DESIGN BRIEF · PART THREE OF A FOUR-PAPER SERIES

The Operating Model the Forecast Can't Break: How a Pull-Based Operating Model Frees Working Capital, Improves Service and Retires the Firefight

A perspective for Boards, CFOs, and Senior Supply Chain Leaders

algo

Why is this brief

The first paper in this series argued that the trillion-dollar inventory burden carried by manufacturers and distributors is the predictable consequence of a planning paradigm built on assumptions that no longer hold. The second showed those assumptions in motion — a forecast waterfall in which a single future week of demand was forecasted thirteen different ways before reality arrived, with a variability range of 116 percentage points between the highest and lowest projections. The diagnosis is settled — forecast error and variability are inevitable and structural. This brief lays out the design of the operating model that creates the required agility to address an unknown future.

Erik Bush
EVP – Internal Operations, Algo

1. First Principles

For three decades, supply chain planning has been dominated by Material Requirements Planning — MRP — and its variants. The reasoning is intuitive. Customers will not wait, so we must plan ahead. We forecast demand, schedule backwards from required-on dates based on published lead times, generate purchase, manufacturing, and distribution orders, and execute. The premise is so embedded in enterprise software, planning curricula, and operating practice that it is rarely questioned.

MRP rests on three assumptions, each of which is demonstrably false at the level where supply decisions are actually made.

Assumption 1: The demand is known

MRP treats the forecast as input — a number to be planned against. The first two papers in this series have established what thirty years of inventory data already show: at the SKU and item level, the forecast is not accurate, cannot be made accurate, and the more it is decomposed toward the buying unit the more error it absorbs.

Assumption 2: There is no variation.

MRP schedules backwards from lead times as if those lead times will always be met. It treats supplier deliveries, manufacturing completions, and distribution movements as deterministic. In practice, every link in the chain introduces variation — late shipments, quality holds, capacity constraints, distribution delays. MRP has no native mechanism for absorbing that variation; only for re-planning around it after the fact.

Assumption 3: The only way to ensure material availability is to make the forecast more accurate.

Because MRP is forecast-driven, the response to any service failure is to invest more in forecasting — better software, more planners, finer segmentation, longer history, machine learning, demand sensing. Three decades of those investments have failed to bend the U.S. manufacturing inventory-to-sales curve. The assumption that better forecasting is the path is the most consequential error in supply chain practice.

A first-principles redesign begins by setting these assumptions aside and asking a different question: what would we build if we accepted that demand cannot be predicted accurately at the unit level, that variation is structural, and that the goal is not forecast precision but high material availability at the lowest working-capital cost? The answer rests on three counter-principles.

> “Variation cannot be forecast away. It can only be absorbed.”

2. Three Counter-Principles

Principle 1: Pace to actual demand.

The most accurate demand signal in any supply chain is the one that has already happened. When a part is consumed off an assembly line, when a unit is sold off a shelf, when a case is picked from a warehouse, that consumption is the real demand. Replenishment paced against actual consumption uses the only demand signal that cannot be wrong — and it is the operating principle Toyota proved at industrial scale half a century ago with Kanban. The forecast may say one thing; the consumption signal says what actually happened. A pull-based operating model gives consumption the authority that MRP gives the forecast.

Principle 2: Use material, time, and capacity buffers to absorb the variation MRP pretends does not exist.

Variation is inevitable. Whether it emerges from the market or emanates from suppliers or production issues, it is an ever-present reality in supply chains. It certainly cannot be forecast away.

A more effective approach is to assume variation will be present and then build compensating mechanisms — buffers — that mitigate and minimize the damage the variation causes to the supply chain. The pull-based operating model uses three types of buffers, each sized for a specific category of variation:

- **Material buffers** at strategic stocking positions, sized by the item’s average daily usage, replenishment lead time, and demand variability. When demand spikes, the buffer absorbs the surge while replenishment catches up; when demand falls, replenishment slows naturally and no excess inventory accumulates. For Engineer-to-Order (ETO) or Make-to-Order (MTO) environments these may represent core materials and items that go into finished goods. For Make-to-Stock (MTS) environments these may represent both finished goods and a majority of the subordinate items.
- **Time buffers** sit in front of critical operations or strategic decoupling points on the shop floor. The time buffer helps to ensure that no production time is lost at the resource that constrains the entire flow. It also serves to improve promise date performance by absorbing upstream supply and production variability. A work order that arrives at the constraint inside its time buffer is not a crisis; it is the buffer doing its job. It also signals a potential need to apply continuous improvement efforts to reduce upstream variability. The status of the ongoing stream or work orders approaching the constraint provide visibility to upstream issues before they impact the constraint.

- **Capacity buffers** in production resources ensure there is adequate 'sprint capacity' to absorb upstream variation without disrupting the schedule at constraint resources. While capacity buffers may appear to be unutilized capacity, they play a critical role in ensuring that variation does not amplify through the entire production flow.

The three buffer types are designed, sized, and deployed into the operating model to effectively address the core issues of unreliable demand signals and the ineffective mechanisms conventional planning and scheduling tools provide.

Principle 3: Reposition the forecast as a stress test for the operating model.

If replenishment is driven by actual consumption and variation is absorbed by buffers, what is the forecast for? Its role becomes evaluative, not directive. The forecast feeds simulations of the operating model under expected demand: will buffer configurations hold through a seasonal ramp? Will rough-cut capacity accommodate a promotional surge? Will time buffers protect promise date performance? The output is a diagnosis of where the model is at risk and the configuration changes needed in anticipation — buffers re-calibrated, an item moved to a forward-looking demand rate, capacity expanded, a new product staged. The forecast is not used to drive committed orders; it tests whether the operating model can flex to the outcomes the business may face.

3. The Operating Model in Practice

The three principles fit together as a single operating system. Day to day, the model runs without forecast intervention. Buffers replenish automatically against actual consumption. Materials managers monitor buffer status and health, not forecast accuracy. Production scheduling pulls from real demand, not projected demand. Exceptions are buffer-status alerts — items running below their red zone, capacity buffers exhausted, time buffers consumed. Like a pilot in a commercial airliner — not flying the plane, but monitoring the conditions and performance and only intervening on an exception management basis.

Periodically, typically as part of the Demand-Driven S&OP cycle, the forecast or forecasts are fed into a simulation of the operating model under one or more future-state scenarios. The simulation surfaces projected exposures: where buffers will fail, where capacity will strain, which items will need a configuration change before the demand window arrives. In parallel, observations from the past period inform additional configuration changes for the upcoming cycle. Those configuration changes are made in anticipation.

The system then returns to consumption-paced operation until the next cycle. The two cadences — continuous execution against actual demand, and periodic model tuning against projected demand — together produce a system that is both responsive in the short run and resilient in the long.

Dimension	MRP-led model	Pull-based operating model
Demand signal	Forecast (predicted)	Actual consumption (observed)
Replenishment trigger	MRP run against forecast	Buffer status against actual draw
Variation handling	Static or manually maintained safety stock	Strategic buffers sized for demand rate and variability
Forecast role	Drives the Master Production Schedule	Stress test; informs buffer and capacity design
Planner role	Improve forecast accuracy; expedite exceptions, pull-in/push-out to address forecast volatility	Configure and tune buffers; monitor model fitness
When demand surges	Forecast revised; supply chain whipsaws, expedite spending	Buffers flex upward and absorb the surge; replenishment accelerates
When demand falls	Excess accumulates; downward revisions rarely trigger drawdown	Buffers flex downward and replenishment slows naturally; no excess builds
Working capital	Bloats over time; hard to release	Sized to actual demand; capital released progressively as the model tunes
Service performance	Bimodal: stockouts and excess simultaneously	High availability with lower carrying inventory

MRP-led planning compared with a pull-based operating model across nine operating dimensions.

"

“The shift is not a software upgrade. It is a redesign of how the supply chain operates.”

4. Design the Operating Model Deliberately

Everything the previous sections describe is deliberate. That is the point — and it is what most companies operating on MRP-led planning have not, in fact, done.

In a conventional planning organization, the operating model has usually accumulated rather than been engineered. The forecast drives replenishment because it always has. Safety stocks and safety lead times were inserted at points where service failed in the past — and have not been questioned since. Lead times, sources of variation, decoupling points, and risk-absorbing mechanisms have not been examined as a coherent whole. The result is an operating model built by accretion rather than by intention, running against assumptions no one currently remembers making.

The first step of every successful deployment of a pull-based model is not the installation of software. It is a deliberate step back — an exercise with the client's leadership team to design the operating model itself. In this step, it's critical to acknowledge where should the strategic stocking positions sit? Which production flows warrant time or capacity buffers? Where is variation entering the system, and which buffer type is best suited to absorb it? What are the actual lead times, and where does the planning system assume lead times that no longer hold?

That exercise is always revealing. Companies discover that MRP assumptions are out of date, safety stocks are sized by rule of thumb instead of recognizing the range of demand patterns that exist, and no one has a documented rationale for how the model was set up. The design work is not a preface to the "real" implementation — it is the real implementation. Everything else is instrumentation.

Once the operating model has been deliberately designed, S&OP takes on a different role. Rather than a monthly reconciliation of the forecast against reality, it becomes a continuous revision of the operating model itself. Are the buffer configurations still fit for purpose? Is the variation pattern shifting? The team's focus shifts from "what is our forecast accuracy?" to "what is our adaptation capability?" — a question the operating model can actually answer, and one that produces action rather than argument.

5. What the Design Produces

The three principles deliver the three outcomes named in this paper's title.

Working capital is freed.

Material buffers sized to actual demand and variability typically yield inventory reductions of twenty five to thirty five percent — often in line with the forecast inaccuracy companies are experiencing. The working capital previously trapped in inventory becomes available for capacity expansion, debt reduction, or growth investment. The trillion-dollar inventory base named in paper one shrinks not by a fractional percent, but by a measurable, sustained, balance-sheet-visible amount.

Service improves.

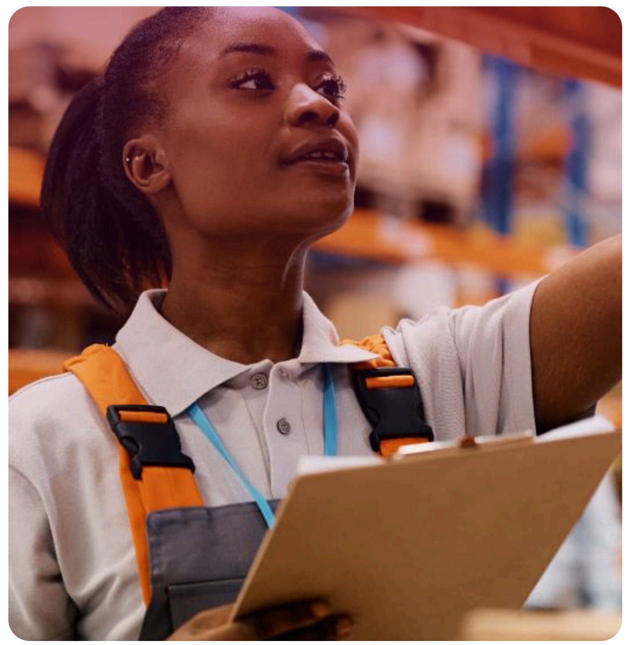
Because buffers are sized for variability rather than for forecast accuracy, items are available when consumption hits — including the items that previously stocked out because the forecast missed them. The bimodal inventory distribution that haunts MRP-led operations — too much of the wrong stuff, too little of the right — resolves.

The firefight is retired.

Planners stop chasing the forecast, stop expediting exceptions, and stop reconciling the gap between the MRP plan and reality. Their work shifts to model configuration, buffer tuning, and S&OP simulation — work that compounds in value over time rather than dissipating into the next week's exception queue. Turnover in the planning function falls. The job becomes a craft, not a treadmill.

6. What This Sets Up

The design above is necessary but not sufficient. Any senior leader who has lived through an enterprise transformation knows that the architectural diagram is the easy part; the organizational change required to operate the new design is where most efforts stall. The pull-based operating model challenges thirty years of MRP orthodoxy, threatens the comfort zone of a planning function that has been measured on forecast accuracy, and exposes the local customizations that have grown up around the old way. None of those will yield without senior leader sponsorship and direct confrontation of the resistance. That is the subject of the fourth and final paper in this series.



About this brief

This is the second paper in a four-part series on supply chain operating model design. The first paper made the strategic case for shifting from forecast-led to demand-driven planning. This brief equips senior leaders to diagnose the problem in their own data. The third paper covers the design of the demand-driven operating model; the fourth addresses deployment and the evolution of the planning function.