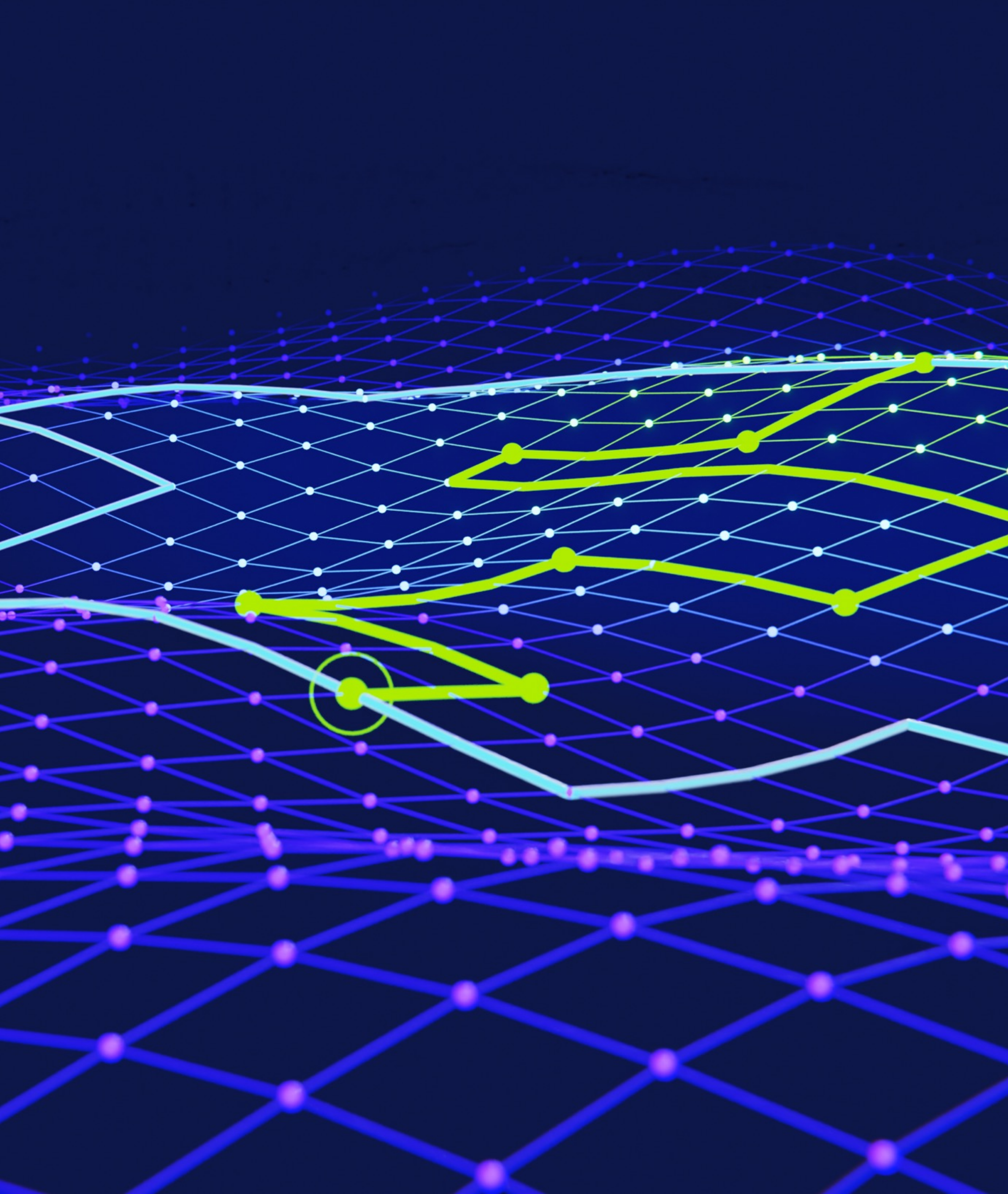




S62495 - Advances in Optimization AI

Alex Fender Ph.D. | GTC 2024



Agenda

- Latest on cuOpt Vehicle Routing

- cuOpt AI agent concept

- Accelerating Mathematical Optimization

NVIDIA cuOpt

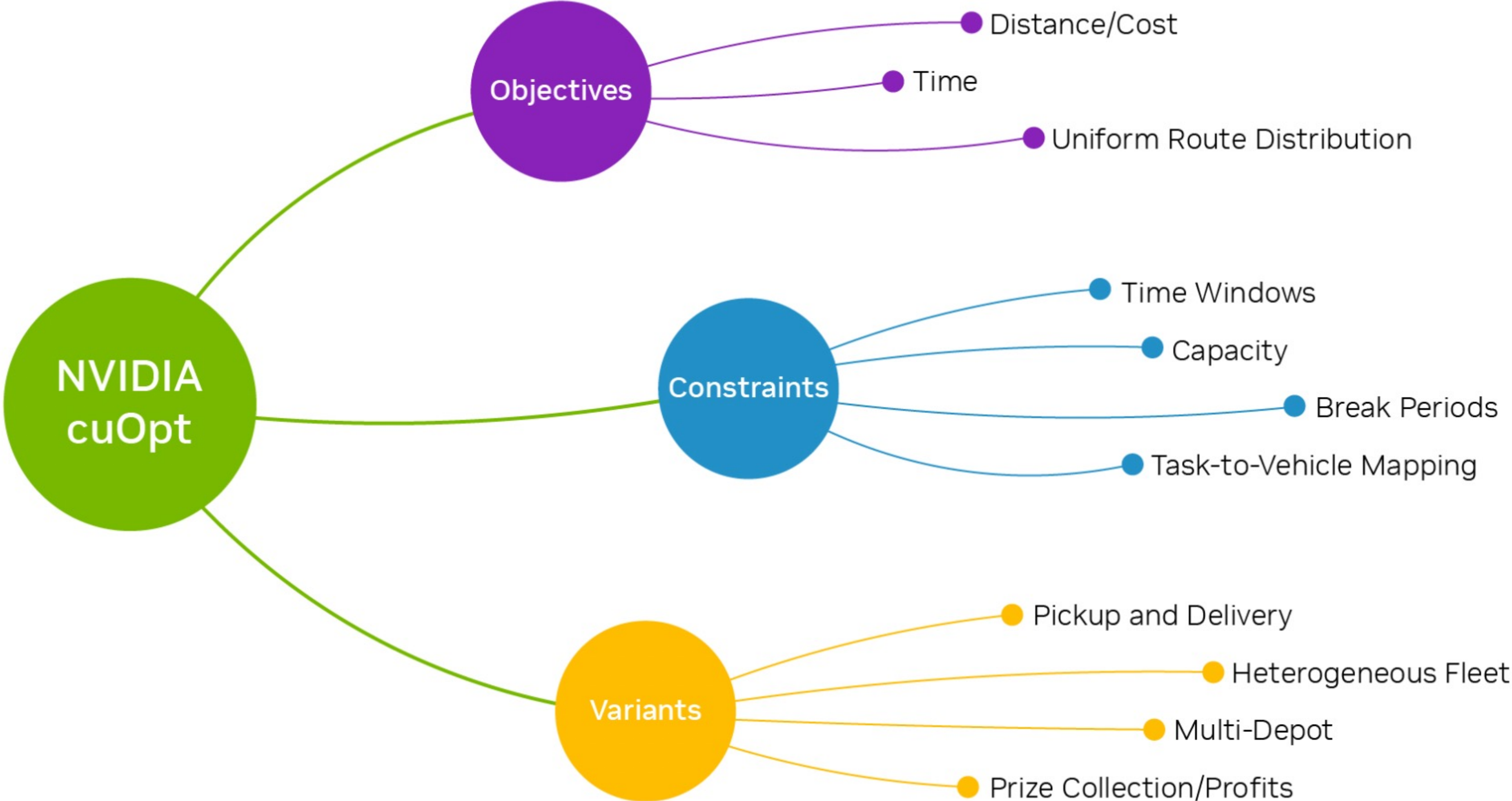
Accelerated Optimization Engine

Vehicle routing optimization with **world's best** planning

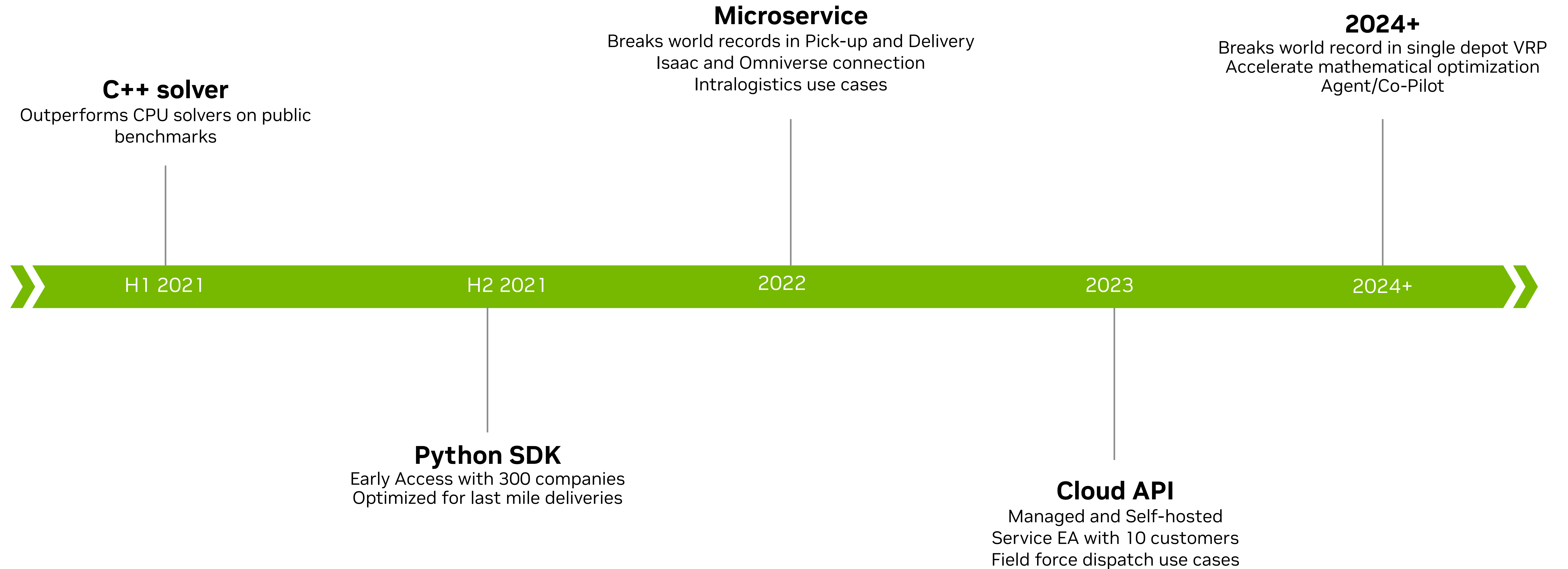
Dynamic re-planning in near real time

Supports the largest problems with GPU acceleration (>15k tasks at once)

Multi constraint, multi objective
Hundreds of problem variants

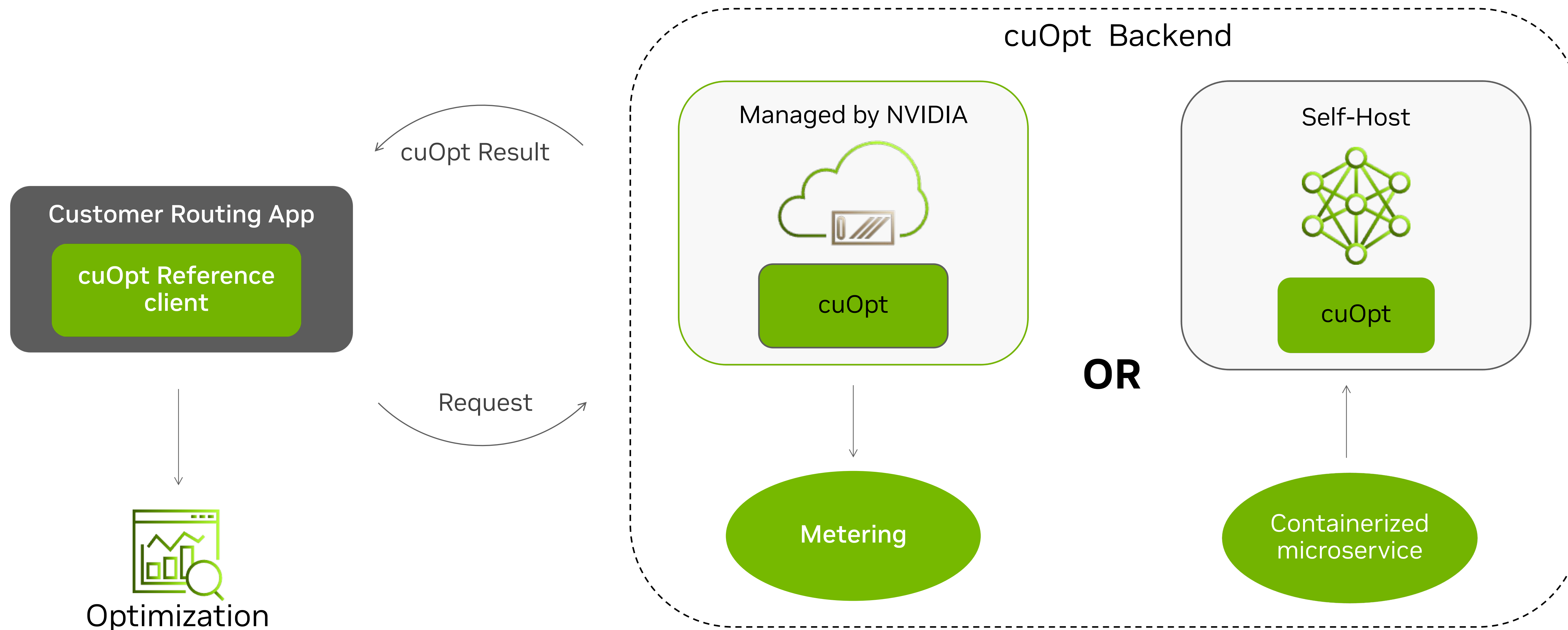


The cuOpt journey

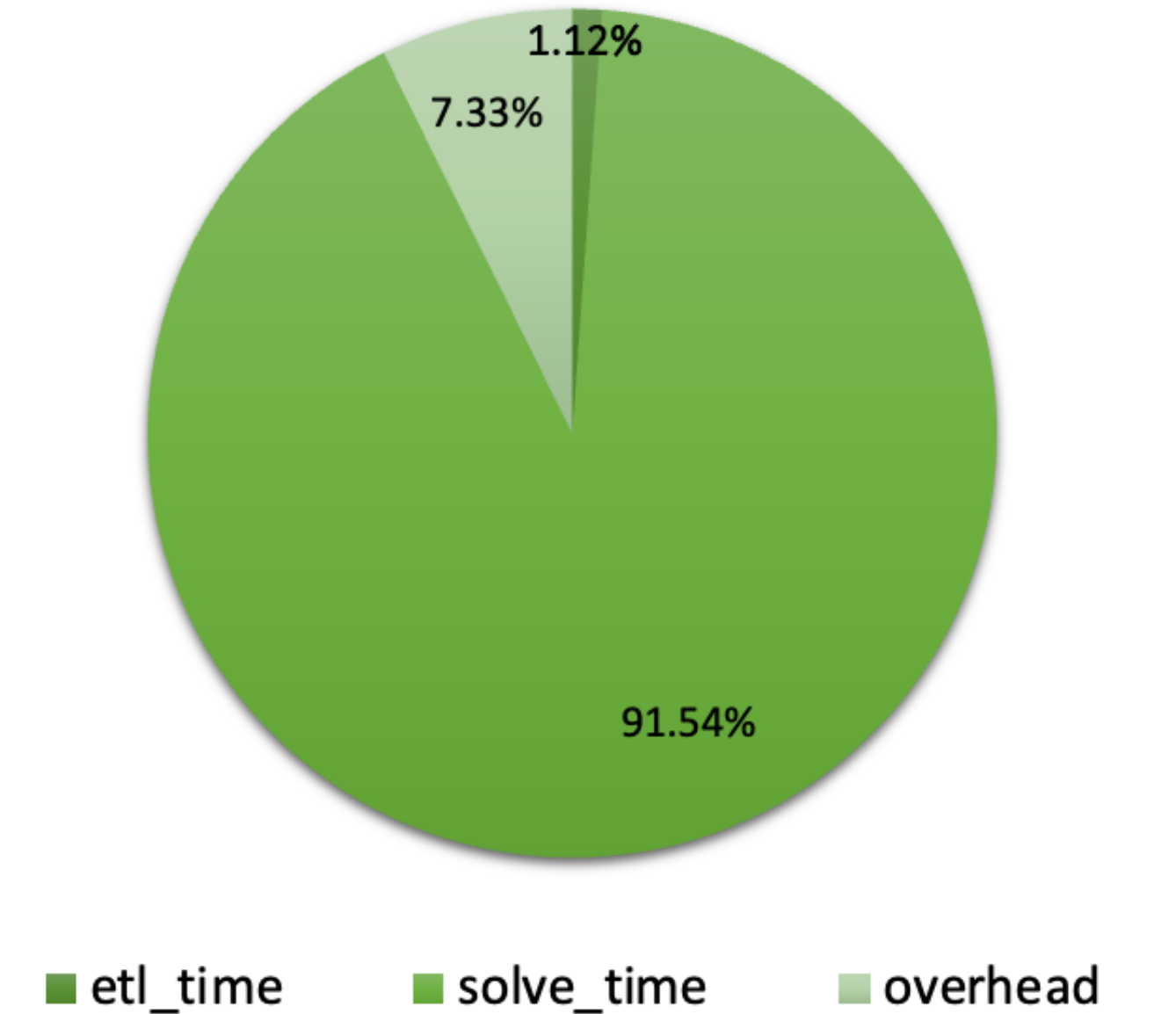


Now available as an API

Managed or Self-hosted



API performance profile



Performance

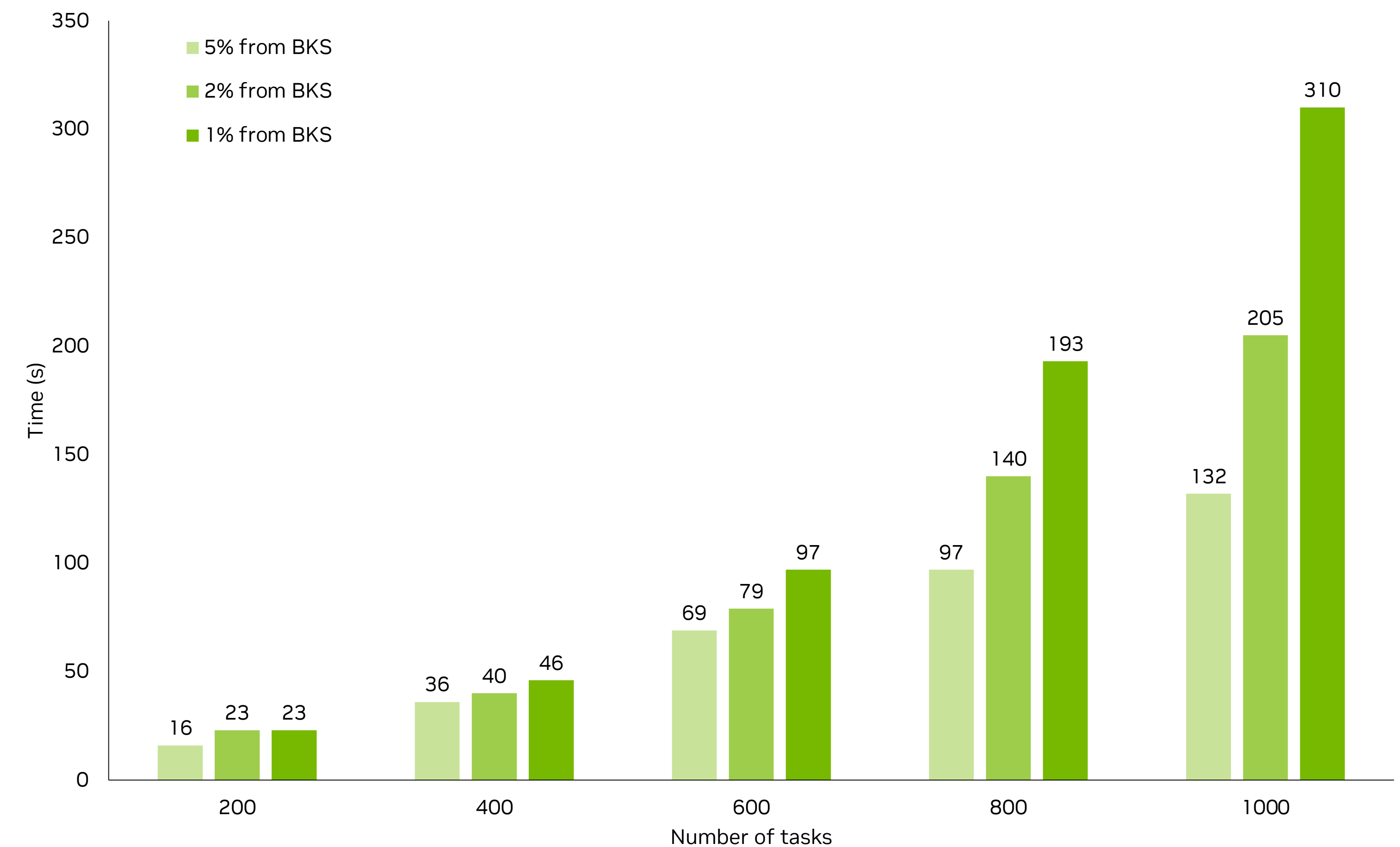
Optimize in seconds

- Comparing to the World's Best Known Solutions (BKS)
- Hierarchical objective, minimize:
 - number of vehicles: BKS found in seconds
 - total distance: displayed for various problem sizes

cuOpt finds the **best number of vehicle in all cases**

cuOpt times include finding the world's best number of vehicles

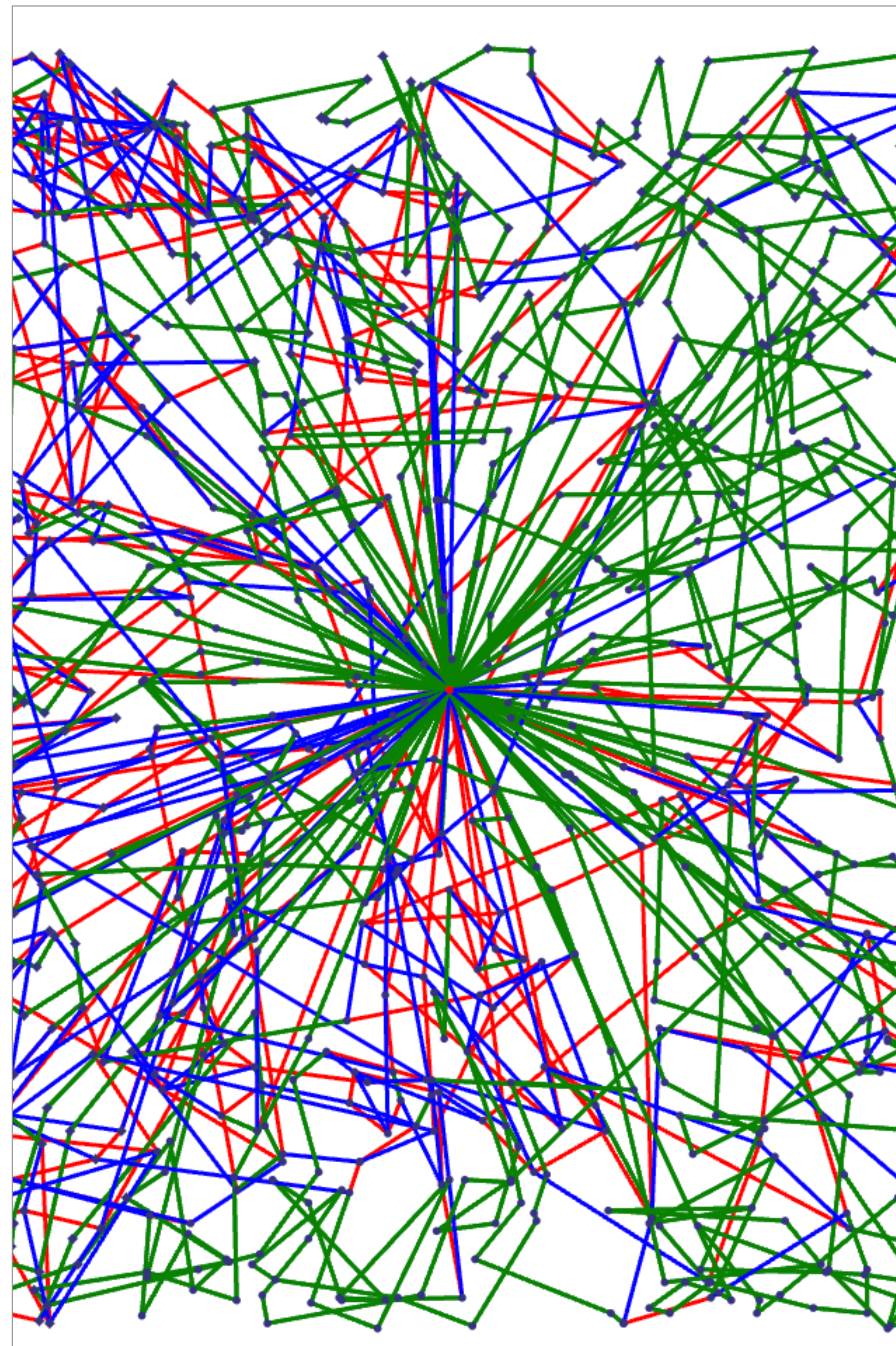
Time to gradually approach the BKS distance on Li&Lim



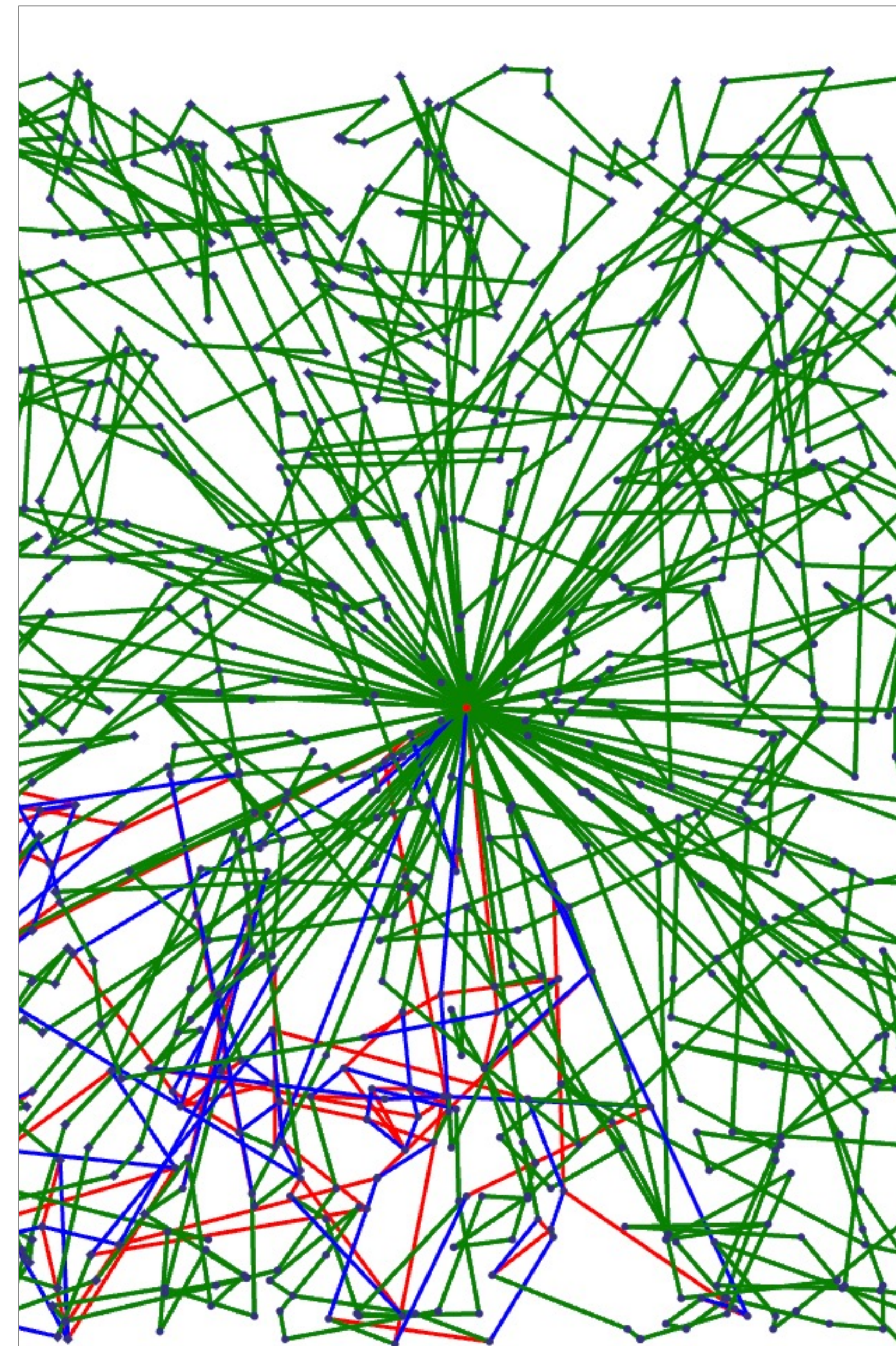
World's best routes

23 world records on Gehring&Homberger and Li&Lim

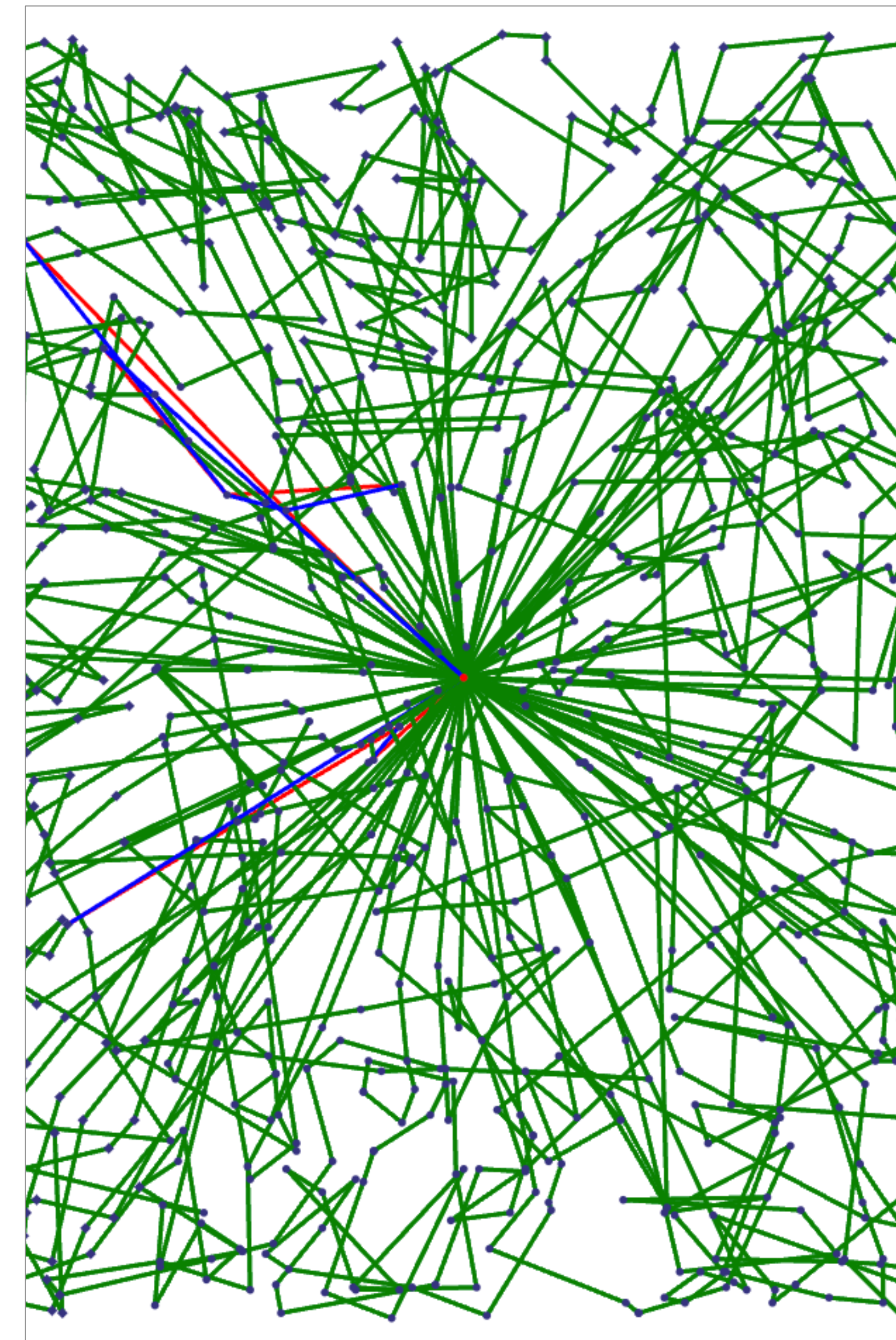
Ir1_10_6 : 4 world record claimed YoY, +.85% compounded improvement



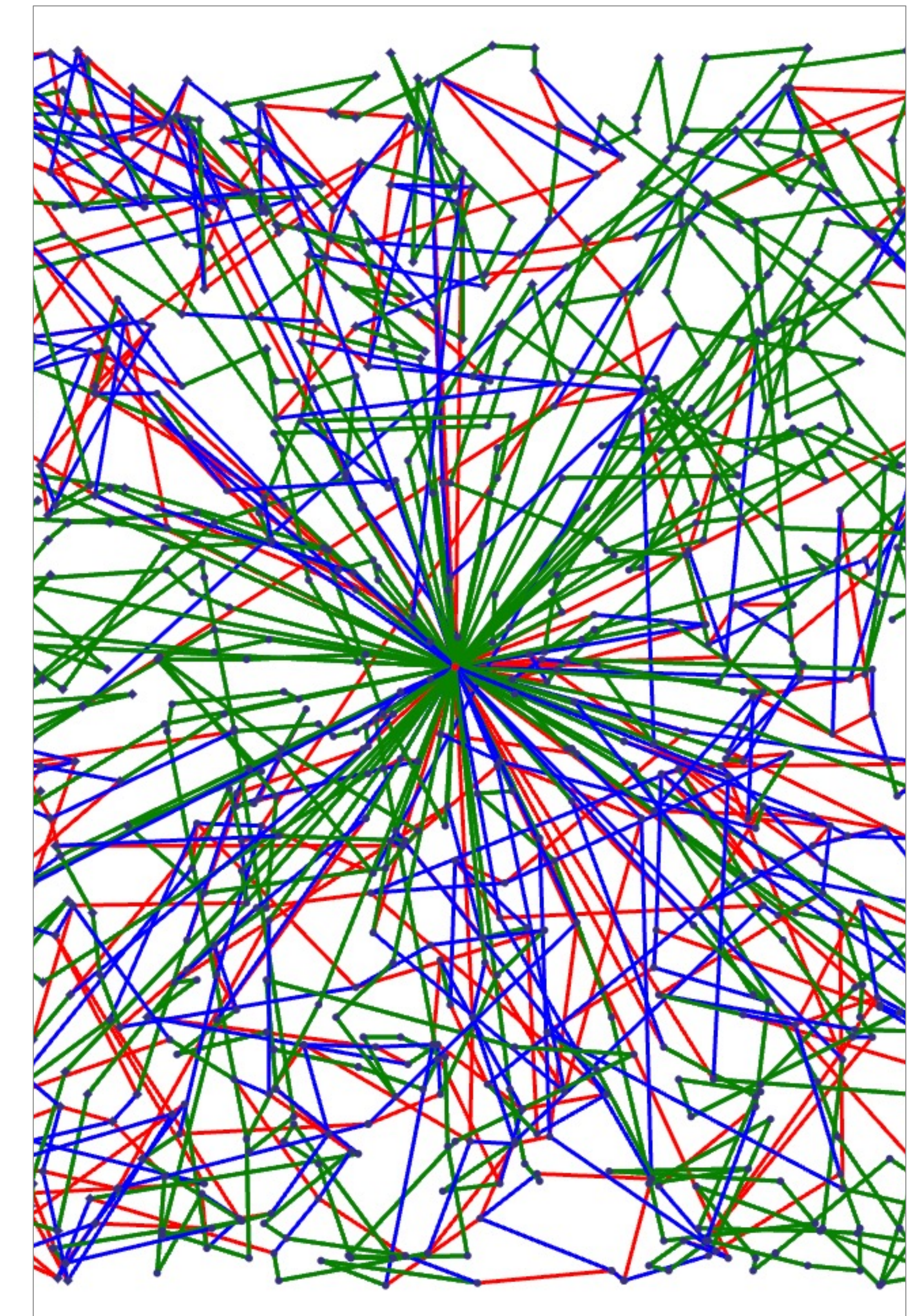
442 days, +.22%
NVIDIA 03/2023



71 days, +.01%
Veeroute 05/2023



4 days, +0.30%
Otimo 05/2023



256 days, +0.31%
NVIDIA 02/2024

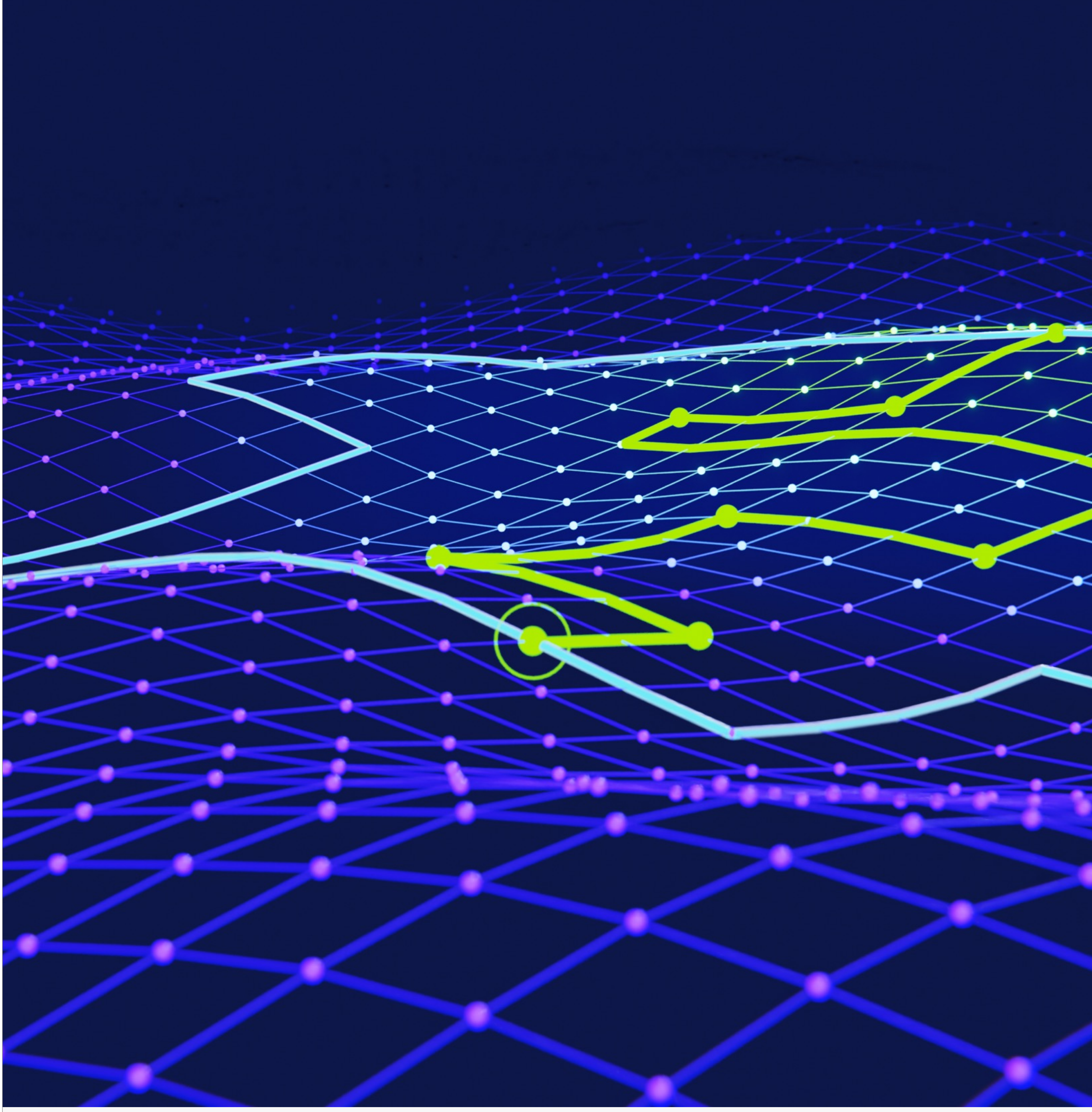
■ both solutions ■ only in previous solution ■ only in new solution

Best solutions and datasets available on SINTEF

Viz credit: combopt.org

Mathematical Optimization

GPU acceleration



Mathematical Optimization

Linear Programming

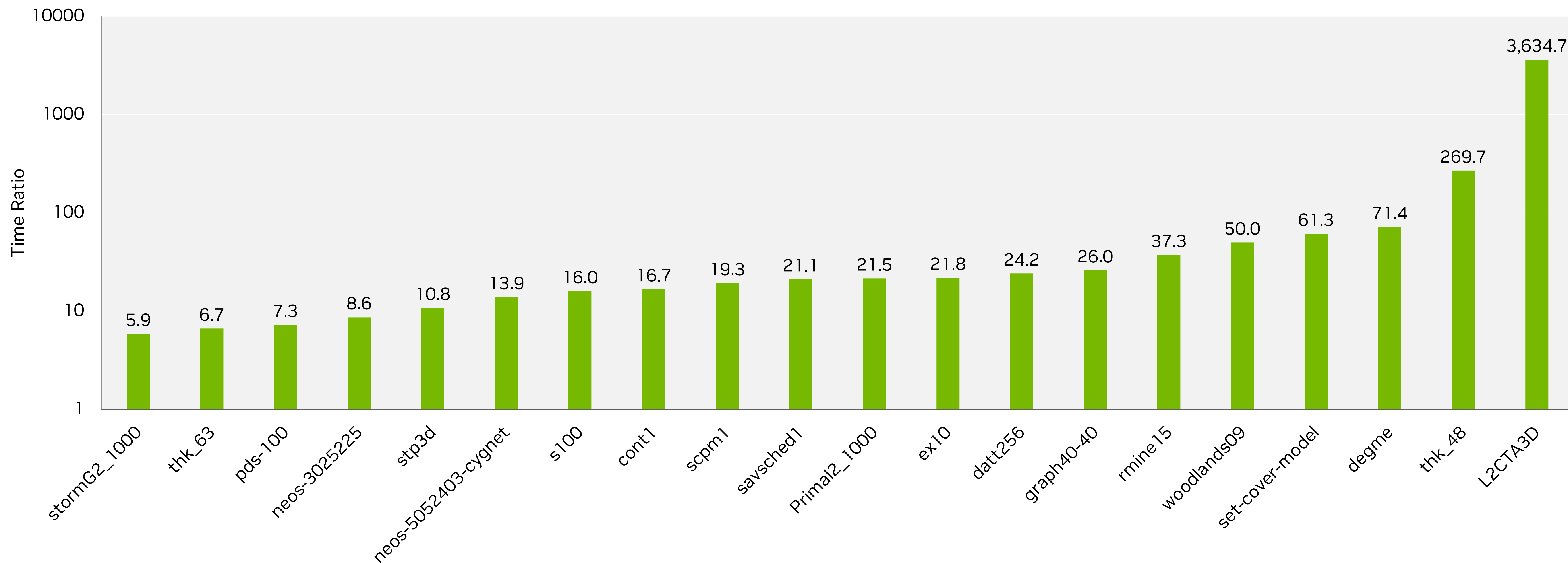
- Optimize an objective function subject to constraints
 - Decision Variables (X_1, X_2, \dots): Quantities to determine (e.g., production levels).
 - Objective Function (Z): The function to optimize (e.g., maximize profit).
 - Constraints ($Ax \leq b$): Linear inequalities that limit the decisions (e.g., resource availability).
-
- GPU can accelerate LP, enabling more efficient and faster optimization at scale



GPU acceleration of Linear Programming

Comparison against Commercial LP on selected instances of Mittleman's Benchmark

cuOpt acceleration compared to state-of-the-art CPU LP (higher is better)

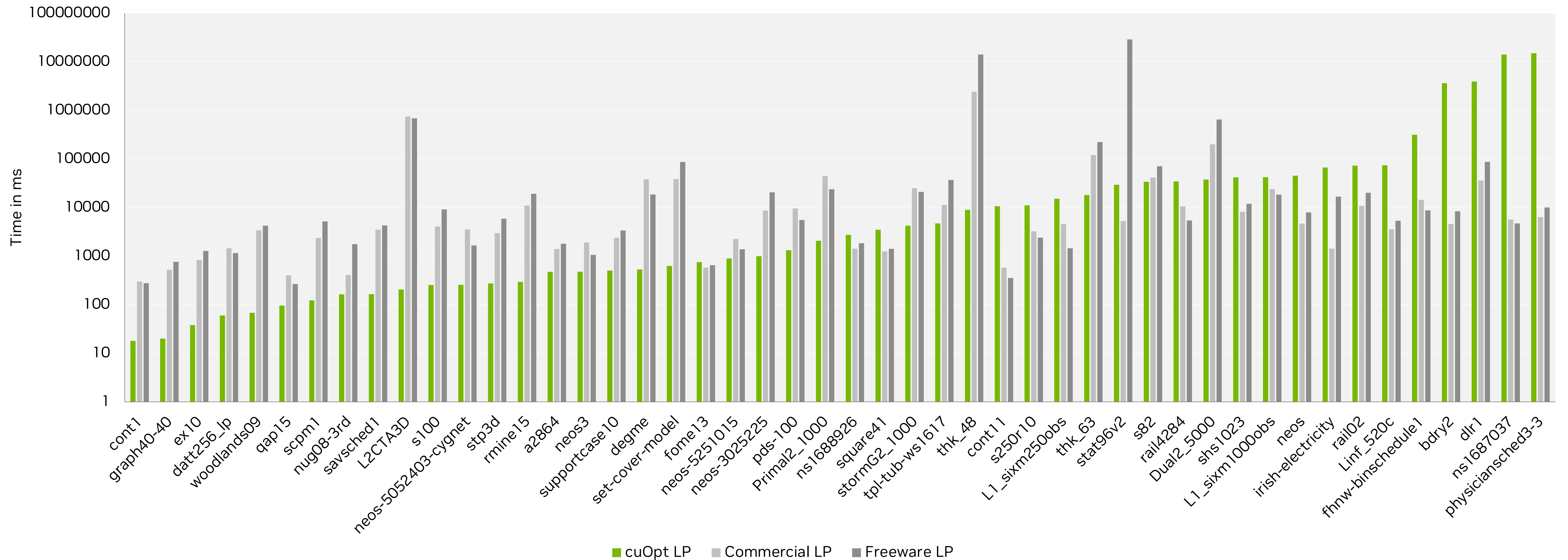


GPU: NVIDIA H100-HBM3
CPU: AMD EPYC 7313P - 16 Cores @ 3.0GHz
Tolerance : 1e-3

NVIDIA cuOpt Linear Programming Solver

Comparison against Commercial LP solvers on [Mittleman's Benchmark](#)

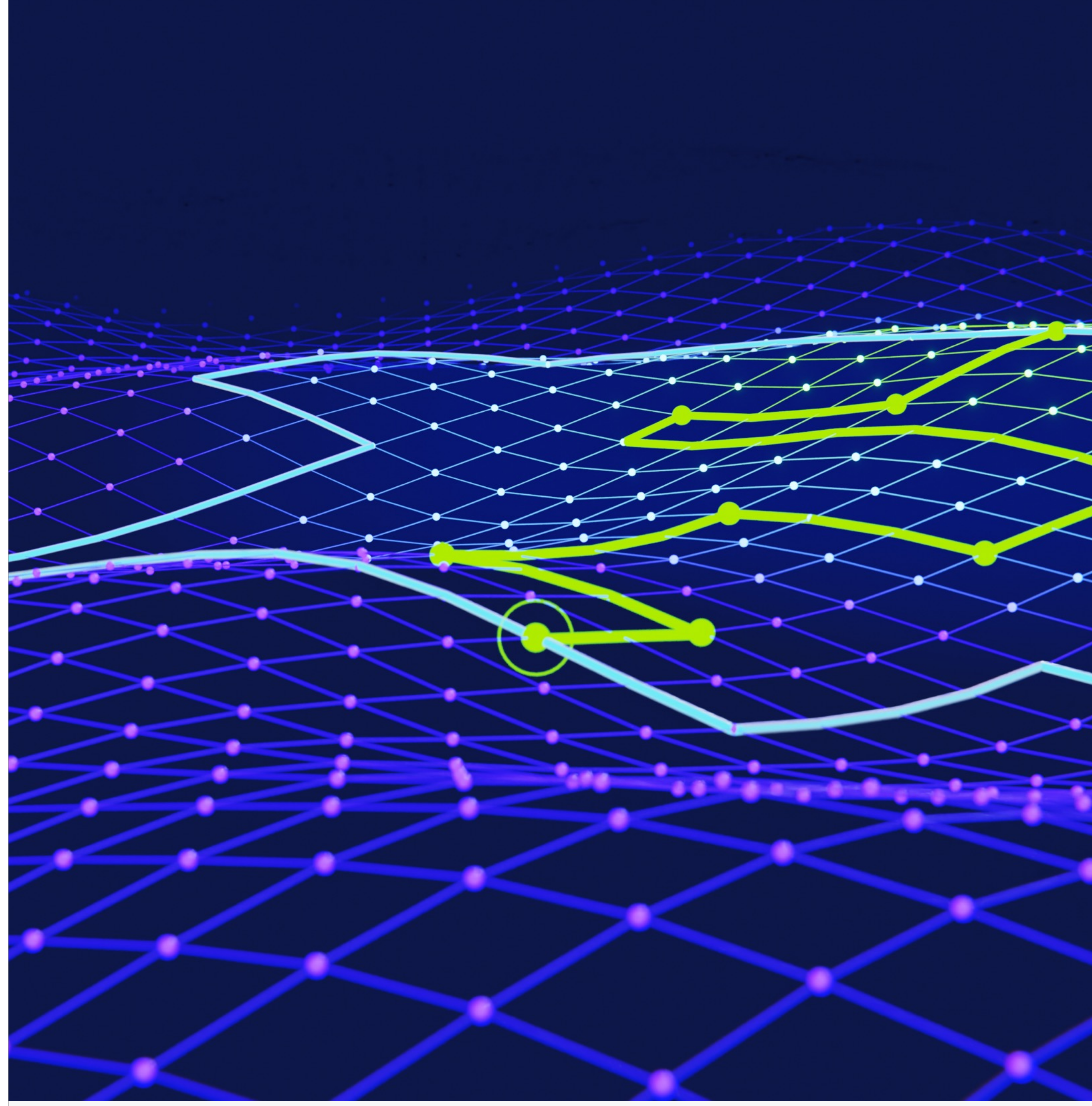
Time Comparison, cuOpt with state-of-the-art solvers (lower is better)



GPU: NVIDIA H100-HBM3
 CPU: AMD EPYC 7313P - 16 Cores @ 3.0GHz
 Tolerance : 1e-3

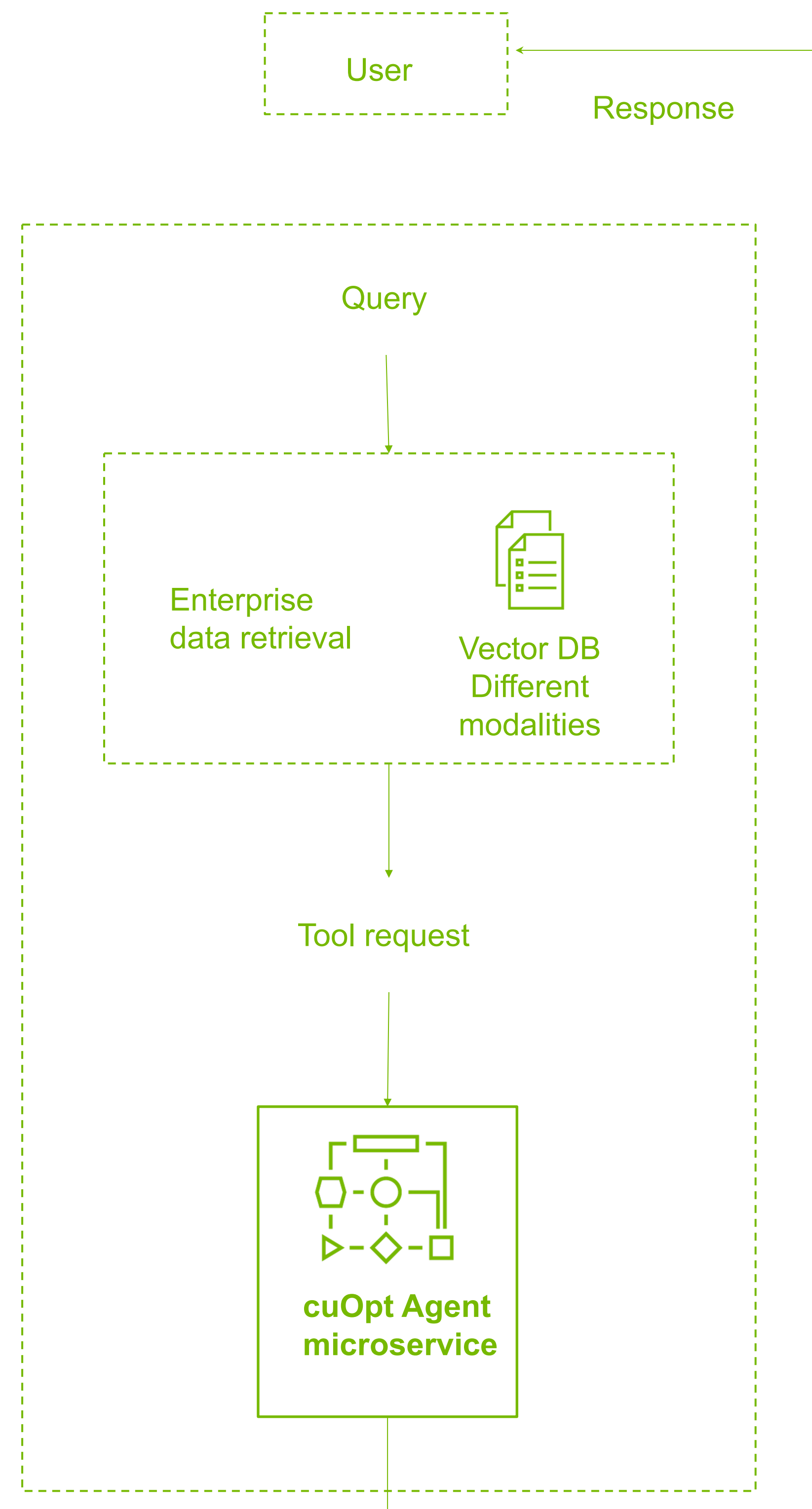
cuOpt AI Agent

Concept

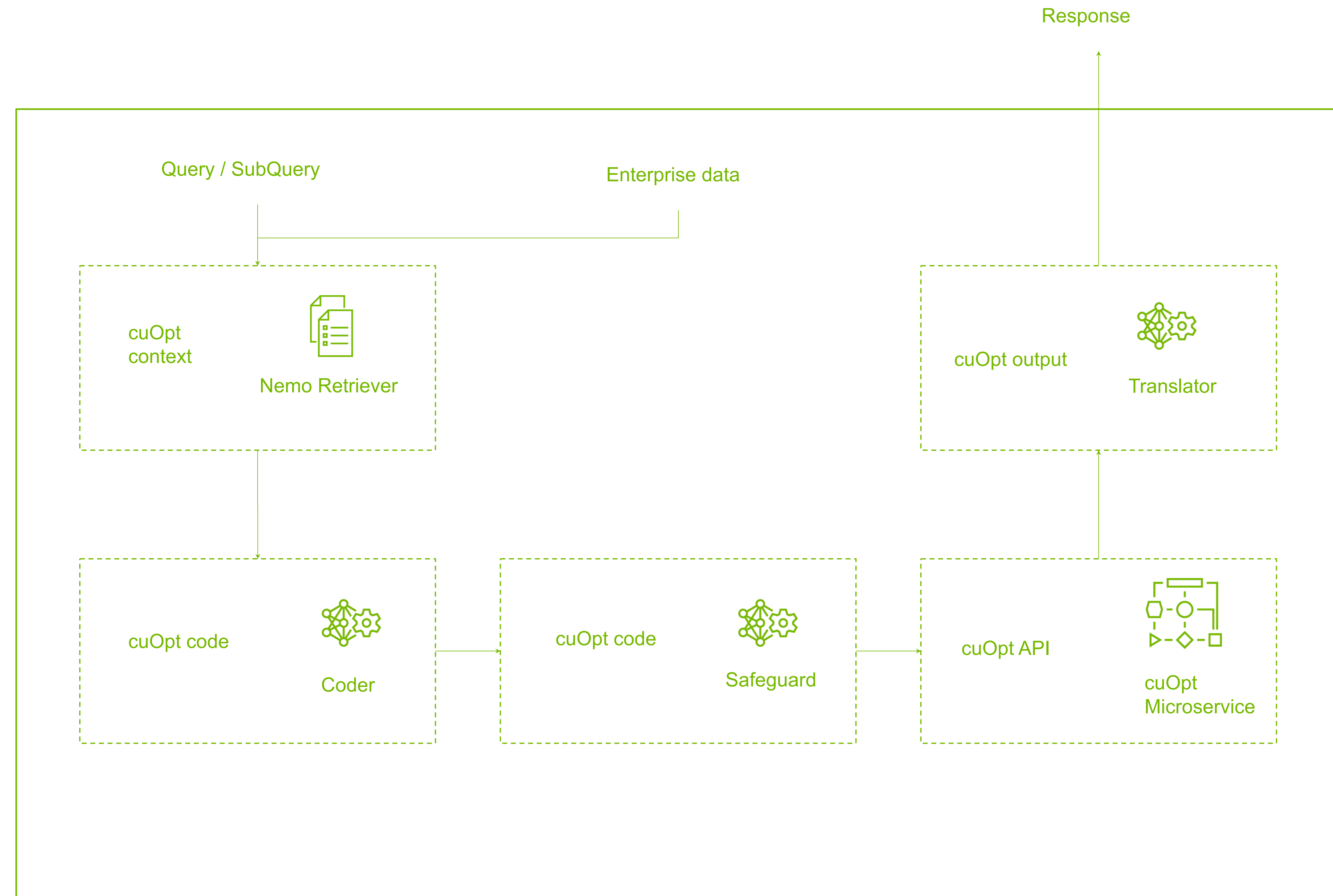


cuOpt agent concept

LLM-based framework



Nemo Retriever Zoom



cuOpt Agent microservice

Legend

- LLM Model based fine tunable Microservice
- Execution Engine Microservice
- Data Microservice (DB/VectorDB)

cuOpt agent concept

LP modeling example conversation

cuOpt OFF

cuOpt ON

User

The following NVIDIA GPUs can only be produced in batches of 500 units. One batch of RTX 4090 yields a \$90,000 profit, require a production time of 150 hours, and there is a demand for 10,000 units. One batch of RTX 4080 yields a \$60,000 profit, require 125 hours to make, and the demand is 18,000 units. There can be only one batch in the production line at a time, and the maximum time window is 200 days. How can we maximize profit?

LLM

To maximize profit, produce 20 batches of RTX 4090 and 0 batches of RTX 4080. This yields a total profit of \$1,800,000. 🚀💰

User

What if we could produce one batch of RTX 4080 in 20% less time.

LLM

If we could produce one batch of RTX 4080 in 20% less time, the optimal production strategy would be to manufacture 20 batches of RTX 4090 and 0 batches of RTX 4080. This configuration would yield a total profit of \$1,800,000. 🚀💰

User

Would it now be profitable to lower the price of the 4080 by 5% which would increase the demand by 2,000 units ?

LLM

Yes, it would be profitable to make a 5% discount on the RTX 4080, resulting in an increased demand of 2,000 units. The new profit would be \$1,140,000. 🚀💰

User

The following NVIDIA GPUs can only be produced in batches of 500 units. One batch of RTX 4090 yields a \$90,000 profit, require a production time of 150 hours, and there is a demand for 10,000 units. One batch of RTX 4080 yields a \$60,000 profit, require 125 hours to make, and the demand is 18,000 units. There can be only one batch in the production line at a time, and the maximum time window is 200 days. How can we maximize profit?

cuOpt Agent

To maximize the profit, the optimal solution will be to produce 20 batches of RTX 4090 and 14 batches of RTX 4080. This will yield a total profit of approximately \$2,640,000.

User

What if we could produce one batch of RTX 4080 in 20% less time.

cuOpt Agent

Now, it would be better to produce approximately 14 batches of RTX 4090 and 27 batches of RTX 4080. This would yield a total profit of approximately \$2,880,000. This approach would increase the profit by \$240,000.

User

Would it now be profitable to lower the price of the 4080 by 5% which would increase the demand by 2,000 units ?

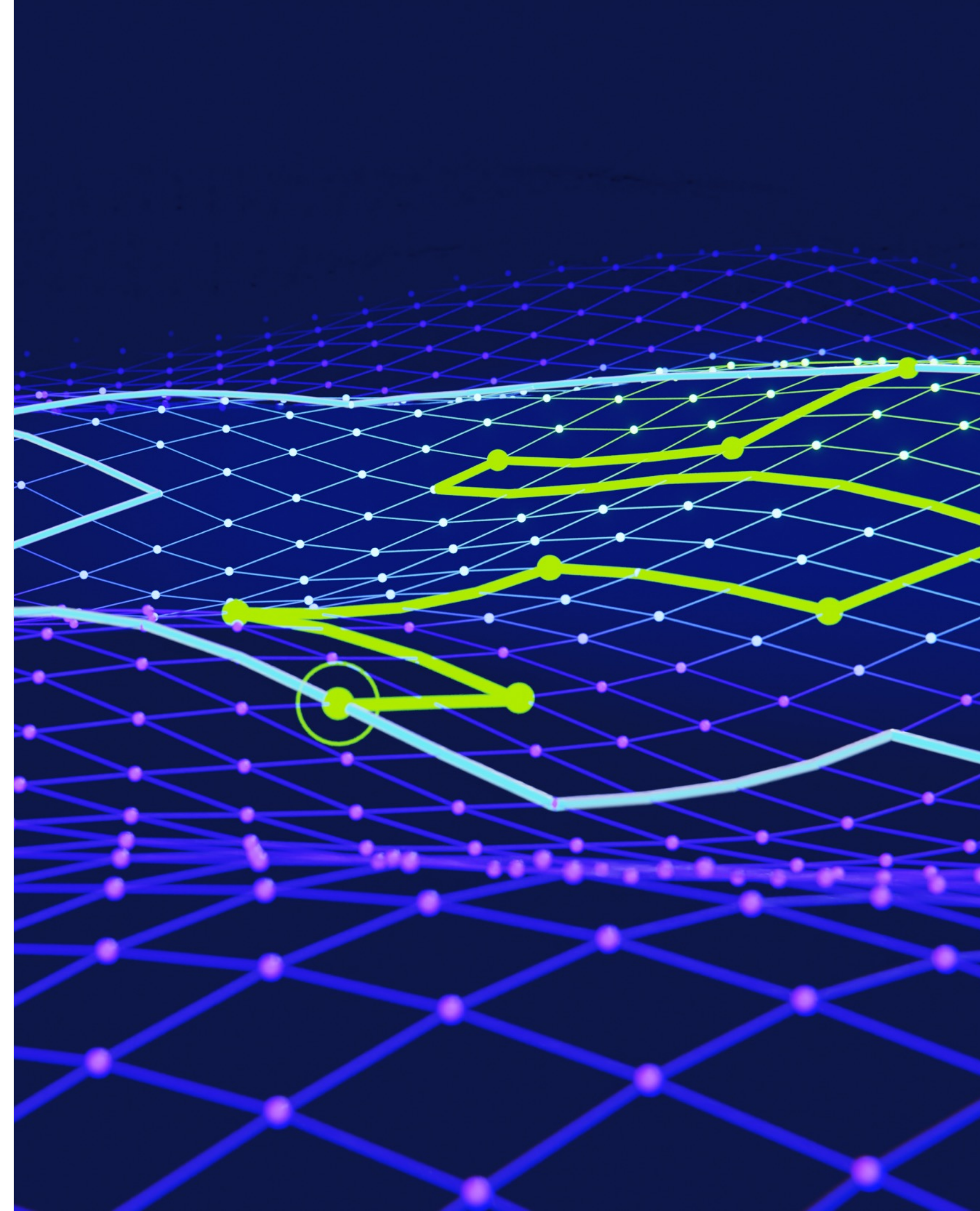
cuOpt Agent

The optimal solution would be to produce 20 batches of RTX 4090 and 18 batches of RTX 4080. This would yield a total profit of approximately \$2,826,000, which is \$54,000 lower. Hence, applying a 5% discount on RTX 4080 is not a profitable decision in this case.

Try cuOpt today!

Free option, enterprise version, managed or self-hosted

- cuOpt combines world's best accuracy, speed, and scale for multi constrained routing problems
- Early signs that GPUs accelerate mathematical optimization
- LLMs help model operations and facilitate optimization software adoption
- Play with the cuOpt service API on [API catalog](#) for free
- cuOpt is in NVIDIA AI Enterprise, deploy it as a managed service or self host





S62495 - Advances in Optimization AI

Alex Fender Ph.D. | GTC 2024

cuopt@nvidia.com