OPERAÇÃO • ARTIGO 06

Como identificar gargalos operacionais

Onde o processo perde tempo, controle e margem e como mapear isso na prática.

Ruan Planeja | Revisão narrativa e aplicação gerencial | Outubro de 2026

Resumo

Gargalos operacionais limitam a capacidade de uma sequência de atividades e podem gerar atrasos, filas e perdas comerciais. Este artigo utiliza revisão narrativa sobre restrições de recursos, sistemas de filas e medição de desempenho, complementada por simulações originais. Diferenciam-se capacidade, carga requerida, tempo de execução e tempo de espera. Um exemplo de processo mostra que ampliar uma etapa já suficiente não aumenta necessariamente a entrega final. Outro exemplo compara contribuição por minuto do recurso restrito. Apresenta-se um procedimento de observação, registro de causas e acompanhamento de intervenções. Conclui-se que o diagnóstico deve identificar a restrição efetiva no período e avaliar o resultado do processo completo, evitando confundir uma tarefa demorada com a causa principal dos atrasos. As aplicações têm caráter ilustrativo e dependem de condições operacionais definidas.

Palavras-chave: Gargalos; Capacidade; Processos; Produtividade.

1 Introdução

Pedidos em atraso nem sempre resultam da falta de pessoas em toda a empresa. A limitação pode estar em uma aprovação, em um equipamento, na informação enviada pelo cliente ou em um horário concentrado de demanda. Contratar ou automatizar sem localizar essa restrição pode elevar o custo e deixar o prazo de entrega praticamente igual.

O objetivo deste artigo é apresentar um método de identificação de gargalos que combine observação e medidas simples. O foco está no fluxo completo: o que chega, o que pode ser processado e o que efetivamente é entregue com qualidade.

2 Metodologia

A revisão narrativa reúne uma fonte universitária sobre recursos restritos, o resumo público do artigo de Little sobre filas e o resumo de uma revisão acadêmica sobre medição de desempenho. Os exemplos foram elaborados para este texto e não resultam de cronometragem de uma empresa. A aplicação das relações de filas exige médias compatíveis e estabilidade; o procedimento proposto não é um instrumento de diagnóstico validado experimentalmente.

3 Capacidade carga e espera

Capacidade é o volume que uma etapa consegue processar em determinado intervalo e sob condições definidas. Carga requerida é o trabalho que a demanda impõe ao recurso. Uma restrição pode limitar a produção ou a prestação de serviço; em uma sequência, o gargalo é a etapa que limita o fluxo nas condições observadas [1]. Ele pode mudar conforme produtos, horários e disponibilidade.

Tempo de execução é o período em que o trabalho está sendo realizado. Tempo de atravessamento inclui execução, espera e movimentação entre etapas. Um atendimento que requer dez minutos de trabalho pode demorar dois dias se permanece em filas de aprovação. Medir apenas o trabalho ativo deixaria de explicar o prazo percebido pelo cliente.

A utilização observada também precisa ser distinguida da carga requerida. Uma etapa não opera continuamente a 150% de sua capacidade real; quando a demanda exige esse nível, forma-se uma diferença que terá de ser absorvida por fila, atraso, perda de pedidos ou aumento efetivo de capacidade.

4 Simulação de um processo sequencial

Considere pedidos equivalentes que passam por recepção, conferência, execução e revisão. Não há retrabalho ou recurso compartilhado nesta primeira simulação. As capacidades já refletem o tempo disponível e as condições normais de cada etapa. A demanda é de 18 pedidos por hora.

Tabela 1 Capacidade do processo simulado

Etapa	Capacidade por hora	Carga requerida pela demanda
Recepção	30 pedidos	60%
Conferência	24 pedidos	75%
Execução	12 pedidos	150%
Revisão	20 pedidos	90%

Fonte da tabela: elaboração própria. A execução limita a entrega a 12 pedidos por hora. Enquanto entrarem 18 pedidos por hora e a capacidade permanecer constante, o acúmulo líquido será de seis pedidos por hora, sem considerar perdas ou abandono. Elevar a recepção de 30 para 40 não resolve essa diferença.

Se a capacidade efetiva da execução subir para 18 pedidos por hora, o processo poderá atender a demanda média nesse cenário determinístico. Na prática, operar exatamente no limite deixa pouco espaço para variações, pausas e retrabalho. O gestor deve observar a distribuição das chegadas e dos tempos, em vez de concluir que a igualdade entre duas médias elimina as filas.

5 Uma relação útil para interpretar o fluxo

A relação de Little conecta o número médio de itens no sistema, a taxa média de fluxo e o tempo médio de permanência, sob as condições estabelecidas no estudo [2]. A aplicação gerencial precisa usar o mesmo limite do processo e o mesmo período de observação para todas as medidas.

$$L = \lambda \times W$$

L é a quantidade média de itens em processo, λ é a taxa média de fluxo e W é o tempo médio de permanência. Em um fluxo estável de oito pedidos por hora, com permanência média de 45 minutos, o cálculo é 8 × 0,75 = seis pedidos em processo, em média. Esse exemplo não corresponde à situação anterior de acúmulo crescente; nela, não se deve aplicar a relação como se o sistema estivesse estável.

Uma fotografia de seis pedidos na mesa não valida esse cálculo. É preciso acompanhar médias ao longo do tempo. Mudanças de fronteira, perda de pedidos e unidades diferentes também podem tornar incompatíveis as medidas. A equação ajuda a interpretar registros, mas não identifica sozinha a causa do gargalo.

6 Contribuição por recurso restrito

Quando vários produtos disputam um mesmo recurso, a comparação deve considerar a contribuição por unidade desse recurso, e não somente o ganho por produto [1]. A análise só pode orientar a programação depois de respeitar demanda, contratos, sequência técnica e requisitos de qualidade.

Tabela 2 Comparação de dois serviços no recurso restrito

Serviço	Contribuição por pedido	Tempo no recurso	Contribuição por minuto
A	R$ 40,00	10 minutos	R$ 4,00
B	R$ 60,00	20 minutos	R$ 3,00

Fonte da tabela: elaboração própria com dados simulados. Com 240 minutos disponíveis e demanda máxima de 12 pedidos de A, uma programação de 12 A e seis B utiliza todo o tempo e produz R$ 840,00 de contribuição. Destinar todo o recurso a 12 B produz R$ 720,00. O exemplo presume divisibilidade do tempo em pedidos inteiros, ausência de preparação e inexistência de outra restrição.

7 Procedimento de identificação e intervenção

O levantamento começa pelo desenho das etapas e de suas transferências. Para cada pedido, registram-se chegada, início, conclusão, devolução e motivo de espera. A observação deve incluir períodos de pico e produtos diferentes. Capacidade e fila são calculadas por etapa, com registro de pausas e disponibilidade de recursos compartilhados.

Depois, procura-se distinguir causas. Um atraso pode resultar de informação incompleta, falta de insumo, aprovação concentrada ou excesso de retrabalho. A intervenção deve ser específica: melhorar a entrada de dados, reorganizar a agenda, reduzir interrupções ou ampliar o recurso que limita o fluxo. A decisão precisa conservar o critério de qualidade da entrega.

A medição do desempenho deve apoiar a avaliação do processo, em vez de valorizar apenas números locais [3]. Após a mudança, comparam-se vazão de pedidos concluídos, prazo, retrabalho e quantidade em espera em condições semelhantes de demanda. Um período curto ou com procura muito diferente não comprova que a ação foi eficaz.

8 Discussão e considerações finais

Os exemplos não contemplam filas complexas, variabilidade elevada, restrições simultâneas ou preparações dependentes da sequência. Nesses casos, o diagnóstico exige registros mais detalhados e, eventualmente, modelagem específica. A busca de utilização máxima em todos os postos pode aumentar itens em processo sem aumentar entregas.

Identificar gargalos significa localizar o recurso que limita o fluxo e compreender por que ele opera dessa forma. A análise de capacidade, espera e contribuição ajuda a priorizar mudanças. O critério de sucesso é o desempenho da operação completa: mais entregas adequadas, com prazos sustentáveis e uso dos recursos compatível com a demanda.

Referências

[1] OPENSTAX. Principles of Accounting Volume 2 Managerial Accounting. Houston: Rice University, 2019. Seção 10.6, decisões com recursos restritos.

Consultar fonte original | Acesso em 13 out. 2026.

[2] LITTLE, John D. C. A Proof for the Queuing Formula L = λW. Operations Research, v. 9, n. 3, p. 383-387, 1961. DOI: 10.1287/opre.9.3.383. Consulta ao resumo público.

Consultar fonte original | Acesso em 13 out. 2026.

[3] NEELY, Andy; GREGORY, Mike; PLATTS, Ken. Performance measurement system design A literature review and research agenda. International Journal of Operations & Production Management, v. 15, n. 4, p. 80-116, 1995. DOI: 10.1108/01443579510083622. Consulta ao resumo público.

Consultar fonte original | Acesso em 13 out. 2026.