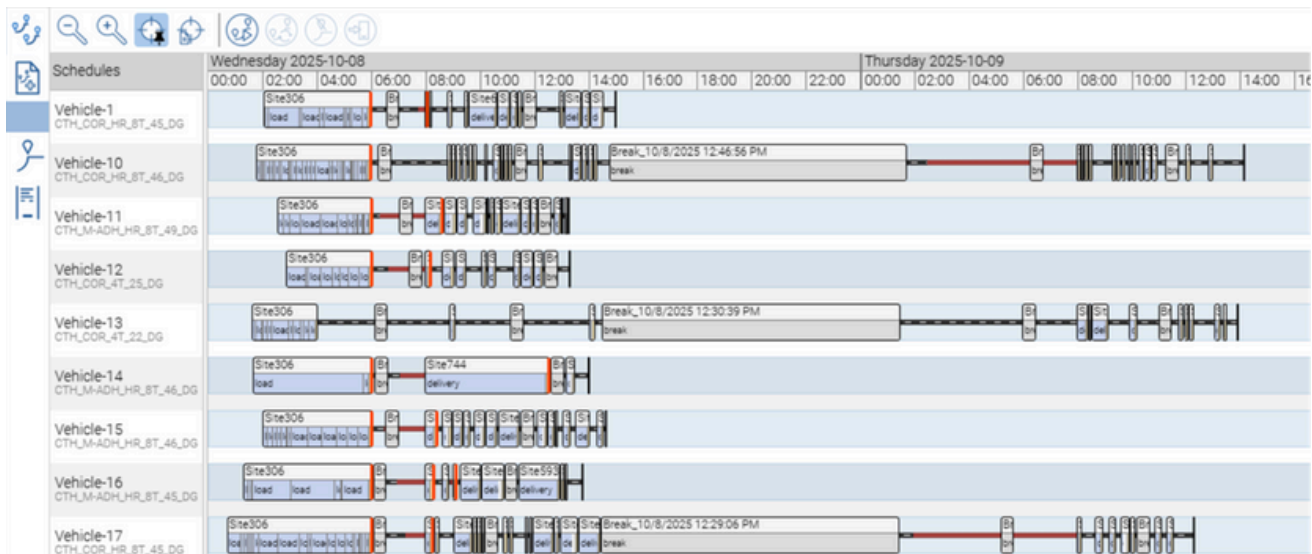
¹ Doan, T. T., Bostel, N., & Hà, M. H. (2021). The vehicle routing problem with relaxed priority rules (VRP-RPR). EURO Journal on Transportation and Logistics, 10, Article 100039. <https://doi.org/10.1016/j.ejtl.2021.100039>

Solo managed to plan 196 out of the 417 jobs in a solve time of 60 minutes. Chora planned 416 of the 417 jobs in a solve time of 1 minute. See the results in the table below:

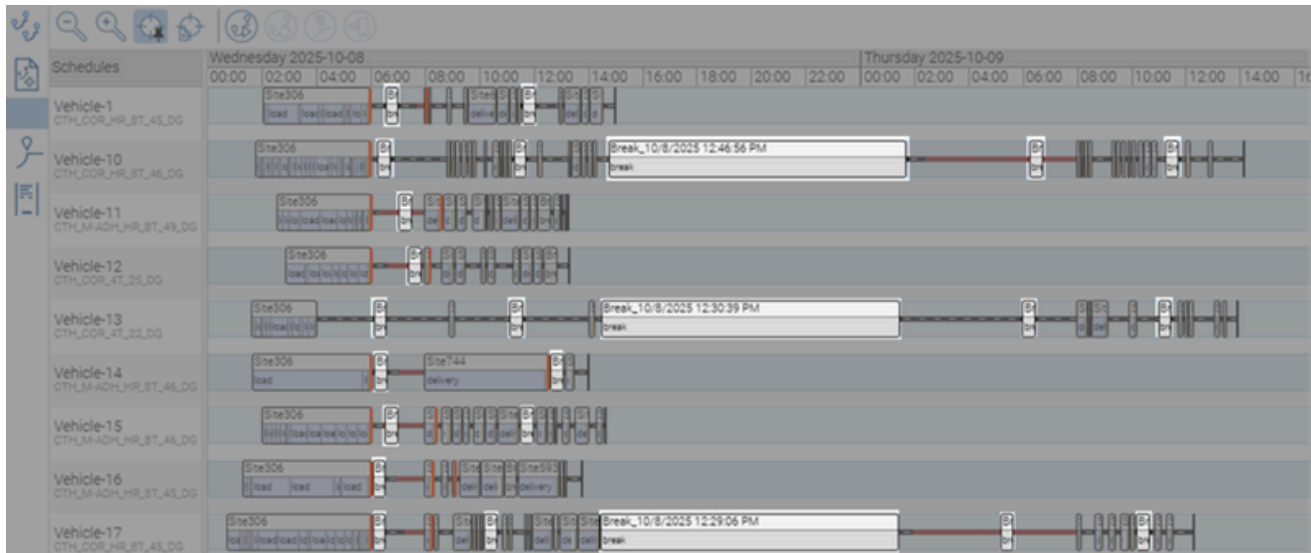
Solver	Runtime	Jobs Planned
Solo	60 min	196/417
Chora	1 min	416/417 (+112%)

This is a completely dominant result highlighting the inability of the Solo solver to adequately handle fatigue breaks.

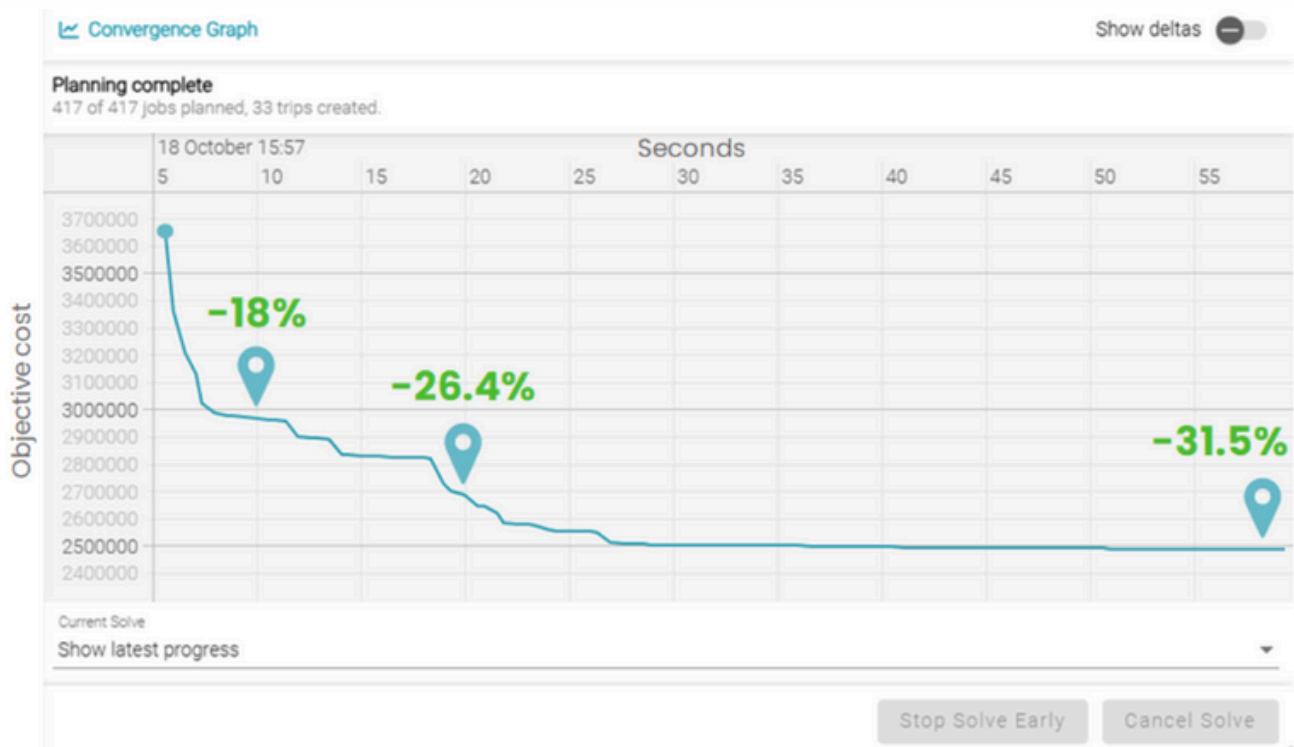
The Gantt chart below illustrates the distribution of the breaks:



The 30-minute and 4.5-hour breaks are highlighted in the image below:



It's also interesting to view Chora's convergence graph to understand how results became optimal over the course of the 60 second solve time:



Elapsed Time	Objective Cost	% Reduction
5 seconds	3635714	N/A
10 seconds	2967346	-18%
20 seconds	2676530	-26.4%
60 seconds	2491795	-31.5%

Scenario 4: Scenario 1 Constraints Plus Loading Bays

The fourth scenario includes the constraints from Scenario 1 (single day planning with time windows and capacity constraints) but included the constraint of having a single loading dock with only 8 loading bays at the depot. These loading bays are set up to allow reservation slots in 15-minute units.

Note that there aren't enough loading bays to load all vehicles at once so the loading bays must be staggered in such a way as to minimize the total waiting time of the vehicles.

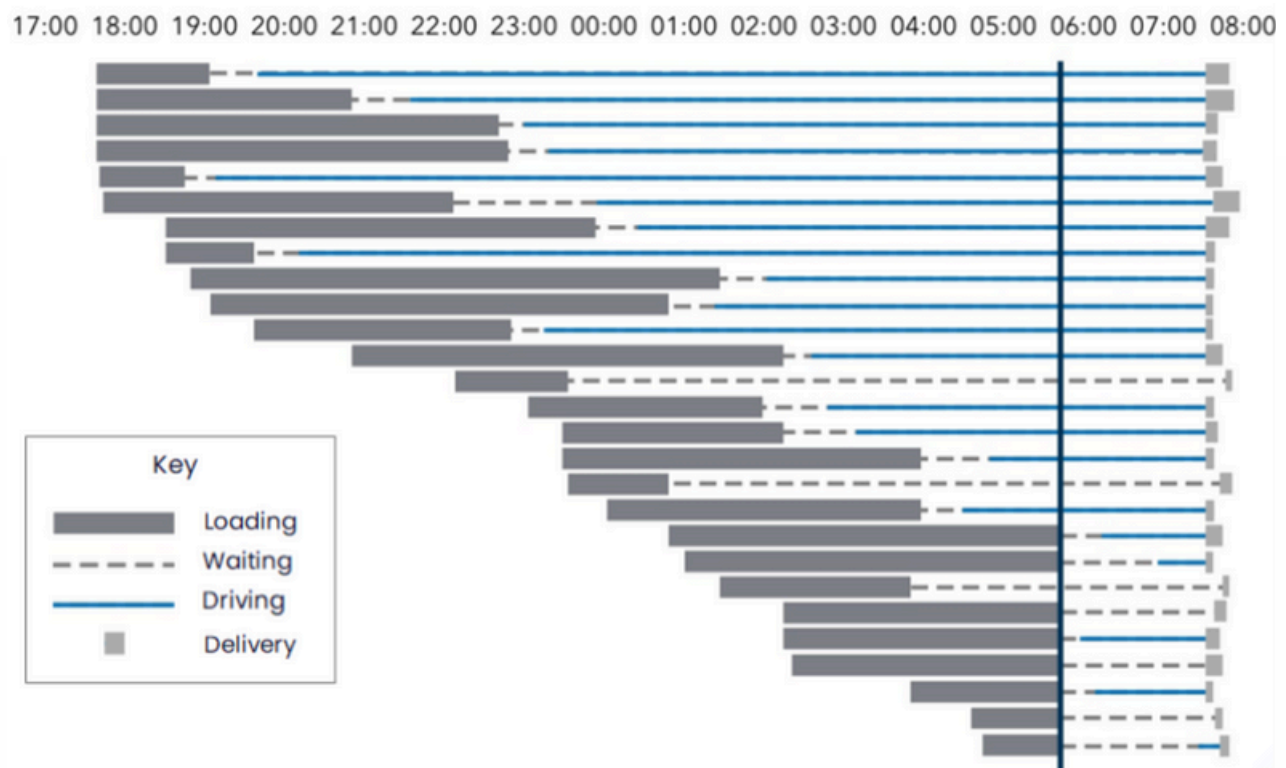
Solo did not manage to find a solution after a 60-minute runtime. Chora was able to find a solution in 1 minute and managed to plan 383 out of the 451 jobs while achieving 95.3% loading dock utilization and 100% loading slot utilization.

Some of the jobs could not be planned since they could not all be loaded due to the individual loading durations combined with the limited number of available loading bays, despite the 12-hour loading window that was available (18:00-06:00).

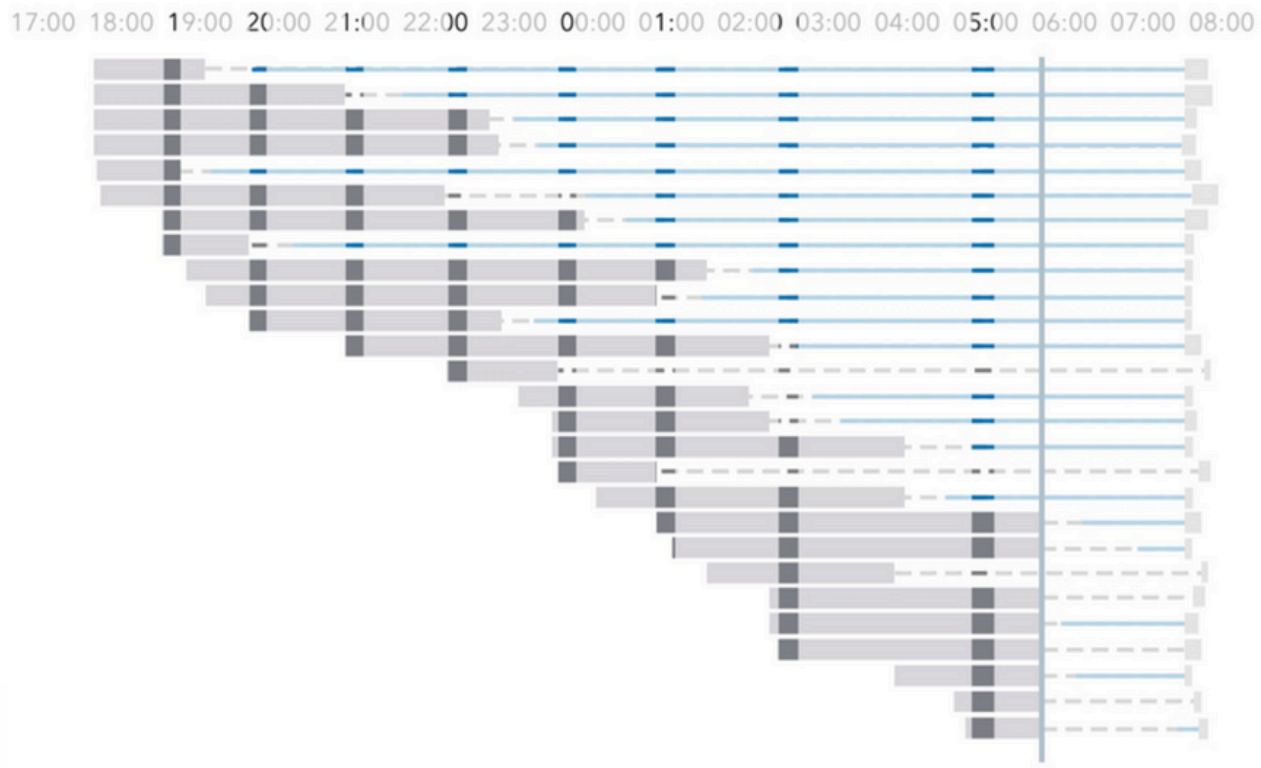
See the results in the table below:

Solver	Runtime	Solution Found	Jobs Planned	Loading Dock Utilisation	Loading Slot Utilisation
Solo	60 min	No	N/A	N/A	N/A
Chora	1 min	Yes	383/451	95.3%	100%

The table below illustrates 27 vehicles where only 8 can be loaded at a time. It is apparent that trips that start with long driving legs are loaded first, which minimizes the total waiting time.



The table below illustrates the 8 loading bays and emphasizes that only 8 trips may be loaded at a time:



Real-World Validation: Leading U.S. Manufacturer Study

JBF Consulting conducted a benchmarking study of Chora for a leading U.S. manufacturer. The study compared Chora's GPU-based solver to vendor solvers from two leading providers under realistic operational conditions including multi-day routing, capacity, regional constraints, and fatigue rules.

The objective was to solve for optimal trip distance (as opposed to trip duration) while maintaining optimal vehicle utilization.

Real-World Validation: Leading U.S. Manufacturer Study

Regions

The data defined four distinct operating regions, and the solver was required to respect these boundaries – meaning no route was allowed to cross into another region. The four regions were:

- NY - New York (Upstate)
- DE - Delaware & Maryland
- VA - Virginia
- PA - Pennsylvania

Fatigue Rules

Two fatigue-related constraints were enforced:

- Working Break: Between any two 14-hour work windows (driving + labor), a 10-hour break must occur. A driver may drive a maximum of 11 hours within that 14-hour window, after which the break is required regardless of total labor time.
- Driving Break: After a maximum of 8 consecutive hours of driving, a 30-minute break is mandatory

Vendor Data Differences

The key difference between Vendor-A and Vendor-B was the planning horizon:

- Vendor-A: Operated over 3 days, with vehicles required to return to the depot by 19:30 on Day 3 (soft constraint – some trips returned slightly later).
- Vendor-B: Operated over 4 days, with the same soft 19:30 return time applied to Day 4.

Vehicle Cap

No maximum fleet size was provided. To ensure feasibility, we supplied 30 vehicles for Chora to utilize during the solve.

Key Indicative Results

Vendor A vs. Chora

Solver	Runtime	Jobs Planned	Total Trip Duration	Total Trip Distance	Total Vehicles Used
Vendor-A	~2 hours	845/845	1237 hr 42 min	18 239,10 miles	27
Chora	1 min (-99.1%)	845/845 (0%)	1154 hr 22 min (-6.7%)	15 622,63 miles (-14.4%)	26 (-3.7%)

Vendor B vs. Chora

Solver	Runtime	Jobs Planned	Total Trip Duration	Total Trip Distance	Total Vehicles Used
Vendor -B	18 min	845/845	1204 hr 01 min	16 373,98 miles	24
Chora	1 min (-94.4%)	845/845 (0%)	1280 hr 19 min (+6.1%)	15 219,51 miles (-7.1%)	23 (-4.2%)

The vendors were using CPU-based optimizers, and, like any vehicle routing solver, they were subject to the exponential complexity inherent in the Capacitated Vehicle Routing Problem (CVRP)². As the number of jobs and sites increases, the complexity grows exponentially. This problem set contained 478 unique sites and 845 jobs (465 collections and 380 deliveries), with the requirement that collections and deliveries at the same site could not be separated.

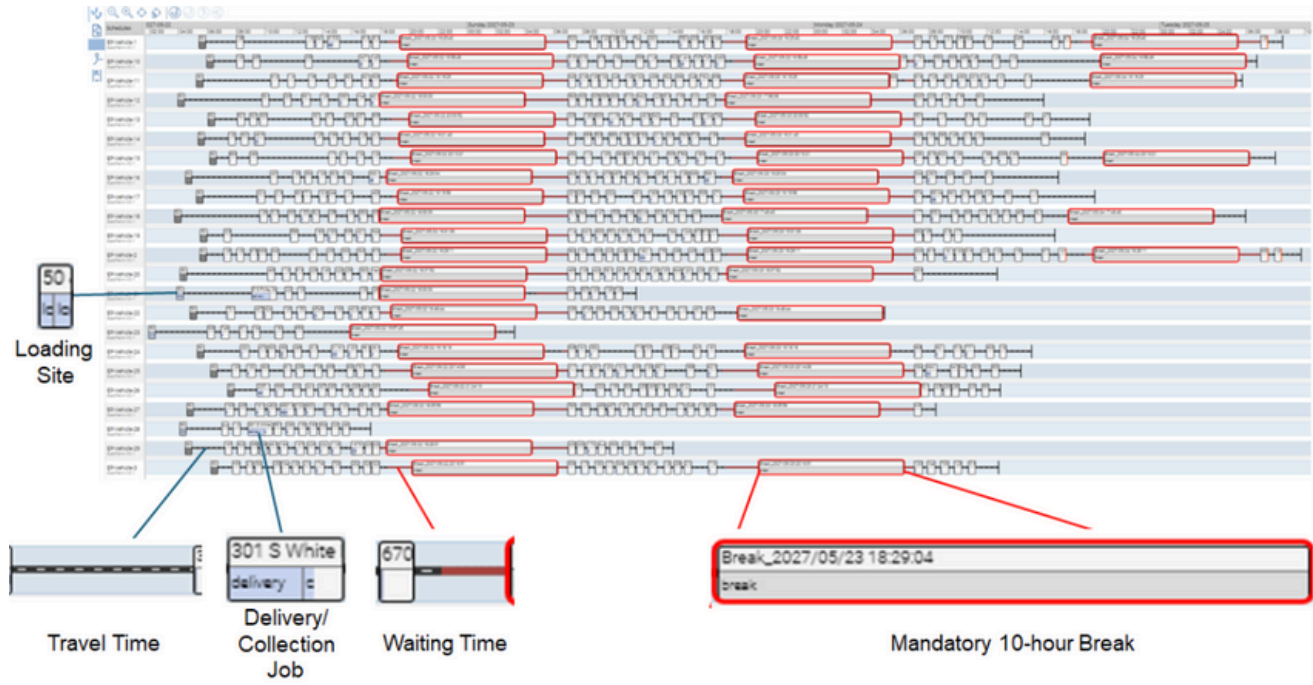
All loading also had to be completed within the first two days of the window; fatigue rules needed to be applied and respected, and strict regional constraints were enforced. A problem of this size, combined with these operational constraints, represents a significant challenge for any CPU-based solver.

Vendor-A struggled the most, taking 2 hours to produce its solution. Vendor-B achieved a stronger result, with the extra planning day making the problem easier to solve, which still required 18 minutes to complete.

Chora, benefiting from massive parallel execution on the GPU, produced an even better solution than both Vendor-A and Vendor-B in a fraction of the time – representing a 99.2% speed-up over Vendor-A and a 94.4% speed-up over Vendor-B. Importantly, Chora ran on only 1/6 of a GPU for these tests; using a full GPU would improve run-time by up to an additional factor of 6.

² Chen, Y. (2023). Approximation Schemes for Capacity Vehicle Routing Problems: A Survey. arXiv. <https://doi.org/10.48550/arXiv.2306.01826>

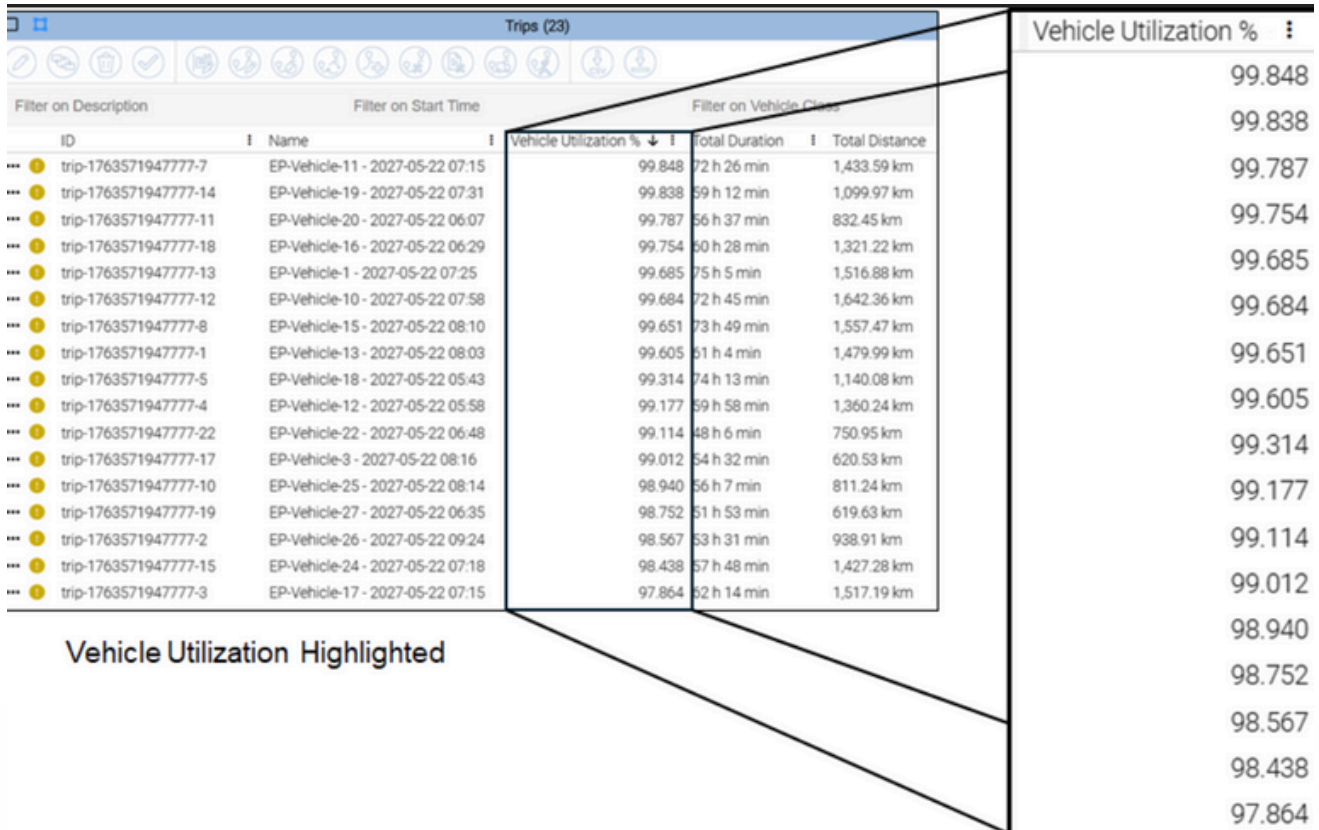
The Gantt chart below shows the routes from Chora over a four-day period, including sleep-outs and breaks for fatigue rule adherence:



“GPU-accelerated optimization represents the next frontier in transport planning efficiency.”

- CTO, Opsi Systems

The below screenshot from Tramm illustrates how Chora was able to achieve extremely high vehicle utilization across the board:



ID	Name	Vehicle Utilization %	Total Duration	Total Distance
trip-1763571947777-7	EP-Vehicle-11 - 2027-05-22 07:15	99.848	72 h 26 min	1,433.59 km
trip-1763571947777-14	EP-Vehicle-19 - 2027-05-22 07:31	99.838	59 h 12 min	1,099.97 km
trip-1763571947777-11	EP-Vehicle-20 - 2027-05-22 06:07	99.787	56 h 37 min	832.45 km
trip-1763571947777-18	EP-Vehicle-16 - 2027-05-22 06:29	99.754	50 h 28 min	1,321.22 km
trip-1763571947777-13	EP-Vehicle-1 - 2027-05-22 07:25	99.685	75 h 5 min	1,516.88 km
trip-1763571947777-12	EP-Vehicle-10 - 2027-05-22 07:58	99.684	72 h 45 min	1,642.36 km
trip-1763571947777-8	EP-Vehicle-15 - 2027-05-22 08:10	99.651	73 h 49 min	1,557.47 km
trip-1763571947777-1	EP-Vehicle-13 - 2027-05-22 08:03	99.605	51 h 4 min	1,479.99 km
trip-1763571947777-5	EP-Vehicle-18 - 2027-05-22 05:43	99.314	74 h 13 min	1,140.08 km
trip-1763571947777-4	EP-Vehicle-12 - 2027-05-22 05:58	99.177	59 h 58 min	1,360.24 km
trip-1763571947777-22	EP-Vehicle-22 - 2027-05-22 06:48	99.114	48 h 6 min	750.95 km
trip-1763571947777-17	EP-Vehicle-3 - 2027-05-22 08:16	99.012	54 h 32 min	620.53 km
trip-1763571947777-10	EP-Vehicle-25 - 2027-05-22 08:14	98.940	56 h 7 min	811.24 km
trip-1763571947777-19	EP-Vehicle-27 - 2027-05-22 06:35	98.752	51 h 53 min	619.63 km
trip-1763571947777-2	EP-Vehicle-26 - 2027-05-22 09:24	98.567	53 h 31 min	938.91 km
trip-1763571947777-15	EP-Vehicle-24 - 2027-05-22 07:18	98.438	57 h 48 min	1,427.28 km
trip-1763571947777-3	EP-Vehicle-17 - 2027-05-22 07:15	97.864	52 h 14 min	1,517.19 km

“Not only did we see the quality and cost benefits of GPU-based solvers like Chora, the dramatic reduction in solve time marks a paradigm shift in logistics optimization.”

- Principal, JBF Consulting

Implications for the Logistics Industry

Chora's success signals a structural change in how optimization can be performed. The enormous computing power in GPU's allows for consistent performance even in large datasets with highly complex multi-constraint environments. This technology empowers planners to re-optimize in near real-time, improving fleet utilization, reducing emissions, and enabling dynamic responsiveness to operational variability.

Future Vision

Chora demonstrates how AI and GPU acceleration can fundamentally redefine logistics optimization. Opsi Systems is expanding the framework to include traffic and weather constraints, continuous real-time optimization, and predictive analytics for end-to-end supply chain efficiency.

The Authors

Chris Doersen

Principal, Client Engagement
JBF Consulting

Chris.doersen@jbf-consulting.com
www.jbf-consulting.com

Dan Palay

Director of Sales & Marketing
Ops Systems

Dan.palay@opsi.co.za
www.tramm.global

About JBF Consulting

JBF Consulting is a leading logistics strategy advisory and technology integration firm that partners with shippers to transform their logistics and supply chain execution operations. We empower clients to achieve operational efficiency and scalable, sustainable value through strategy development, roadmap orchestration, unbiased technology selection, expert implementation, data-driven insights, and ongoing managed services. For over two decades, our client-centric approach and alliances with best-of-breed solution providers have ensured that every strategy and solution we deliver drives measurable impact, long-term success, and customer satisfaction. For more information, visit www.jbf-consulting.com.

About Ops Systems

Ops Systems is a global provider of advanced transportation management and optimization solutions. Its flagship platform, Tramm, combines planning, execution, real-time visibility, and analytics into a single integrated environment. Powered by Chora, Ops's proprietary GPU-accelerated optimization engine, Tramm delivers near-real-time routing performance and high-quality results at scale.

Serving leading retailers, manufacturers, logistics providers, and transporters, Ops helps organizations improve fleet utilization, reduce costs, and operate with greater agility across both owned and contracted transport networks. Learn more at www.tramm.global.

JBF Consulting and Ops Systems invite logistics software vendors, 3PLs, and supply chain innovators to engage with the findings of this white paper. Organizations seeking to enhance solver performance or explore strategic partnerships in GPU-powered optimization are encouraged to contact either company for further collaboration. We welcome companies with difficult optimization problems to contact us, and we will perform sample runs of optimization exercises for them at no cost.*

*Contact us for terms and conditions.