Євген В. Шоломій*

МОДЕЛЬ ПРИЙНЯТТЯ РІШЕНЬ НА ОСНОВІ ТЕОРІЇ МАСОВОГО ОБСЛУГОВУВАННЯ ДЛЯ ВИЗНАЧЕННЯ МОМЕНТУ ПРИДБАННЯ ДОДАТКОВОЇ ВИРОБНИЧОЇ ПОТУЖНОСТІ МАЛИМИ ВИРОБНИКАМИ ПЕРСОНАЛІЗОВАНОЇ ПРОДУКЦІЇ: ІНДЕКС МАСШТАБОВАНОСТІ ВИРОБНИЦТВА (PSI)

Малий виробник персоналізованої продукції, який уже перевів виготовлення у власні потужності, стикається з питанням, на яке література з вибору «виробляти чи купувати» відповіді не дає: у якому саме місяці слід замовити наступну одиницю обладнання. Література з розширення потужностей пропонує правила визначення моменту інвестування за умов зростаючого попиту, проте побудована для великих об'єктів із неперервно подільною потужністю; література з теорії масового обслуговування описує залежність між завантаженням і тривалістю виробничого циклу, але не дає дати інвестиції. У статті запропоновано Індекс масштабованості виробництва (Production Scalability Index, PSI) — кон'юнктивний двоплечовий індекс, що поєднує плече потужнісного тиску з плечем граничної окупності та повертає єдиний місяць-тригер для придбання (N+1)-ї одиниці обладнання у вузькому місці. Критичне завантаження не постулюється, а виводиться: обернення співвідношення типу Кінгмана відносно обіцяного строку виконання замовлення дає замкнену форму, яка для будь-якої скінченної обіцянки повертає стелю, строго нижчу за насичення, — формальне вираження правила «купуй, доки верстат ще не заповнений». На матеріалах діючого виробництва УФ-друку базовий сценарій дає критичне завантаження 0,838, безпечну стелю 838 замовлень на місяць проти ефективної потужності 1000 і припис замовити другий принтер у 3-му місяці для запуску в 6-му. Окремо обґрунтовано два результати. По-перше, необхідний місяць запуску обладнання є інваріантним щодо тривалості циклу постачання, доки остання його не перевищує, тож цикл постачання зміщує лише дату підписання договору. По-друге, кількісно визначено розрив між потребою у потужності та її фінансовою спроможністю: за темпів зростання нижче 1,07% на місяць гранична одиниця обладнання не здатна окупитися в момент вичерпання потужності, і приписом стає тимчасове повернення надлишкового потоку замовлень до зовнішнього виконавця. Модель відкалібровано на одному виробництві: її структура є переносною, значення параметрів — ні.

Ключові слова: *розширення потужностей; момент розширення; теорія масового обслуговування; наближення Кінгмана; управління тривалістю виробничого циклу; завантаження; друк на замовлення; персоналізована продукція; малі та середні підприємства; теорія обмежень.*

Формл. 13. Табл. 10. Літ. 15.

DOI: 10.32752/1993-6788-2023-2-268-136-148

Yevhen Sholomii

A QUEUEING-GROUNDED DECISION MODEL FOR TIMING THE ACQUISITION OF ADDITIONAL PRODUCTION CAPACITY IN SMALL PERSONALIZED-PRODUCT MANUFACTURERS: THE PRODUCTION SCALABILITY INDEX (PSI)

A small personalized-product manufacturer that has already internalized production faces a question the make-or-buy literature does not answer: in which month should it order the next

* National Technical University "Kharkiv Polytechnic Institute". Ukraine. ORCID ID: https://orcid.org/0009-0000-3916-5564

machine? The capacity-expansion literature supplies timing rules under growing demand, but for large facilities with continuous capacity; the queueing literature supplies the utilization-lead-time law but yields no investment date. This study proposes the Production Scalability Index (PSI), a two-armed conjunctive index combining a capacity-pressure arm with a marginal-payback arm, returning a single trigger month for acquiring the (N+1)-th machine at the bottleneck. The critical utilization is derived rather than assumed: inverting a Kingman-type lead-time law against a promised lead time yields a closed form returning a ceiling strictly below saturation for every finite promise, which is the formal statement of the rule to buy before the machine is full. A methodological finding accompanies the derivation: the scale of aggregation at which queueing bites is the working day rather than the individual order. Applied to a UV-printing operation, the base case gives a critical utilization of 0.838, a safe ceiling of 838 orders per month against an effective capacity of 1,000, and a prescription to order the second printer in month 3 for installation in month 6. Two further findings are emphasized: the required installation month is invariant to the acquisition lead time, provided that lead time does not exceed it, so the lead time shifts only the signature date; and a capacity-finance gap is quantified, in which capacity is needed before the marginal machine can repay itself. Below 1.07% monthly growth it cannot repay at the month capacity binds, and the prescription is to route overflow back to print-on-demand until it can. The model is calibrated on a single operation; its structure is transferable, its parameter values are not.

Keywords: *capacity expansion; expansion timing; queueing theory; Kingman approximation; lead-time management; utilization; print-on-demand; personalized products; small and medium enterprises; theory of constraints.*

Peer-reviewed, approved and placed: *17.10.2023.*

Problem statement and its relation to important scientific and practical tasks. A small manufacturer of personalized physical products that has completed the transition from outsourced print-on-demand (POD) fulfilment to in-house production typically owns one machine at its constraining resource. Demand then grows, and at some month the owner must decide to buy a second — a decision taken well before the capacity is needed, because equipment must be procured, shipped, installed, and staffed. In practice it is made by watching utilization and waiting for the machine to look full. This paper argues that the intuition is quantitatively wrong and supplies a computable replacement.

The reason it fails is queueing. Production lead time is not linear in utilization; it is convex and diverges as utilization approaches one, so by the time a machine is full the delivery promise on which the business depends has already been destroyed. The promise therefore implies a utilization ceiling strictly below saturation, and the correct expansion trigger is that ceiling rather than the machine's rated output. Capacity pressure alone, however, is not sufficient to justify buying: a machine that is needed may still be unable to repay its cost within the horizon the owner accepts. There the correct action is neither to buy nor to absorb the congestion, but to route the marginal order flow back to the outsourcing provider. A usable rule must therefore contain both a capacity gate and a financial gate, and must say which binds.

Analysis of recent publications. Capacity-expansion timing originates with Manne, who framed it as a trade-off between economies of scale in construction and the penalty for unserved demand under probabilistic growth [1]. Luss established the field's characteristic setting — large facilities, continuous divisible capacity, cost minimization [2] — and Van Mieghem organized its modern form around capacity sizes, types, and timing under uncertainty [3]; Chaouch and Buzacott analysed how an acquisition lead time interacts with the timing and size of a plant [4]. Ryan is the closest prior work [5]: exponential demand growth, a deterministic expansion lead time, and a rule that expands when demand reaches a fixed proportion of the capacity position. Ryan's proportion, however, comes from an exogenous shortfall-based service level, whereas here it is derived from a queueing law and a delivery promise; and Ryan has no financial gate, no integer fleet, and no outside option.

The queueing tradition supplies the mechanism. Kingman's heavy-traffic approximation gives the waiting time of a general single-server queue as a product of a variability term, a utilization term diverging as utilization approaches one, and a service-time term [6]; Whitt developed the two-moment multi-server generalization [7]; Hopp and Spearman rendered the same relation as the VUT equation and formalized effective process times [8]; Suri drew the managerial consequence that shops should deliberately operate below full utilization [9]. What this literature does not supply is an investment trigger. Palaka, Erlebacher, and Kropp are the key theoretical anchor [10]: modelling a make-to-order firm as an M/M/1 queue facing lead-time-dependent demand, they establish that utilization should be held lower when customers are more lead-time sensitive, but as a static optimum rather than a purchase date. So and Song study capacity selection undertaken to sustain a delivery-time guarantee [11], and Fisher, Gallino, and Xu supply the empirical elasticity: online sales rose on average by 1.45% per business-day reduction in delivery time [12]. Two further strands frame the decision: Dixit and Pindyck established that irreversibility plus uncertainty raises the investment trigger above the NPV-positive point [13], a conclusion departed from here because the option to defer is truncated by the acquisition lead time; and the theory of constraints supplies the conjunctive form, Watson, Blackstone, and Gardiner giving the scholarly treatment of protective capacity [14] without saying how much or when to buy it, while Song and Zhang offer a contrast in which an on-demand printer is lightly utilized yet valuable as a stockout hedge [15]. Each parent literature thus supplies one component and misses the join: we are aware of no computable, single-index trigger telling a micro-manufacturer which month to order machine N+1, jointly gated by a queueing-derived critical utilization and by the marginal machine's payback, with an explicit outside option when the two gates disagree.

Aim of the study. The aim is to construct and calibrate a computable index returning the calendar month in which a small personalized-product manufacturer should order the (N+1)-th machine at its bottleneck. Four objectives follow: to derive a lead-time-critical utilization from a queueing law and a delivery promise rather than assuming one; to combine that gate with a marginal-payback gate in weakest-link form and characterize the regime in which the two disagree; to apply the model to data from an operating UV-printing manufacturer; and to establish its structural properties under variation of the acquisition lead time.

Main results and their substantiation

Demand, utilization, and the lead-time law

Demand at the bottleneck grows deterministically, as in formula (1), where $\lambda(t)$ is monthly order volume, λ_0 the current volume, g the forecast monthly growth rate, and t the month index. With N machines each of effective capacity μ, the offered load is given by formula (2).

$$\lambda(t) = \lambda_0(1 + g)^t \tag{1}$$

$$\rho(t) = \frac{\lambda(t)}{N \cdot \mu} \tag{2}$$

Production lead time is modelled by a Kingman-type law applied at the daily work-release scale, formula (3), with W in working days, W_0 the irreducible handling and release time as load approaches zero, and V the variability term of formula (4), formed from the squared coefficients of variation of daily order inflow and of daily effective capacity [6, 8].

$$W(\rho) = W_0 + V \cdot \frac{\rho^2}{1 - \rho} \tag{3}$$

$$V = \frac{{c_a}^2 + {c_e}^2}{2} \tag{4}$$

The station is aggregated to a daily job batch: batch interarrival time is deterministic, while batch service time is the day's work content, with mean days and squared coefficient of variation equal to the sum of the two component coefficients. Substituting into Kingman's G/G/1 form yields the second term of formula (3), equivalently the reflected-Brownian workload result. The choice of scale matters: applied at the individual-order level, where effective process time is roughly ten minutes against a two-day promise, the same law returns a critical utilization near 0.99 — an artefact of matching a fine-grained variability model to a coarse-grained commercial promise. Aggregating to daily work release recovers realistic values. This is offered as a methodological observation, the effect of the modelling unit on cycle-time estimates being treated in related form in the factory-physics literature [8].

Critical utilization, derived rather than assumed

Let W* denote the promised production lead time. The lead-time-critical utilization ρ_c is defined by $W(\rho_c) = W^*$. Substituting and rearranging gives the quadratic of formula (5), with R defined by formula (6) and the admissible root given by formula (7).

$${\rho_c}^2 + R \cdot \rho_c - R = 0 \tag{5}$$

$$R = \frac{W^* - W_0}{V} \tag{6}$$

$$\rho_c = \frac{-R + \sqrt{R^2 + 4R}}{2} \tag{7}$$

For every finite W* this returns $\rho_c < 1$ strictly. That inequality is the formal statement of the rule to buy before saturation: no finite delivery promise is compatible with running the constraint at full utilization. The lead-time-safe monthly ceiling follows as formula (8).

$$\Lambda_c(N) = N \cdot \mu \cdot \rho_c \tag{8}$$

The two arms of the index

The capacity-pressure arm evaluates projected load at the month the machine would become live, normalized by the critical utilization, with L the acquisition lead time in months, as in formula (9). The payback-adequacy arm applies a simple payback ratio to the marginal asset: the incremental serviceable throughput of machine N+1 in month s is the volume it lets the firm serve within the promise that it otherwise could not, formula (10), and the arm is that flow's contribution over the payback horizon divided by the marginal capital expenditure, formula (11), where m is contribution margin per order, I_m the marginal capital expenditure, and P_{max} the maximum payback horizon.

$$\Pi(t) = \frac{\lambda(t + L)}{N \cdot \mu \cdot \rho_c} = \frac{\rho(t + L)}{\rho_c} \tag{9}$$

$$\Delta Q(s) = min(\lambda(s), (N + 1) \cdot \mu \cdot \rho_c) - min(\lambda(s), N \cdot \mu \cdot \rho_c) \tag{10}$$

$$F(t) = \frac{m \cdot \sum_{s = t+L+1}^{t+L+P_{max}} \Delta Q(s)}{I_m} \tag{11}$$

The index is conjunctive — a weakest-link rule in the theory-of-constraints tradition [14] — and the trigger is the first month it clears the threshold, formula (12). For the capacity arm alone a closed form is available, floored at the present month, formula (13).

$$PSI(t) = min(\Pi(t), F(t)), \;\; t* = min\{ t \in \mathbb{N} : PSI(t) \geq 0.85 \} \tag{12}$$

$$t^*_{cap} = max\left(0, \left\lceil \frac{ln(0.85 \cdot N \cdot \mu \cdot \rho_c / \lambda_0)}{ln(1 + g)} \right\rceil - L \right) \tag{13}$$

Decision bands and the binding arm

The index is read through four decision bands (Table 1).

Table 1. **PSI decision bands and prescribed action**

PSI(t)	Band	Action
< 0.50	MONITOR	No expansion action; re-evaluate quarterly
0.50 – 0.85	PREPARE	Obtain quotes, reserve financing, confirm the bottleneck, pre-qualify a POD overflow partner
0.85 – 1.00	COMMIT	Place the order for machine N+1 now
≥ 1.00	OVERDUE	Order immediately and deploy an interim bridge; expect lead-time overrun until the machine lands

Wherever the projected load at the live date is at or below saturation, Π is bounded above by $1/\rho c = 1.193$ in the base case; beyond saturation the station is unstable and the band OVERDUE regardless. The binding arm is reported as a model output. If $\Pi < F$ the firm is capacity-bound: the machine is needed and self-financing, so buy. If $F < \Pi$ it is finance-bound: capacity is needed but the marginal machine cannot repay within P_{max}, and the prescription is not to buy but to route overflow back to print-on-demand and extend shifts. This is the capacity-finance gap.

Two justifications support a trigger at 0.85 rather than 1.00. It tolerates the realized growth factor exceeding the forecast by $(1/0.85)^{(1/L)} - 1 = 5.57\%$ per month over the lead time — at base-case growth, a realized 7.61% against a forecast 1.93% — before the machine lands late; and it lands the machine at $\rho = 0.713$, below the crossover at which lateness cost overtakes the capital charge of earliness.

Data, parameters, and assumptions. The model is calibrated on an operating manufacturer that uses UV printing to decorate personalized consumer-electronics accessories, sells through Etsy, produces in Ukraine and ships to the United States via consolidated logistics. Table 2 lists every input with its status — canonical, observed in the author's own practice, or a model parameter. Model parameters are assumptions and are not presented as data.

Table 2. **Model inputs, values, and status**

Symbol	Value	Status	Note
μ	1,000 orders/mo	canonical	Effective capacity in shipped orders; stated as a minimum, which is conservative since higher μ defers the trigger
λ_o	700 orders/mo	model parameter	Opening volume of the illustrative trajectory, N = 1
g	1.9311%/mo	model parameter	Monthly equivalent of a 25.8% annual growth scenario; reported unrounded because F is computed from it
operators	1 per 2 machines	canonical	Machine #2 needs no new operator
scrap	7–8%	canonical	Observed defect rate; already booked into unit costs and inside μ, not re-applied to capacity
$P_{▯ax}$	12 months	model parameter	Undiscounted simple payback horizon
m	USD 16.39/order	derived	See Table 3
W_o	1.0 working day	model parameter	Irreducible handling and release time
W*	2.0 working days	model parameter	Etsy-typical 1–3 business-day window
c_a^2	0.36	model parameter	Daily inflow variability
c_e^2	0.10	model parameter	Daily effective-capacity variability
V	0.23	model parameter	Least observable input; carries the headline sensitivity
L	3 months	model parameter	Procurement, shipping, installation; training is 1–2 weeks and not binding
$I_{▯}$	USD 12,000	model parameter	Marginal printer, tooling and installation; not first-asset capital expenditure
r	15%/year	model parameter	Used only in the earliness computation

Further assumptions, revisited under Limitations: the bottleneck is the UV printer and remains so; growth is deterministic with no seasonality; orders are homogeneous in bottleneck time; each machine is an independent station under dedicated rather than pooled routing [7]; the approximation holds in steady state within each month; and reversible capacity options are excluded from the base model.

Table 3 derives the contribution margin over the representative basket mix of 40% phone cases, 25% T-shirts, 20% mugs, and 15% laptop cases. Selling prices, in-house unit costs and the logistics figure are the author's operating data. The fee line equals 10.8% of price, inside the observed blended platform take of 10.6–11% excluding Offsite Ads, which serves as an internal cross-check; the logistics figure is calibrated on the phone case and applied flat, so it overstates m and biases the trigger early.

Table 3. **Derivation of contribution margin per order, USD**

Component	Basis	Value
Blended selling price P	0.40×29.99 + 0.25×34.99 + 0.20×24.99 + 0.15×54.99	33.99
Blended in-house unit cost C	0.40×6.50 + 0.25×9.20 + 0.20×6.30 + 0.15×13.00	8.11
Platform fees	9.5% × P + 0.45 fixed	3.68
Cross-border logistics	consolidated model, per unit	5.00
Import duty	10% × declared value (C)	0.81
Contribution margin m	P – C – fees – logistics – duty	16.39

The lead-time law and the critical utilization

With V = 0.23 and R = (2.0 – 1.0)/0.23 = 4.348, the closed form returns ρc = 0.8383, giving Лc (1) = 838 orders per month against an effective capacity of 1,000. Table 4 reports lead time against utilization.

Table 4. **Production lead time as a function of bottleneck utilization (W_0 = 1.0 day, V = 0.23, W* = 2.0 days)**

Utilization ρ	Lead time W(ρ), working days	Multiple of the promise
0.60	1.21	0.60×
0.70	1.38	0.69×
0.80	1.74	0.87×
0.8383 (= ρc)	2.00	1.00×
0.90	2.86	1.43×
0.95	5.15	2.58×
0.99	23.54	11.77×

Convexity drives the result. Between ρ = 0.60 and ρ = 0.80 lead time rises by half a day; between 0.95 and 0.99 it rises by more than eighteen days. An operator who waits for the machine to be full destroys the two-day promise long before utilization reaches one: at 90% the promise is exceeded by 43%, at 95% by 158%. Because V is the least observable parameter, Table 5 reports ρc over a two-way grid of the promise and the total variability.

Table 5. **Two-way sensitivity of the critical utilization ρc to the promise W* and total variability**

W* (working days)	$c^2 = 0.10$	$c^2 = 0.20$	$c^2 = 0.46$	$c^2 = 0.60$	$c^2 = 1.00$
1.5	0.916	0.854	0.745	0.703	0.618
2.0	0.954	0.916	0.838	0.805	0.732
2.5	0.969	0.941	0.881	0.854	0.791
3.0	0.976	0.954	0.906	0.883	0.828

Across the plausible region ρc spans 0.618 to 0.976 — everywhere below 1. The exact trigger is parameter-sensitive; the qualitative rule is robust across the grid.

Fleet ceilings, base-case trajectory, and growth scenarios

Table 6. **Lead-time-safe ceilings and labour steps by fleet size (μ = 1,000; one operator per two machines)**

N	Effective saturation	Λc(N)	Operators	Expansion step
1	1,000	838	1	baseline
2	2,000	1,677	1	machine only
3	3,000	2,515	2	machine + new operator
4	4,000	3,353	2	machine only

A structural finding follows: capacity expands in steps of μ but labour in steps of 2μ, so expansions alternate between machine-only and machine-plus-operator; the second printer is a machine-only step.

Table 7. **Base-case monthly trajectory (λ_0 = 700, g = 1.9311% per month, N = 1, L = 3, I_m = USD 12,000)**

Month t	Π(t)	F(t)	PSI(t)	Band	Binding arm
0	0.884	0.419	0.419	MONITOR	finance
1	0.901	0.573	0.573	PREPARE	finance
2	0.919	0.751	0.751	PREPARE	finance
3	0.937	0.955	0.937	COMMIT	capacity
4	0.955	1.185	0.955	COMMIT	capacity
5	0.973	1.442	0.973	COMMIT	capacity
6	0.992	1.725	0.992	COMMIT	capacity
7	1.011	2.024	1.011	OVERDUE	capacity

The trigger is month 3, machine live in month 6. The interplay of the arms is the point: the capacity arm alone would have committed at month 0, but at that date the marginal machine would have returned only 42% of its cost within the payback horizon, so the finance arm delays the order by three months. From month 3 the arms switch and capacity binds.

Table 8. **Purchase timing by demand growth rate (base parameters otherwise; the g = 1.93% row is the base case)**

g per month	t* capacity arm only	t* (PSI ≥ 0.85)	t* (PSI ≥ 1.0)	Live month	ρ at order date	Binding arm	Gap, months
0.50%	1	39	41	42	0.850	finance	38
1.00%	0	15	16	18	0.813	finance	15
1.93%	0	3	7	6	0.741	capacity	3
3.00%	0	0	4	3	0.700	capacity	0
5.00%	0	0	1	3	0.700	capacity	0
8.00%	0	0	0	3	0.700	capacity	0
10.00%	0	0	0	3	0.700	capacity	0

The headline result is in the lower rows: at 70% utilization, with growth at or above 3% per month, the model prescribes ordering the second printer immediately. The upper rows carry the complement: at low growth the finance arm binds and the prescription is to defer — in the 0.50% case by over three years.

Two growth thresholds must not be conflated. The first, g_{min} = 1.07% per month, is the rate below which the marginal machine cannot repay USD 12,000 within twelve months even in the month projected load first reaches the safe ceiling (Π = 1.0); below it the purchase must be deferred past the capacity trigger and overflow routed to POD in the interim. Because the month at which Π reaches 1.0 is integer-valued, F at that month is not monotone in g, so g_{min} is reported as the rate above which F $\geq$1 for every higher growth rate, its continuous-month analogue being 1.08%. The second, g_gap = 2.36% per month, is the rate above which both arms trigger in the same month; below it the gap widens rapidly — three months at 1.93% growth, fifteen at 1.00%, thirty-eight at 0.50%.

Earliness versus lateness

The capital charge of holding the machine a month early is $I_m \cdot r/12$ = USD 150 per month. Against it stands the cost of lateness, $C_{late} = \theta \cdot (\lambda - Лс) \cdot m$, where θ is the fraction of the over-ceiling flow lost to lead-time attrition. Table 9 reports the break-even attrition at which the two costs are equal.

Table 9. **Break-even lead-time attrition by overshoot of the safe ceiling Лс(1) = 838**

Overshoot of Λc	Shortfall, orders/month	C_late at θ = 1, USD/month	Break-even θ, fraction of the over-ceiling flow
2%	16.8	275	0.546
5%	41.9	687	0.218
10%	83.8	1,374	0.109
20%	167.7	2,748	0.055

Anchoring attrition at the published elasticity of 1.45% of sales per business-day [12] gives a fraction of total sales rather than of the over-ceiling flow, so the comparison is made in USD. At ρ = 0.90 the 0.86-day overrun implies about 1.25% of sales, or USD 185 per month; at $\rho = 0.95$ the 3.15-day overrun implies about 4.57%, or USD 712. Lateness therefore overtakes earliness at and beyond utilization of 0.90 and then grows convexly, while committing at $\rho = 0.85$ lands the machine at $\rho = 0.713$, below that crossover. It is the convexity, not the level, that justifies the early bias.

Secondary sensitivities and the invariance property

Table 10. **Sensitivity of the trigger to marginal capital expenditure and to acquisition lead time (g = 1.9311% per month)**

Varied parameter	Value	Order month t*	Live month T*
I⍰, USD	8,000	1	4
I⍰, USD	12,000	3	6
I⍰, USD	20,000	5	8
L, months	1	5	6
L, months	3	3	6
L, months	6	0	6
L, months	9	0	9

The lead-time rows establish the invariance property. Because the index depends on t only through the live date $T = t + L$, the required installation month is invariant to the acquisition lead time; L shifts only the signature date. The property holds while $L \leq T^*$; for a longer lead time the trigger clamps to $t = 0$ and the machine necessarily lands late, at $T = L$, which is why the $L = 9$ row returns $T^* = 9$ rather than 6. Within the property's range, lead times of one, three, and six months all give $T^* = 6$, with order months of 5, 3, and 0. This is both an internal-validity check and the formal basis for a practitioner rule: decide when the machine must be running, then subtract the lead time.

A second shift is the cheaper first move, and the model is tested against it: with $\mu = 1{,}500$ at zero capital expenditure and a per-order labour premium, the capacity trigger moves from month 0 to month 20 at base-case growth, and to month 6 at 5% growth. A second reversible option in the same operation is a zero-capital change in the product mix at the constraint, lifting the ceiling to 1,250 orders per month, which defers the trigger to order month 14 for installation in month 17. The base case therefore presupposes that reversible options have been evaluated and the apparent constraint found genuine.

Numerical verification confirms that (t) is strictly increasing in t and F(t) non-decreasing: F is flat at zero while $\lambda(t+L+P_{max})$ lies below the safe ceiling, a region the base case never enters; strictly increasing across the stable region ($\rho < 1$, months 0–18); and flat again once $\lambda(t+L+1)$ exceeds twice the ceiling, saturating at $m \cdot P_{max} \cdot$ Лc/I_m = 13.74 from month 42. The index is therefore non-decreasing and t* a unique first crossing. Three further checks hold: $W(\rho c) = 2.000000$, the definitional identity; the closed-form capacity trigger reproduces the numerical search across the growth grid of Table 8 when floored at $t = 0$; and $\Pi(t) \leq 1/\rho c = 1.193$ wherever projected load at the live date is at or below saturation.

Interpretation and managerial implications. Three results follow. First, the critical utilization is derived: translating a two-day promise into a ceiling of 0.838 requires no target-utilization assumption, and the general result $\rho c < 1$ for all finite W* is the formal content of a doctrine that quick-response manufacturing states qualitatively [9], that the theory of constraints calls protective capacity [14], and that Palaka et al. [10] and So and Song [11] obtain as a static optimum, here converted into a purchase date. Second, the capacity-finance gap is a quantified regime rather than a corner case: it occupies the whole growth range below 2.36% per month and reaches thirty-eight months in width at 0.50%. The closest antecedent, Ryan's fixed-proportion trigger [5], has no financial gate, and that gate binds across the entire range below g_gap; the prescription there is neither obvious option, but to bridge via print-on-demand overflow. Third, the invariance result separates two questions practitioners conflate: when the machine must be running, and when the cheque must be signed. Only the second depends on the lead time, which answers the question Chaouch and Buzacott posed [4] — how far ahead of demand an acquisition lead time forces the decision — namely, L months, and no further.

Four managerial implications follow: track the ceiling rather than the rated output, since the operating limit is 838 orders per month, not 1,000; report the binding arm, because capacity-bound means buy and finance-bound means bridge to POD; plan the installation month first and subtract the lead time to obtain the signature date; and exhaust reversible options before an irreversible purchase.

Limitations

Growth is deterministic, with no within-horizon seasonality, although Etsy's fourth-quarter peak is real and large; a stochastic-growth extension in the manner of Ryan [5] is the natural next step. The variability parameters are assumed rather than measured and are the least observable inputs, which is why they carry the headline sensitivity of Table 5; W_0, W*, L, I_m and λ_0 are likewise model parameters, and Table 8 reports the trigger across the growth range so the result can be read without committing to λ_0.

Orders are assumed homogeneous in bottleneck time, and the dedicated-routing assumption is conservative, since pooling [7] would raise ρc and defer the trigger. The approximation is heavy-traffic and two-moment, so accuracy degrades away from heavy traffic, though the decision region is where it is strongest. Labour is step-fixed while the model books it per unit, and the index is undiscounted; the contribution margin is optimistic for the reason given under Table 3, so the trigger is biased early. The attrition elasticity is transferred from US omnichannel apparel retail [12] rather than POD personalized goods, which is why break-even values are reported in Table 9; and the growth rate is a scenario parameter rather than an estimate, so g should be anchored on the operation's own order history. The model is calibrated on a single operation with one bottleneck technology and one channel mix: the structure is transferable in principle, the parameter values are not.

Conclusions and prospects for further research

This study proposed the Production Scalability Index, a two-armed conjunctive model returning the month in which a small personalized-product manufacturer should order the (N+1)-th machine at its bottleneck, and reporting which gate binds. The capacity arm rests on a lead-time-critical utilization derived from a Kingman-type law and a delivery promise, yielding $\rho c < 1$ for every finite promise; the payback arm applies a simple payback ratio to the marginal asset. Applied to a UV-printing operation, the model gives a safe ceiling of 838 orders per month against an effective capacity of 1,000 and prescribes ordering the second printer in month 3 for installation in month 6 — a decision the capacity arm alone would have taken three months too early.

Four contributions are claimed: the critical utilization is derived rather than assumed; the aggregation scale at which queueing bites is shown to be the working day rather than the individual order; the capacity-finance gap is named and quantified, with its boundary at 2.36% monthly growth and a deferral region below 1.07%; and the invariance of the installation month to the acquisition lead time gives formal basis to the rule that one should decide when the machine must be running, then subtract the lead time.

Future work should replace deterministic growth with a stochastic process and seasonality, measure the variability parameters directly from work-release records, and integrate a complexity-weighted capacity measure so the homogeneous-order assumption can be relaxed. Empirical validation across firms and technologies remains necessary before the parameter values reported here are treated as anything other than one operation's calibration.

1. Manne, A. S. (1961). Capacity expansion and probabilistic growth. Econometrica, 29(4), 632–649.

2. Luss, H. (1982). Operations research and capacity expansion problems: a survey. Operations Research, 30(5), 907–947.

3. Van Mieghem, J. A. (2003). Capacity management, investment, and hedging: review and recent developments. Manufacturing & Service Operations Management, 5(4), 269–302.

4. Chaouch, B. A., & Buzacott, J. A. (1994). The effects of lead time on plant timing and size. Production and Operations Management, 3(1), 38–54.

5. Ryan, S. M. (2004). Capacity expansion for random exponential demand growth with lead times. Management Science, 50(6), 740–748.

6. Kingman, J. F. C. (1961). The single server queue in heavy traffic. Mathematical Proceedings of the Cambridge Philosophical Society, 57(4), 902–904.

7. Whitt, W. (1993). Approximations for the GI/G/m queue. Production and Operations Management, 2(2), 114–161.

8. Hopp, W. J., & Spearman, M. L. (2011). Factory Physics (3rd ed.). Long Grove, IL: Waveland Press.

9. Suri, R. (1998). Quick Response Manufacturing: A Companywide Approach to Reducing Lead Times. Portland, OR: Productivity Press.

10. Palaka, K., Erlebacher, S., & Kropp, D. H. (1998). Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. IIE Transactions, 30(2), 151–163.

11. So, K. C., & Song, J.-S. (1998). Price, delivery time guarantees and capacity selection. European Journal of Operational Research, 111(1), 28–49.

12. Fisher, M. L., Gallino, S., & Xu, J. J. (2019). The value of rapid delivery in omnichannel retailing. Journal of Marketing Research, 56(5), 732–748.

13. Dixit, A. K., & Pindyck, R. S. (1994). Investment under Uncertainty. Princeton, NJ: Princeton University Press.

14. Watson, K. J., Blackstone, J. H., & Gardiner, S. C. (2007). The evolution of a management philosophy: the theory of constraints. Journal of Operations Management, 25(2), 387–402.

15. Song, J.-S., & Zhang, Y. (2020). Stock or print? Impact of 3-D printing on spare parts logistics. Management Science, 66(9), 3860–3878.

1. Manne, A. S. (1961). Capacity expansion and probabilistic growth. Econometrica, 29(4), 632–649.

2. Luss, H. (1982). Operations research and capacity expansion problems: a survey. Operations Research, 30(5), 907–947.

3. Van Mieghem, J. A. (2003). Capacity management, investment, and hedging: review and recent developments. Manufacturing & Service Operations Management, 5(4), 269–302.

4. Chaouch, B. A., & Buzacott, J. A. (1994). The effects of lead time on plant timing and size. Production and Operations Management, 3(1), 38–54.

5. Ryan, S. M. (2004). Capacity expansion for random exponential demand growth with lead times. Management Science, 50(6), 740–748.

6. Kingman, J. F. C. (1961). The single server queue in heavy traffic. Mathematical Proceedings of the Cambridge Philosophical Society, 57(4), 902–904.

7. Whitt, W. (1993). Approximations for the GI/G/m queue. Production and Operations Management, 2(2), 114–161.

8. Hopp, W. J., & Spearman, M. L. (2011). Factory Physics (3rd ed.). Long Grove, IL: Waveland Press.

9. Suri, R. (1998). Quick Response Manufacturing: A Companywide Approach to Reducing Lead Times. Portland, OR: Productivity Press.

10. Palaka, K., Erlebacher, S., & Kropp, D. H. (1998). Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. IIE Transactions, 30(2), 151–163.

11. So, K. C., & Song, J.-S. (1998). Price, delivery time guarantees and capacity selection. European Journal of Operational Research, 111(1), 28–49.

12. Fisher, M. L., Gallino, S., & Xu, J. J. (2019). The value of rapid delivery in omnichannel retailing. Journal of Marketing Research, 56(5), 732–748.

13. Dixit, A. K., & Pindyck, R. S. (1994). Investment under Uncertainty. Princeton, NJ: Princeton University Press.

14. Watson, K. J., Blackstone, J. H., & Gardiner, S. C. (2007). The evolution of a management philosophy: the theory of constraints. Journal of Operations Management, 25(2), 387–402.

15. Song, J.-S., & Zhang, Y. (2020). Stock or print? Impact of 3-D printing on spare parts logistics. Management Science, 66(9), 3860–3878.