Network Optimization Thinking

Idea In Short

Improving every part of a network does not improve the network. A factory that speeds up every station still ships at the pace of its slowest one, and a supply chain that tunes each link for local efficiency can produce wild swings in orders and inventory. Network optimization thinking looks at flows through connected structures and asks where constraints, delays, and handoffs limit the whole. Eliyahu Goldratt's theory of constraints, John Sterman's studies of supply chain dynamics, and research on the Toyota Production System all point the same way. This article explains how to find the real limit, how to improve flow without creating fragility, and how to avoid optimizing the wrong thing.

A logistics firm invests in faster trucks, quicker loading, and better route software. Deliveries still arrive late, because the warehouse that sorts parcels works at a fixed pace, and all the new speed upstream piles up outside its doors. The firm optimized parts of the network and left the limit untouched. Network optimization thinking directs attention to where the limit actually sits.

Constraints set the pace

Eliyahu Goldratt popularized the idea in his 1984 novel The Goal. The theory of constraints, as an encyclopedia entry on management theory explains, begins with the goal of the organization and then sets out five iterative steps for identifying and addressing the constraints, also known as bottlenecks, that limit performance.1 The key claim is that throughput rises only when the constraint improves.

A reference entry defines a constraint as anything that prevents the system from achieving its goal, and it describes the ongoing improvement process in which efforts stay centered on the constraint.2 A constraint can be physical, such as a machine. It can also be a policy, such as an approval rule, or a market limit, such as customer demand.

A research summary from EBSCO adds a practical clue for finding it: business owners should look for areas with an excess of work in progress.3 Piles of waiting work usually sit in front of the slowest step. Following the pile upstream shows where the pace is set.

The five focusing steps

The first step is to identify the constraint. The second is to exploit it, which means getting the most output from the resource as it exists, for example by ensuring it never idles and never works on poor-quality inputs. The third is to subordinate everything else, so that other steps work at the pace that the constraint can absorb, even if this leaves them with idle time.

The fourth step is to elevate the constraint, through investment, added capacity, or redesign. The fifth is to repeat, since once the constraint is relieved, another step becomes the limit. A lean production guide notes that the process is cyclical, and that the approach is meant for continuous improvement and not for a one-time fix.4

Subordination is the hardest step culturally. Managers dislike seeing resources idle, and many performance systems reward utilization at every station. Yet running a non-constraint at full speed only increases inventory and waiting. Leaders must change the measures, so that people are rewarded for helping the whole network and not for keeping their own station busy.

Delays and the dynamics of supply chains

Networks that move goods or information over time face a second problem: delays. John Sterman explains in a chapter on supply chain dynamics that supply chains are governed mainly by negative feedback, and that because they contain substantial time delays, they are prone to cycles. He names oscillation, amplification, and phase lag as pervasive features.5

The consequence is that tuning each link on its own can make the whole worse. A retailer who orders generously when stock is low, a wholesaler who does the same, and a factory who responds to both create swings that none of them intended. Each decision is locally sensible, and the system produces waste.

Optimization therefore includes information. Share end-customer demand with every link,

show the orders already in the pipeline, and agree on simple ordering rules. These steps cost little, and they reduce the amplification that wastes capital.

Efficiency and flexibility together

Many managers believe that a network must choose between efficiency and flexibility. A study of model changeovers at the NUMMI auto plant, a Toyota subsidiary in California, challenges the belief. Its authors report that the plant used four mechanisms to support an exceptional combination of flexibility and efficiency.6

The lesson for network design is that standard routines can enable change when they are designed with the people who use them. Rigid rules imposed from above create brittleness. Rules built by the workers and updated as they learn create a network that is efficient on a normal day and adaptable on a changing one.

Slack deserves a place in the design. A network tuned to zero idle time has no capacity to absorb a surprise, and the first disruption spreads through every link. A modest buffer at the constraint protects the whole system, and it usually pays for itself.

Measures that fit a network

The measures you choose steer behavior. Three network-level measures capture most of what matters: throughput, which is the rate at which the system delivers value to customers, inventory or work in progress, which is what the system holds in the form of unfinished work, and operating expense, which is what it costs to run. Improving the first while holding the other two steady is the aim.

Station-level measures, such as machine utilization or individual output, can mislead. A team that maximizes its own output when it is not the constraint creates piles of unfinished work, longer waits, and more rework. Replace such measures with signals that show how well each step supports the flow, such as the time work waits for the step and the share of items that pass to the next step without correction.

Review the measures regularly with the people involved. Ask what behavior each one encourages, and whether that behavior helps the network. Where it does not, change the measure before you blame the people.

How to run an optimization project

Map the process from request to delivery, and note the time work spends in each step and in each queue. Most processes show far more waiting than working. Identify the step where queues are longest, and test whether it truly limits output by watching what happens when you add capacity there or remove it.

Next, improve flow at that step with low-cost changes: remove rework, schedule maintenance outside peak periods, and prevent starved or blocked states with small buffers. Align upstream release of work to the constraint's pace so that queues do not grow without limit. Measure throughput of the whole network, cycle time, and work in progress, and avoid station-level efficiency measures that encourage overproduction.

Finally, review the goal. A network optimized perfectly for the wrong goal produces the wrong results quickly. Ask whether speed, cost, quality, or reliability is what customers value, and let that answer decide where to push.

A hypothetical example

Consider a hypothetical insurer whose claims process takes thirty days. Intake, assessment, approval, and payment each look efficient on their own metrics. A map shows that claims wait an average of nine days for the assessor pool, which has the fewest people and handles every complex case.

The insurer routes simple claims to a fast lane that bypasses the pool, trains two more staff for the complex lane, and limits the intake rate to what assessors can handle. Median time falls to nineteen days, and the intake team reports less pressure as the queue shrinks. The example is invented, yet it shows how improving flow at the constraint outperformed speeding up every step.

Cautions

Optimization can erode resilience. Removing every buffer and duplicate may raise efficiency and reduce the ability to survive shocks. Keep a deliberate margin.

Optimization can also distort behavior. If you measure one metric, people will improve it at

the cost of others. Use a balanced set, and review for side effects.

Finally, remember that networks include people. A plan that treats staff as nodes will meet resistance. Involve them early, and explain how the changes serve the goals they share.

- 1Theory of constraints (Encyclopedia of Management Theory)
- 2Theory of constraints
- 3Theory of constraints (TOC)
- 4Theory of constraints
- 5Booms, busts, and beer: understanding the dynamics of supply chains
- 6Flexibility versus efficiency? A case study of model changeovers in the Toyota Production System

Summary

Network optimization thinking measures improvement at the level of the whole. The theory of constraints directs effort to the one step that limits throughput, supply chain research shows that delays and local decisions create cycles that waste capital, and the Toyota case suggests that efficiency and flexibility need not be traded against each other. Leaders should find the constraint, align other steps to its pace, share information along the chain, and keep enough slack to absorb shocks. They should also check that the goal itself is right. Map one process from request to delivery this week, mark where work waits, and test one change at the constraint before touching any other step.