Gest. Prod., São Carlos, v. 21, n. 1, p. 65-76, 2014

Produzindo para disponibilidade: uma aplicação da Teoria das Restrições em ambientes de produção para estoque

Making to availability: an application of the Theory of Constraints in make-to-stock environments

Fernando Bernardi de Souza[1]
Silvio Roberto Ignácio Pires[2]

Resumo: O objetivo deste artigo é apresentar a forma pela qual a Teoria das Restrições vem recentemente tratando a gestão de ambientes de produção para estoque. Esse novo formato traz interessantes inovações tanto em termos de uso simplificado do método Tambor-Pulmão-Corda quanto na forma de puxar a produção em ambientes que produzem de forma antecipada à demanda. Ele inova também ao criar as condições que permitam oferecer oportunidades mercadológicas baseadas na declaração explícita de garantia de disponibilidade de produtos, daí seu nome produzir para disponibilidade (*make to availability* - MTA). Embora contenha importantes contribuições à área de planejamento e controle da produção, não foram identificados até a presente data artigos em periódicos, nacionais ou internacionais, que tivessem tratado diretamente este tema.

Palavras-chave: Teoria das Restrições. Produção para estoque. Produção para disponibilidade. Tambor-Pulmão-Corda simplificado.

Abstract: *The objective of this paper is to present how the Theory of Constraints has recently been dealing with make to stock environments. This new format introduces interesting innovations in terms of the use of simplified Drum-Buffer-Rope method as well as ways to pull production in environments that produce in advance of demand. It also innovates by creating conditions for providing market opportunities based on the explicit assurance of product availability, hence its name: make to availability (MTA). Although it provides important contributions to production planning and control, articles that have directly addressed this issue have not been identified to date in both national and international journals.*

Keywords: *Theory of Constraints. Make to stock. Make to availability. Simplified Drum-Buffer-Rope.*

1 Introdução

O desenvolvimento industrial nos últimos séculos acarretou duas formas clássicas de um sistema produtivo atender aos seus clientes. A primeira a surgir foi a chamada produção sob encomenda (*Make to Order* - MTO), característica do período inicial de desenvolvimento industrial, mais marcantemente durante a chamada produção artesanal. A lógica é relativamente simples: produzir um produto conforme o que fora prometido e vendido ao cliente. Posteriormente, com o advento da chamada produção em massa para se atender à demanda das emergentes e então rotuladas "sociedades de consumo", se consolidou uma segunda forma. A lógica nesse caso é produzir para estoque (*Make to Stock* – MTS) produtos padronizados (*standard*) para atender a uma demanda futura. Vollmann et al. (2005) realçam que, desde seu início, questões como a previsão de demanda/vendas, determinação do nível dos estoques e garantia de níveis de serviço ao cliente são inerentes aos sistemas que trabalham na forma MTS.

Mais recentemente, na década de 1990, a chamada gestão por processos de negócios ganhou espaço como uma alternativa à tradicional gestão por funções praticada desde o início do século XX. Assim, produzir e vender foram redefinidos como dois processos de negócios-chave dentro do ambiente industrial, independentemente da estrutura organizacional utilizada pela empresa. Consequentemente, as duas formas clássicas de interação de um sistema produtivo com seus clientes puderam ser redefinidas de uma forma mais simples e objetiva. Portanto, quando o processo vender acontece antes do produzir, tem-se

[1] Universidade Estadual Paulista Júlio de Mesquita Filho – UNESP, Av. Engenheiro Luiz Edmundo Carrijo Coube, 1401, CEP 17033-360, Bauru, SP, Brasil, e-mail: fbernardi@feb.unesp.br

[2] Universidade Metodista de Piracicaba – UNIMEP, Rod. do Açúcar, Km 156, CEP 13400-911, Piracicaba, SP, Brasil, e-mail: sripires@unimep.br

Recebido em 12/1/2011 — Aceito em 21/7/2012

Suporte financeiro: Nenhum.

o caso MTO. Quando acontece o contrário, tem-se o MTS (PIRES, 2004), que consiste no objeto de estudo deste artigo.

Produzir para vender no futuro traz consigo algo fisicamente impossível de ser "resolvido plenamente": a imprevisibilidade do futuro. Qual variedade (mix) de produtos deve ser produzida no período de planejamento considerado? Qual o volume que deve ser produzido? Assim, na atualidade, geralmente produzir na forma MTS de forma efetiva requer sempre um árduo trabalho de gestão da demanda baseada em processos de previsões de demanda (*forecasting*), os quais fazem uso de *softwares* e algoritmos cada vez mais sofisticados. Entretanto, mesmo complementados com práticas de gestão colaborativa e de envolvimento de diversas áreas funcionais e de fornecedores (como é o caso do Planejamento de Vendas e Produção – *Sales and Operations Planning*), uma meta primária na gestão da demanda continua sendo minimizar o erro absoluto médio entre a demanda prevista e real. A isso deve ser somada a dificuldade de muitos setores industriais de manter uma série histórica de demanda que contemple também as vendas perdidas, ou seja, a demanda real e não apenas uma série histórica de vendas efetivadas.

Também, em termos de competitividade industrial, o chamado desempenho das entregas (*delivery performance*) se transformou nas últimas décadas em um forte elemento diferenciador e ganhador de pedidos. Isso requer cada vez mais fornecimentos com prazos (*lead times*) menores possíveis e cumprimento dos prazos acordados dentro das chamadas "janelas de entrega". Em muitos setores, como o automotivo e o grande varejo, o OTIF (*On Time In Full* – no prazo e completo) é atualmente um forte indicador de desempenho do fornecedor.

Adicionalmente, o custo atual do capital acentua ainda mais o tradicional *trade-off* entre nível/custo dos estoques e o nível de atendimento aos clientes. Assim, cada vez mais cresce o desafio de melhorar o nível de atendimento ao cliente (em especial na questão do desempenho das entregas) e diminuir os custos com estoques, especialmente de produtos acabados.

Na atualidade, deve-se somar a isso também novos requisitos advindos dos crescentes processos de gestão das cadeias de suprimentos, especialmente os de abastecimento rápidos e integrados aos clientes, como ocorre nas práticas de ECR (*Efficient Consumer Response* – Resposta Eficiente ao Consumidor) e VMI (*Vendor Managed Inventory* – Estoque Gerenciado pelo Fornecedor) (PIRES, 2004).

Por sua vez, os métodos Tambor-Pulmão-Corda (TPC) e Gerenciamento do Pulmão (GP) para planejamento, programação e controle da produção, baseados na Teoria das Restrições (*Theory of Constraints* – TOC) e originalmente apresentados no livro "A Meta" (GOLDRATT; COX, 1984), foram, especialmente nas três últimas décadas, bastante estudados e confrontados com outras abordagens de planejamento e controle da produção.

Na última década, ganhou destaque uma abordagem simplificada do método TPC, chamado Tambor-Pulmão-Corda Simplificado ou simplesmente TPC-S (SCHRAGENHEIM; DETTMER, 2001; SCHRAGENHEIM, 2012). O método TPC-S, assim como o seu tradicional antecessor, foi originalmente voltado para ajudar as companhias a deixar de produzir na forma MTS e satisfazer a ordens firmes de seus clientes (MTO). O pressuposto era que, em muitas situações, o fluxo de produção mais rápido, possível com esses métodos, reduziria os prazos de entrega ao ponto dessas companhias poderem produzir sob encomenda, reduzindo seus estoques de produtos acabados (SCHRAGENHEIM; DETTMER; PATTERSON, 2009). De fato, produzir sob encomenda parece ser a forma mais adequada de se produzir, especialmente dadas as atuais exigências mercadológicas, como: elevada variedade de produtos ofertados, constante pressão por novos modelos e customização e necessidade de se manter baixos níveis de estoque e reduzir custos.

Ainda que a maior quantidade de aplicações dos métodos TPC e TPC-S esteja voltada para ambientes MTO, diversas companhias, incluindo muitas de grande porte que vendem seus produtos por meio de redes varejistas, têm dificuldades em produzir sob encomenda, pois tudo o que produzem está estocado em algum lugar. Assim, ao mesmo tempo que gerenciar ambientes de produção para estoque de forma altamente competitiva passou a ser uma necessidade a ser satisfeita, por razões que serão vistas posteriormente neste artigo, o TPC clássico se manteve com dificuldades de adaptação a esse tipo de ambiente (SCHRAGENHEIM; DETTMER; PATTERSON, 2009).

Assim, a proposta deste artigo é apresentar a forma pela qual a TOC planeja e controla a produção em ambientes MTS, cujo objetivo é garantir alta disponibilidade de produtos acabados com relativos baixos níveis de estoque. Aqui, portanto, dentro do conceito geral de produzir para estoque, é introduzido um novo subgrupo: *make to availability* (MTA), ou fazer para disponibilidade. MTA é uma declaração geral do produtor no sentido de prover imediato fornecimento sempre que necessário. Geralmente, garantir disponibilidade aos clientes não é parte das políticas de empresas que produzem para estoque. Por isso, é interessante conhecer a abordagem operacional sustentada no método TPC-S necessária para oferecer tal garantia de disponibilidade e suas consequências em termos de práticas mercadológicas.

Apesar de não ser propriamente recente, pois os primeiros trabalhos e publicações sobre este tema – ainda que incompletos ou com algumas

defasagens conceituais - remontam ao início dos anos 2000, ainda são escassas as publicações em periódicos científicos internacionais ou nacionais sobre o tema. Quando encontradas, elas mencionam rapidamente sua existência, não focando uma discussão sobre o assunto.

Acrescentam-se a isso os bons resultados permitidos por esta abordagem e relatados especialmente em congressos da *Theory of Constraints International Certification Organization* – TOCICO (http://www.tocico.org/) e a presença de algumas características que podem ser consideradas como inovadoras na área de conhecimento de planejamento e controle da produção (PCP) aplicado em ambientes MTS. Assim, este artigo se justifica em seu objetivo de apresentar a abordagem de produção para disponibilidade baseada na TOC, discutindo suas técnicas e características peculiares.

Na sequência deste trabalho, são apresentados os critérios metodológicos adotados para o desenvolvimento da pesquisa e suas delimitações, são discutidas as especificidades de um ambiente MTS e como o TPC-S é aplicado com a intenção de garantir disponibilidade de produtos acabados aos clientes da operação.

2 Método e delimitação da pesquisa

Este artigo reflete uma pesquisa desenvolvida com base em uma revisão bibliográfica com o intuito de recolher informações que permitam apresentar o conhecimento atualmente disponível sobre a proposta da TOC para ambientes MTS, a qual também objetiva a eliminação de efeitos indesejáveis típicos deste tipo de ambiente, como a coexistência de excessos e faltas de itens no estoque de produtos acabados.

Desta forma, a proposta deste artigo encontra consonância com a visão de Webster e Watson (2002) sobre trabalhos fundamentados na revisão da literatura. Os autores afirmam que uma revisão eficaz da literatura cria uma base sólida para o avanço do conhecimento, facilitando o desenvolvimento da teoria e descobrindo áreas em que a pesquisa se faz necessária. Tais autores acrescentam que trabalhos baseados em revisões poderiam tratar um problema emergente que se beneficiaria da exposição de fundamentos teóricos. Neste caso, a revisão da literatura atual sobre o tema emergente seria, necessariamente, mais curta, enfatizando-se os fundamentos teóricos que potencialmente beneficiariam a solução daquele problema emergente. Neste artigo, a revisão da literatura é especialmente voltada à proposta da TOC para gestão de ambientes MTS, e menos sobre as especificidades e problemas deste tipo de ambiente.

Nesse sentido, este trabalho é metodologicamente tratado como uma pesquisa que visa uma contribuição teórica (WHETTEN, 1989), tendo como instrumento de coleta de dados a pesquisa bibliográfica.

Assim, para seu desenvolvimento, foram tomadas como fontes de dados algumas literaturas clássicas relativas à TOC e mais especificamente seu método denominado MTA. Para tanto, foram pesquisados artigos publicados em periódicos, em especial internacionais, além de alguns livros, *webcasts* e vídeos que, por serem de autoria dos criadores da técnica, têm grande importância para a condução da pesquisa. Apesar de atemporal, as bibliografias referentes ao método MTA estão restritas a períodos relativamente recentes (uma década ou menos), devido a sua contemporaneidade. O levantamento bibliográfico foi conduzido a partir de pesquisas nas bases Informs, Elsevier, Taylor & Francis, Wiley, Emerald e Scielo, fazendo uso de palavras-chave como "TPC-S" (ou "S-DBR" e "SDBR", em sua tradução para o inglês) em combinação com "MTA" ou "MTS", em áreas afins ou relacionadas com engenharia e tecnologia, negócios, gestão, operações e ciências da decisão. As palavras "Goldratt", "Schragenheim" ou "TOC" foram utilizadas como filtro quando um número muito grande de resultados não relacionados ao tema era gerado. Tais procedimentos metodológicos não permitiram encontrar artigos que abordassem diretamente a técnica objeto deste trabalho.

Também não é intenção deste artigo revisar a TOC como abordagem de gestão, tampouco caracterizar todo o espectro de seus métodos, técnicas e ferramentas. Ainda que também pouco documentada pela literatura, mesmo a técnica TPC-S, base sobre a qual a abordagem MTA se sustenta, não será detalhada neste trabalho, sendo apresentados apenas alguns de seus aspectos essenciais para a condução da pesquisa. Por fim, por seu ineditismo, optou-se aqui por não confrontar os conceitos apresentados com outros geralmente aplicados em ambientes MTS e já devidamente documentados pela literatura. Outras abordagens aqui mencionadas são utilizadas para fins exclusivos de melhor compreensão das características e peculiaridades da abordagem MTA. Um aprofundado entendimento destas abordagens está fora do escopo do artigo. O foco está estritamente em apresentar e discutir alguns aspectos técnicos que fundamentam o modo de operação MTA.

3 Produção para estoque na ótica da TOC

Fazer para estoque é uma opção tentadora para muitas empresas, especialmente para aquelas que vendem produtos padronizados. Nesses casos, tais empresas podem produzir esses produtos padrões sem antes ter recebido qualquer pedido formal, pois partem da premissa de que uma venda é altamente provável em algum momento não muito distante no futuro. Contudo, a opção de se fazer para estoque faz

com que muitas dessas empresas acabem por manter elevados níveis de estoque.

Ademais, fazer para estoque ou segundo previsões de venda tem sido um meio para se alcançar eficiências mais altas e maior nivelamento de carga, evitando a manutenção de excesso de capacidade. Consequentemente, a não ser que haja alguma urgência, essas empresas relutam em "desperdiçar as valiosas capacidades dos recursos", optando pela alternativa de se produzir sempre, nivelando a carga ao longo do ano, estocando itens nos períodos de demanda mais baixa para serem consumidos quando a demanda superar a capacidade de produção (SCHRAGENHEIM; DETTMER; PATTERSON, 2009).

Os autores argumentam que, geralmente, não há reais benefícios em se nivelar a carga sobre os recursos. Para eles, se existe suficiente capacidade protetiva ou excesso de capacidade, não faz sentido desperdiçar um ativo (capacidade do recurso) produzindo itens para os quais não há uma demanda confirmada. Por outro lado, quando bem utilizados, tais recursos poderiam permitir à empresa cumprir os prazos de entrega e responder mais rapidamente às necessidades da demanda. Ainda para esses autores, enquanto fazer para estoque é uma prática comum, muitas vezes ela se justifica em função de uma premissa incorreta, fundamentada na ideia de que recursos ociosos constituem um grande desperdício.

Existem, contudo, algumas situações que exigem que se produza para estoque. A primeira é quando há períodos de pico de venda em que o Recurso com Restrição de Capacidade (RRC) não tem capacidade de atendê-los. Se a empresa optar pela alternativa MTO, os prazos de entrega ficariam excessivamente longos durante esses períodos. Uma segunda situação na qual a opção MTS é realmente necessária e benéfica é quando o tempo de tolerância dos clientes é menor que o *lead time* de produção. Nesse caso, o sistema MTS torna-se uma necessidade para se evitar a perda de vendas, especialmente quando competidores já oferecem entregas imediatas de seus produtos (SCHRAGENHEIM; DETTMER; PATTERSON, 2009; SCHRAGENHEIM, 2010).

Antes da próxima seção tratar de detalhes técnicos inerentes ao método MTA, faz-se necessário comentar dois importantes assuntos voltados à gestão de ambientes de produção para estoque: a) datas finais e/ou intermediárias de ordens de produção e b) entendimento do papel das previsões.

Muitas empresas, incluindo aquelas que produzem para estoque, assumem que cada ordem de produção deve ter uma data de conclusão. Alguns métodos tradicionais de planejamento e controle da produção tratam uma ordem de produção para estoque da mesma forma que as ordens de produção sob encomenda, pois todas estariam baseadas em datas compromissadas de conclusão.

Para Schragenheim, Dettmer e Patterson (2009) e Schragenheim (2010), parece incongruente que sistemas de planejamento e controle da produção (PCP), como o MRPII (*Manufacturing Resource Planning*), tratem os ambientes MTS e MTO da mesma forma. Em muitos ambientes fabris, é até mesmo difícil identificar se certa ordem de produção atenderá um cliente específico ou reabastecerá o estoque. De qualquer forma, todas as ordens de produção são tratadas por esses sistemas de PCP como possuindo uma data de complementação.

Para os autores, isso não faz sentido. Quando o pedido visa atender um cliente específico, é natural que se tenha uma data de conclusão, a qual é gerada a partir de sua data de entrega. Porém, quando um pedido de produção para estoque é colocado no chão de fábrica, não se sabe em que momento algum cliente solicitará o item. Nos sistemas MRPII, por exemplo, diferentes ordens de produção resultantes de necessidades de se repor estoques e de alguns pedidos dos clientes, com datas de entrega eventualmente diferentes, são agrupadas em único pedido com a finalidade de se obter supostas maiores eficiências advindas de uma economia com preparações dos equipamentos. Porém, tal comportamento, que leva a ordens de produção com uma única data de entrega proveniente de pedidos com diferentes necessidades dos clientes, faz com que se perca o significado e o foco no atendimento ao cliente. A prioridade passa a ser atender a essas datas de relevância duvidosa, ao mesmo tempo que se buscam altas eficiências produzindo para estoque. O significado prático das datas de entrega fica ainda mais enfraquecido quando se percebe que as datas de conclusão das ordens para estoque são determinadas a partir de tempos-padrão da produção.

Uma explicação para esse comportamento de se conferir datas de complementação às ordens de produção, ainda que essas sejam para estoque ou que não reflitam exatamente os prazos individuais dos clientes, é que muitos sistemas de PCP dependem de datas de conclusão para estabelecer as datas intermediárias de produção, assim como para definir as prioridades no chão de fábrica. Uma ordem de produção sem data de conclusão não receberia qualquer prioridade da operação.

Isso também gera uma dificuldade adicional ao sistema TPC clássico, que vincula a cada ordem de produção uma data de conclusão estabelecida em função de uma data de conclusão dessa ordem. As datas de disparo da liberação de material são calculadas em função dessas datas, e suas prioridades são também calculadas em função dessas mesmas datas e dos respectivos pulmões de tempo dessas ordens (SOUZA; BAPTISTA, 2010).

Contudo, conforme discutem Schragenheim, Dettmer e Patterson (2009) e Schragenheim (2010),

ao contrário dos ambientes MTO, em sistemas de produção para estoque, as ordens de produção não deveriam vir acompanhadas de datas de conclusão, pois não há datas ou clientes reais vinculados a estas ordens.

Essa visão é compartilhada com os assim chamados sistemas puxados de produção, como o CONWIP (*Constant Work in Proceess*) e o *Kanban* que, ao contrário dos sistemas empurrados que programam o trabalho a ser executado, autorizam a produção segundo o status do sistema (HOPP; SPEARMAN, 2000). Mas não guarda semelhança com a abordagem *heijunka* muitas vezes utilizada por estes sistemas puxados que, a partir de um horizonte de pedidos ou previsão, distribui igualmente o volume e o mix de produção ao longo do tempo, tornando a produção mais nivelada e previsível (ROTHER; HARRIS, 2002).

O segundo aspecto refere-se ao uso das informações provenientes de previsões. Para Schragenheim (2002), produzir para estoque pressupõe sempre algum nível de previsão de vendas. A quantificação do "ponto de pedido", por exemplo, é resultado de uma previsão sobre as vendas brutas esperadas durante o tempo de reposição. Entretanto, mesmo com toda a sofisticação e altos níveis de estoque, indisponibilidades ou rupturas nos estoques ocorrem regularmente.

O processo de prever vendas procura responder à seguinte pergunta: quantas unidades provavelmente serão vendidas durante o próximo período? No entanto, ainda que a resposta dada por qualquer tipo de previsão não seja muito confiável, quando se entende que a demanda não pode ser conhecida de fato, a tendência é que se deixe de olhar muito à frente e que se mantenha um foco em um horizonte de tempo mais curto, quando a qualidade da previsão é significativamente melhor, mantendo um *lead time* de produção também mais curto (SCHRAGENHEIM, 2002).

Para Schragenheim (2010), a informação realmente relevante não é a demanda média prevista, mas qual poderá ser a venda no próximo período ou horizonte. Para Schragenheim (2010), o horizonte real que se deve considerar é o tempo de reposição, isto é, quanto tempo é necessário para repor aquilo que acabou de ser vendido e, dentro desse período, calcular quanto poderia ser vendido do item (demanda máxima prevista). O cálculo deve considerar não somente as flutuações na demanda, mas também as flutuações no tempo de reposição.

Schragenheim, Dettmer e Patterson (2009) apresentam algumas armadilhas em se planejar a produção baseando-se em previsões:

- Decisões baseadas em previsões podem criar simultaneamente rupturas de estoque de algumas linhas de produtos e excessos em outras. Isso é resultado do mal-entendido de se produzir segundo valores médios de uma previsão;
- Quando um estoque de segurança é deliberadamente adicionado ao consumo médio previsto e incluído no plano, sua validade não é monitorada. Assim, sempre que problemas de caixa emergem, o estoque de segurança acaba por ser reduzido devido a uma dificuldade de se justificar quão necessário ele realmente é;
- Uma vez que a prática da previsão é estabelecida, é tentador para a gerência aplicá-la por longos períodos. A motivação para isso está na intenção de garantir eficiência à produção mediante o uso de grandes lotes e de extensos acúmulos de estoque antes de períodos de picos de vendas; e
- A forma como a gerência geralmente lida com a imprecisão das previsões de longo prazo é realizando constantes "reprevisões", ou seja, atualizando ou alterando as previsões originais. Essas mudanças nas previsões originais podem causar grandes distúrbios na execução da produção e quanto mais frequentes são as atualizações nas previsões, mais instável se torna a operação. Os citados autores mencionam a advertência de W. Edwards Deming a respeito de intervir em um sistema que está, na realidade, sob controle estatístico. Esse fenômeno se manifesta, por exemplo, nas frequentes mudanças na programação da produção em resposta às mudanças observadas na média calculada, a despeito do fato de que o sistema poderia ainda estar dentro dos limites de controle.

4 Fazendo para disponibilidade: princípios e técnicas

Há cerca de uma década, a TOC vem disseminando um novo conceito dentro da ideia geral de se produzir para estoque, o qual é denominado de "fazer para disponibilidade" (MTA). MTA é um compromisso com o mercado, ou com alguns clientes específicos, para manter uma disponibilidade boa o suficiente de produtos em um armazém específico e, assim, ser capaz de entregar imediatamente, mediante pedido, todas as vezes (SCHRAGENHEIM; DETTMER; PATTERSON, 2009; SCHRAGENHEIM, 2010). Esta definição é diferente da produção para estoque (MTS) tradicional, na qual nenhum compromisso firme de disponibilidade é dado (SCHRAGENHEIM; DETTMER; PATTERSON, 2009).

Schragenheim (2010) comenta que a definição de MTA possui dois elementos críticos. Um é a mensagem de *marketing*, ao definir o mercado-alvo e os itens que estarão incluídos no compromisso de disponibilidade. O outro é o elemento operacional,

pois, uma vez que o compromisso é assumido, a produção deverá satisfazê-lo.

Schragenheim, Dettmer e Patterson (2009) estabeleceram alguns princípios básicos que devem ser seguidos como um guia na implantação da proposta MTA:

- Princípio 1: Estoque e tempo de reposição são fortemente correlacionados. Tempos curtos de reposição necessitam de estoques muito menores para assegurar disponibilidade e evitar perda de vendas. Tempos de reposição mais curtos significam também projeções mais acuradas da demanda. Esse princípio necessita ser mantido em mente quando não há formas fáceis de produzir com rapidez, como nos casos com tempos de *setup* significativos e altamente dependentes da sequência de produção. Se o *lead time* de produção é alto, a gerência deve saber conviver com níveis mais altos de estoque e com suas consequências negativas. A compreensão dessa interdependência entre estoque e tempo de reposição deve forçar a gestão a encontrar meios de minimizar a limitação;
- Princípio 2: Estoque em processo suplementa a proteção da disponibilidade. A razão por detrás deste princípio é que, ainda que o estoque em processo (*Work in Process* - WIP) não esteja instantaneamente disponível para os clientes, parte dele está "quase concluído". Portanto, uma maneira simples e efetiva para assegurar disponibilidade é manter alguma quantidade fixa de estoque, denominada de estoque alvo, na qual se somam os estoques de produto acabado e "a caminho" (no *pipeline* ou WIP). Ainda que a proporção de cada tipo de estoque (de acabados e WIP) possa variar em função das variações da demanda e do tempo de reposição, o sistema como um todo estaria estável;
- Princípio 3: Amanhã será semelhante a hoje. Esta é a base das previsões de curto prazo. Esse princípio não deveria ser verdadeiro apenas para a demanda, mas também para a combinação demanda e fornecimento. Consequentemente, a menos que se note uma indicação clara de que há uma mudança de tendência, assume-se que o estoque-alvo atual está correto para assegurar disponibilidade no curto prazo. Ao se confiar no papel do WIP (princípio 2) e na estabilidade da demanda de um dia para o outro (princípio 3), então sempre que houver um consumo de produtos acabados, uma nova ordem de produção deverá ser gerada. Desta forma, segundo a combinação desses princípios, o modo MTA de produção está baseado na ideia de produção puxada de acordo com o consumo real;
- Princípio 4: O Status de estoque de acabados dita as prioridades no chão de fábrica. Seja para liberar uma nova ordem de produção (OP) ou apressar o trabalho já liberado, o tamanho do desvio de cada item em relação aos níveis-alvo determina o nível de prioridade que a OP deve receber; e
- Princípio 5: Estagnação é indesejável. Ainda que se possa esperar que diferentes itens estejam nas zonas vermelha ou verde do pulmão (o significado de zonas do pulmão será visto oportunamente) de tempos em tempos, a "permanente residência" em qualquer das situações por muito tempo é um sinal que o nível-alvo necessita de ajuste.

Definidos seus cinco princípios, Schragenheim, Dettmer e Patterson (2009) apresentam o procedimento para operar um sistema de produção no modo MTA. Esse procedimento é formado por quatro passos:

- **Passo 1: Definir os níveis iniciais de estoque-alvo**

O nível-alvo de estoque a ser mantido em cada ponto de armazenagem do sistema (na fábrica, nos armazéns regionais, nos pontos de venda etc.) em um modo MTA deve cobrir a "demanda média durante o tempo de reposição mais um estoque de segurança", tanto para proteger adequadamente possíveis picos de demanda, quanto eventuais atrasos nas ordens de reabastecimento (SCHRAGENHEIM, 2010). Para que o resultado da aplicação desse conceito permita um nível baixo de estoque, o tempo de reposição deve ser o mais curto possível (SCHRAGENHEIM, 2002). Outra maneira de expressar esse conceito é que o estoque-alvo deve equivaler à "máxima demanda dentro de tempo de reposição médio multiplicado por um fator de incerteza" (SCHRAGENHEIM; DETTMER; PATTERSON, 2009).

Um aspecto importante neste passo é que não há necessidade de ser muito preciso na definição dos estoques-alvo. Conforme declarado no princípio 5 e detalhado na sequência, (passo 3), os níveis-alvo de estoque deverão ser dinamicamente ajustados. O que interessa neste passo é estabelecer um nível-alvo de estoque inicial bom o suficiente que, com algum grau de conservadorismo, permita baixas frequências de rupturas no início da implantação do método.

- **Passo 2: Gerar a ordem de produção**

Uma vez definido o nível-alvo de estoque, cabe à produção mantê-lo sempre constante no sistema. Isso significa que, sempre que o estoque real total de qualquer item, isto é, o estoque de produtos acabados mais as ordens de produção em aberto desse item, estiver abaixo do nível-alvo, uma nova ordem deverá ser imediatamente gerada, preferencialmente no

tamanho exato para cobrir a diferença entre o estoque total atual e o nível-alvo de estoque daquele item.

Ainda que simples e baseada no bom senso, essa ideia traz o inconveniente de gerar ordens de produção muito pequenas quando o estoque total de um determinado item no sistema estiver apenas muito pouco abaixo do seu nível-alvo. Para contornar essa situação, Schragenheim, Dettmer e Patterson (2009) sugerem dois mecanismos de geração de ordens de produção.

O primeiro mecanismo está baseado na manutenção de um lote mínimo de produção. Esse lote mínimo não está baseado em cálculos de lotes econômicos e tem como único interesse encontrar o menor lote que evite uma sobrecarga nos recursos devido a excessos de *setup*. Segundo esse mecanismo, sempre que o estoque total no sistema estiver abaixo do nível-alvo, uma ordem de produção é aberta na quantidade equivalente ao seu lote mínimo ou na quantidade referente à diferença entre o estoque-alvo e o estoque real total, predominando o valor maior. Se o maior valor corresponde ao lote mínimo, o estoque total no sistema ficará temporariamente acima do nível-alvo e uma nova ordem de produção será emitida apenas quando o estoque total no sistema voltar para um nível abaixo do estoque-alvo. Dessa forma, quando a demanda média diária de um item for inferior ao seu lote mínimo, ordens de produção não serão geradas todos os dias para esse item. Esse mecanismo, portanto, na intenção de prevenir o surgimento de gargalos no sistema, paga um preço, na forma de mais estoque do que o estritamente necessário e já estabelecido pelo estoque-alvo.

O segundo mecanismo, proposto em Schragenheim (2002), Schragenheim, Dettmer e Patterson (2009) e em Schragenheim (2010), procura dinamicamente definir os tamanhos e momentos de liberação das ordens de produção em função da Carga Planejada. A Carga Planejada equivale à carga imposta pelas ordens de produção abertas sobre o Recurso com Restrição de Capacidade (RRC).

De acordo com este segundo mecanismo, assim como o disposto no primeiro, todos os dias uma verificação deve ser feita para todos os produtos fabricados no modo MTA. Quando o estoque em mãos (na forma de itens acabados) disponível é inferior ao nível-alvo de reposição, uma verificação adicional é necessária para examinar a quantidade de estoque em processo (WIP – *work in process*) existente para esse produto. Se a soma do estoque em mãos com o WIP é inferior ao nível-alvo de reposição, deve-se abrir um pedido de reposição para esse produto. Cada pedido de reposição tem uma prioridade de liberação em cada dia. A prioridade é definida em função do status do pulmão. No modo MTA de operação, o status do pulmão para fins de liberação de uma ordem de produção de um determinado item é expresso como a razão entre a quantidade restante para o nível-alvo deste item – a qual corresponde ao nível-alvo subtraído da soma das quantidades referentes às ordens de produção abertas de reposição deste item com a quantidade em estoque de acabados deste mesmo item – dividido pelo nível-alvo de reposição (SCHRAGENHEIM; DETTMER; PATTERSON, 2009). Quanto maior o status do pulmão para liberação de uma ordem de produção, maior é a prioridade. No entanto, antes de liberar a ordem, a carga de trabalho no RRC precisa ser verificada.

De acordo com o método TPC clássico, o RRC é protegido por um pulmão de tempo. No TPC-S aplicado ao modo MTA de operação, todos os pedidos de reposição são lançados até que a assim chamada Carga Planejada sobre o RRC se aproxime de um limite de tempo equivalente ao pulmão de recurso do RRC do TPC clássico. Quando o RRC é carregado em níveis iguais ou superiores a isso, então o algoritmo poderá atrasar a liberação de material até que a Carga Planejada sobre ele volte aos níveis inferiores ao limite estabelecido. Schragenheim, Dettmer e Patterson (2009) sugerem que a carga máxima aceitável sobre o RRC fique sempre abaixo de 80% do tempo médio de reposição. Assim, se o tempo de reposição é de 5 dias, a Carga Planejada sobre o RRC não deverá ultrapassar 4 dias de trabalho desse recurso.

Conforme novos pedidos são aceitos a cada dia, os pedidos de reposição que aguardam liberação, devido às suas prioridades relativamente baixas e à alta carga no RRC, terão suas prioridades aumentadas no dia seguinte (mais estoque para reabastecer devido às vendas adicionais). Dessa forma, este procedimento ajusta o lote de produção de acordo com a carga no RRC. Em períodos com poucas vendas, a carga sobre o RRC fica baixa, permitindo que a maioria das ordens de reabastecimento seja liberada no dia seguinte ao consumo, ou seja, com lotes relativamente pequenos. Quando a Carga Planejada sobre o RRC é alta, diversos pedidos de reposição concorreriam entre si para serem liberados, fazendo com que alguns pedidos tenham suas liberações autorizadas alguns dias depois, aumentando dessa forma o lote de produção. Desse modo, o sistema fica estável, enquanto o RRC mantém ainda capacidade suficiente para lidar com a demanda do mercado.

Schragenheim, Dettmer e Patterson (2009) sugerem que, especialmente em ambientes em que o consumo de capacidade devido à troca de ferramentas é considerável, se faça uso de ambos os mecanismos, mantendo alguns lotes mínimos ao mesmo tempo que se monitora a liberação de ordens de produção em função da carga sobre o RRC.

- **Passo 3: Gerenciar o pulmão**

O gerenciamento do pulmão (GP) é um mecanismo de controle da produção. A ideia por trás desse método pode ser resumida à identificação de situações em que a proteção prevista está quase esgotada.

Depois de identificadas tais situações, um aviso é emitido, resultando em alta prioridade para as ordens problemáticas e, em seguida, faz-se uso do resto de proteção no pulmão para garantir que a perturbação local não impactará o desempenho do sistema como um todo. No TPC (e no TPC-S) aplicado a ambientes MTO, os pulmões são todos dimensionados na forma de tempo. Nesse caso, esgotar os pulmões significa chegar muito perto da data crítica (data de embarque do pedido, por exemplo). Em ambientes MTA, a principal proteção é o estoque de produtos acabados disponíveis (estoque alvo anteriormente apresentado). Nesse caso, esgotar os pulmões significa ter estoque em mãos (estoque de produtos acabados disponíveis para entrega) em níveis muito baixos, de modo que pudesse ser esgotado antes que qualquer ordem de reposição chegasse. Para Schragenheim, Dettmer e Patterson (2009), o estoque-alvo ou pulmão deve ser dividido em zonas e cada zona corresponde a uma cor específica. O seguinte esquema de cores é definido:

- Verde: Estoque de acabados é 2/3 ou mais do nível-alvo;
- Amarelo: Estoque de acabados está entre 2/3 e 1/3 do nível-alvo. OP amarelas recebem prioridade maior que as verdes; e
- Vermelho: Estoque de acabados é menor ou igual a 1/3 do nível-alvo. Tem prioridade sobre as OPs amarelas. Risco de ruptura de estoque é aumentado. Ação imediata é necessária para restaurar o estoque de acabados para os níveis amarelo ou verde. É também chamado de nível de emergência.

O tamanho do nível de emergência deve obedecer a dois critérios: 1) deve haver um tempo suficiente para acelerar o fluxo do estoque em processo e, assim, alcançar o estoque de produtos acabados a tempo. Se este não for o caso, o nível de emergência deverá ser maior, o que pode, também, ter um impacto semelhante no nível-alvo de reposição; 2) a frequência de cruzamentos do nível de emergência não poderá ser alta nem muito rara. Se isso não ocorrer, o nível de reposição deve ser alterado.

As cores têm uma correspondência direta com o conceito de status do pulmão, conforme visto. Para fins práticos, a cor verde corresponde a um status entre 0% e 33%; a cor amarela entre 33% e 67%; e a cor vermelha corresponde a um status entre 67% e 100%.

Um status do pulmão de 70% significa que o estoque em mãos deste item está em 30% do nível de reposição, sua cor é vermelha e, neste caso, o nível de emergência terá sido atingido. O status do pulmão para cada produto determina, dessa forma, as prioridades no chão de fábrica. Assim, por exemplo, quando duas ordens de produção, referentes à reposição de produtos no modo de operação MTA, aguardarem pela disponibilidade de um mesmo centro de trabalho, sendo uma ordem relativa a um produto com um status do pulmão de 36% e a outra com um status de 50%, esta última deverá ter prioridade sobre a primeira (SCHRAGENHEIM, 2002).

Vale registrar que o nível de prioridade de uma ordem de produção deverá considerar, se for o caso, a existência de ordens de produção mais antigas e abertas de um mesmo produto. Por exemplo, se o nível-alvo de determinado item é 800 unidades, há atualmente em mãos - disponível para entrega - 350 unidades deste item e há uma ordem já aberta, mais antiga e ainda não concluída deste item referente à produção de 250 unidades, então o status do pulmão de uma ordem de produção mais nova é de (800-350-250)/800, ou 25%, e sua cor é verde. A ordem de produção mais antiga (de 250 unidades), por outro lado, tem a sua frente apenas o estoque em mãos disponível e seu status é (800-350)/800, ou 56%, e sua cor é amarela.

Além de identificar as prioridades no chão de fábrica, o GP é uma importante fonte de informação para fomentar um processo de melhoria contínua. Goldratt (2009c) menciona que um elemento fundamental na gestão da cadeia de suprimentos é estabelecer um processo focalizado de balanceamento do fluxo.

O GP pode apoiar neste processo não apenas identificando o que é prioritário para ser produzido no curtíssimo prazo, mas também permitindo uma perspectiva de longo prazo, provendo um mecanismo focalizado para identificar áreas que devem ser melhoradas, pois estão bloqueando o fluxo. Para tanto, no modo de operação MTA, o GP permite que se identifiquem as ordens de produção que, por alguma razão, sofreram com atrasos e, por consequência, alcançaram a cor vermelha. Sob o ponto de vista do fluxo, isso pode ser devido a um atraso na liberação da ordem (falta de matéria-prima ou elevada Carga Planejada sobre o RRC) ou a um fluxo muito lento na fábrica que fez com que o estoque em mãos penetrasse no vermelho (SCHRAGENHEIM; DETTMER; PATTERSON, 2009).

Goldratt (2009b) sugere que se registre qualquer atraso considerado muito longo, definido como aquele cuja duração corresponda a um décimo do tempo de reposição. Falhas em etapas do processo que provocaram simultaneamente atraso desta magnitude e penetração na região vermelha do pulmão deveriam ser candidatas a compor a lista de Pareto que definirá as ações de melhoria voltadas ao balanceamento do fluxo de produção.

- **Passo 4: Manter os corretos níveis de estoque-alvo**

Enquanto os passos anteriores estabelecem um mecanismo de reposição rápida dos estoques a partir das vendas reais e identificação de prioridades no chão de fábrica em função do status do pulmão, o quarto passo visa obter um correto *feedback* para o estágio de planejamento, mais especificamente no que concerne à determinação do estoque-alvo. O

objetivo aqui é definir e obter sinais que indiquem se a estimativa inicial do nível-alvo de certo produto continua ainda válida, dadas eventuais mudanças combinadas no seu tempo de reposição (*lead time* de produção) ou em sua demanda. Tais sinais poderiam indicar a necessidade de se reduzir ou aumentar o estoque-alvo.

O sinal mais óbvio que o pulmão – ou nível-alvo – é muito grande é que frequentemente e por muito tempo o nível de estoque está na região verde. Isso significa que a combinação fornecimento e demanda não exige um estoque-alvo tão grande. Tal situação deveria ser evitada, menos devido aos custos financeiros relativos à manutenção deste estoque, mas principalmente porque se está repondo quando não há uma necessidade real, consumindo capacidade produtiva geralmente escassa em períodos de pico nas vendas. O pulmão é considerado muito alto e deveria ser reduzido quando, durante um período equivalente ao dobro do tempo de reposição – chamado de período de checagem verde – o estoque de acabados está na região verde (status acima de 67%) (SCHRAGENHEIM, 2010). Goldratt (2009b) sugere que a redução deva ser na ordem de uma região do pulmão, isto é, 33% do nível-alvo. Já Schragenheim (2010) sugere uma redução mais modesta, na ordem de 15%. Depois da redução do nível-alvo, é natural que o nível de estoque em mãos supere o próprio novo nível-alvo. Para Schragenheim (2010) e Goldratt (2009b), nenhuma checagem e nenhuma decisão sobre novas reduções no nível alvo deveriam ser consideradas até que o estoque real esteja abaixo do nível-alvo.

Muitas invasões na região vermelha, por outro lado, sinalizam que o nível-alvo pode ser muito baixo. Para estabelecer quais sinais identificariam uma necessidade de aumento do nível-alvo, Schragenheim (2010) sugere que seja feita uma combinação entre tempo gasto na região vermelha e profundidade de invasão nesta região. O algoritmo estabelece que sempre que houver uma penetração na região vermelha, a profundidade de invasão seja registrada e expressa em número de unidades do item abaixo do nível vermelho. Se dentro de um período equivalente ao tempo de reposição a soma de todas as invasões diariamente registradas for igual ou maior que o tamanho atual da região vermelha, então a recomendação é que o nível-alvo seja aumentado. Novamente, para Goldratt (2009b), o aumento deve ser na ordem de 33%, e Schragenheim (2010) sugere que algo em torno de 20% é mais apropriado.

A justificativa dada por Schragenheim (2010) é que, ao contrário de pontos isolados de uma rede distribuição, como em uma loja, por exemplo, no nível da fábrica as flutuações nas vendas são bem menos acentuadas e tais valores seriam suficientes para corresponder às novas tendências. Outro ponto importante é que, após o aumento do nível-alvo, uma ordem de produção será liberada com tamanho equivalente ao consumo real da demanda que necessita ser reposta mais a quantidade aumentada do próprio nível-alvo. Se a carga sobre o RRC já estiver alta, então a geração de grandes ordens de produção devido a aumentos no nível-alvo poderá elevar ainda mais a carga sobre o RRC, aumentando o tempo de reposição desse e de outros itens e provocando novas necessidades de aumento de níveis alvo de estoque. Se este ciclo vicioso não for interrompido, o consumo da capacidade de produção inviabilizará a promessa de disponibilidade imediata de entrega do modo MTA de operação. Portanto, aumentos mais modestos e, às vezes, divididos em mais de uma parcela podem ser necessários.

Ainda no que se refere às eventuais necessidades de aumento em níveis-alvo de estoque, deve-se ter em mente que, após o aumento no nível-alvo, o item específico estará definitivamente na região vermelha. O aumento no nível-alvo causará a liberação de uma nova ordem de reposição, que apenas após um tempo de reposição se converterá em produtos acabados. Durante este período – chamado pelo algoritmo de período de congelamento – nenhum registro de invasões na região vermelha deve ser feito.

Este tipo de análise é uma forma grosseira de se fazer previsão, mas leva em conta todos os parâmetros que impactam o nível-alvo de estoque: a demanda média de mercado, o tempo médio de reposição e o nível de flutuação de ambos. Por isso, é o sensor final para a validade dos parâmetros de planejamento (SCHRAGENHEIM, 2002).

5 A gestão de capacidade em ambientes MTA

Um gargalo emergente, especialmente em um ambiente com compromisso de manter a disponibilidade, poderia prejudicar significativamente o desempenho de toda a fábrica. Em um ambiente MTO, em que já se conhecem os pedidos dos clientes e as datas em que devem ser entregues, o controle da demanda é mais facilmente realizado (para detalhes, recomenda-se a leitura de Souza e Baptista, 2010 e Lee et al., 2010). Em um ambiente MTA, contudo, o controle da demanda é mais complexo, porque nem todas as necessidades de produção, que deveriam ser parte da Carga Planejada, estão nela incluídas.

Como sugerido anteriormente (Passo 2, seção 4), o fluxo de liberação de ordens é monitorado e controlado pela Carga Planejada. Porém, a Carga Planejada, como uma ferramenta geral de gestão de capacidade, não é suficiente, pois não inclui as ordens que "deveriam ter sido" liberadas, e que no momento tiveram sua liberação bloqueada (SCHRAGENHEIM; DETTMER; PATTERSON, 2009; SCHRAGENHEIM, 2010).

A sugestão, segundo estes autores, é manter uma carga planejada que inclua não apenas as ordens de reposição já liberadas, mas todas as necessidades (ou sugestões) de liberação para produção. Em outras palavras, para cada produto estocado, deve-se incluir uma ordem "falsa" (*dummy*) com a quantidade necessária total de reposição. Essa "Carga Planejada Total" representa a carga real sobre o RRC.

Quando a "Carga Planejada Total" estiver com uma tendência de crescimento, levando a um tempo de reposição maior que o tolerado, a administração deverá tomar ações no sentido de adicionar capacidade ou conter a demanda. Para se identificar antecipadamente o momento crítico em que ações efetivas necessitam ser tomadas, sugere-se acompanhar o aumento da carga planejada real e atuar com segurança. Quando a carga planejada total aproximar-se de 80% do maior tempo de reposição tolerável, ações de gestão da capacidade e da demanda deverão ser tomadas (SCHRAGENHEIM, 2010).

Manter ao menos 20% de excesso de capacidade, portanto, permite uma dupla vantagem em termos de maximização da disponibilidade de produtos: 1) menores níveis de utilização de equipamentos críticos reduzem o *lead time* de produção, reduzindo o tempo de reposição ou de resposta do sistema às variações reais da demanda. Essa ideia de manutenção de pelo menos 20% de capacidade ociosa é compartilhada por Suri (1998), em sua proposta da técnica QRM (*Quick Response Manufacturing*); 2) se a capacidade protetiva cair para menos de 10%, o tempo de reabastecimento aumentará fortemente, o que requer aumentar os estoques-alvo e reduzir, assim, ainda mais, a capacidade protetiva. Como consequência, o sistema estaria fortemente sujeito a entrar em uma espiral decrescente de desempenho de entrega (SCHRAGENHEIM, 2010).

Quando as vendas crescem continuamente, e a expectativa das empresas que a adotam é que isso de fato ocorra como resultado da oferta de disponibilidade imediata de produtos a alguns clientes, a carga na fábrica cresce continuamente. Se nenhuma medida preventiva for tomada, a capacidade protetiva irá diminuir até chegar a um nível perigoso. Em uma situação de vendas em expansão, quando a capacidade protetiva cair para um nível abaixo de 20%, um alerta deve ser dado. Quando cair para níveis próximos de 10%, qualquer expansão nas vendas de ofertas MTA é congelada (SCHRAGENHEIM; DETTMER; PATTERSON, 2009).

Schragenheim (2010) sugere uma forma complementar de gestão da capacidade em ambientes MTA, denominada de Pulmão de Capacidade, que é um meio rápido de se adquirir capacidade adicional, quando necessário. O que tipifica este tipo de capacidade é que os custos aumentam toda vez que é utilizado. Ele funciona como um pulmão para proteger a habilidade da companhia em se comprometer com a oferta de disponibilidade e realmente cumprir com o compromisso. Horas extras, turnos extras ou serviços de terceiros podem ser utilizados como Pulmões de Capacidade. A empresa pode construir alguns mecanismos de planejamento e gestão desse tipo de capacidade de forma que poderá estar disponível em um curto intervalo de tempo.

O uso do Pulmão de Capacidade deverá ser disparado quando a Carga Planejada Total estiver acima do admitido (acima de 80% do tempo de reposição tolerado) e/ou o número de ordens vermelhas estiver aumentando e se aproximando de 20% do total de ordens abertas. Assim como o pulmão de estoque, o Pulmão de Capacidade poderá ser também gerenciado segundo as cores verde, amarela e vermelha. O Pulmão de Capacidade deverá estar no verde (dois terços ou mais do Pulmão de Capacidade disponível) a maior parte do tempo. Se estiver regularmente no amarelo, então sua função como pulmão estará comprometida em certo grau e, se estiver na cor vermelha, medidas de contenção da oferta de MTA deverão ser tomadas imediatamente, além de ser um sinal de que investimentos permanentes de capacidade poderão se justificar.

Uma última abordagem, porém não menos importante, necessita ser mencionada quando o assunto é gestão de capacidade em ambientes de oferta de disponibilidade. Essa abordagem está baseada na exploração da capacidade protetiva estrategicamente mantida para garantir a operacionalização de um sistema de produção MTA e exige um efetivo e estreito sincronismo entre os setores de produção e de vendas.

Como relatado, um dos elementos cruciais requeridos em qualquer ambiente MTA é manter um nível suficiente de capacidade protetiva. Isso significa que os centros de trabalho não estarão autorizados a operar 100% do tempo. Na verdade, o recurso mais carregado (RRC) deverá estar ocioso, em média, cerca de 20% do tempo. No entanto, normalmente, as pessoas se sentem muito incomodadas com a ideia de capacidade ociosa. Isso é especialmente verdadeiro em ambientes que utilizam equipamentos caros ou em ambientes em que a ativação total dos recursos é teoricamente possível, pois existem produtos com valor de mercado em diversos estágios intermediários do processo de produção. Exemplos de tais ambientes são as siderúrgicas (GOLDRATT, 2009a).

Para esse autor, a relutância em manter capacidade ociosa poderá levar as empresas a desistirem de operar segundo a lógica MTA, especialmente ao se considerar que, em uma fase inicial de implementação, as empresas não estarão totalmente cientes de que tal modo de produção ajudará na obtenção de mais vendas.

Para que uma quantidade adequada de capacidade protetiva seja mantida e, simultaneamente, a necessidade de se manter recursos ociosos seja minimizada, Goldratt (2009a) recomenda gerar ordens de produção de produtos para as quais não haja qualquer compromisso

de disponibilidade ou de data de entrega. Nesse tipo de venda, deve-se vender o que há no estoque de produtos acabados e não gerar uma ordem para atender a um pedido ou um cliente específico.

Tal decisão permitirá a utilização da capacidade protetiva para esses tipos de ordens de produção. Essas ordens deverão ter seu fluxo interrompido a qualquer momento e em qualquer estágio do processo produtivo, para dar lugar às ordens regulares MTA. Dessa forma, aquelas ordens, não prioritárias, seriam "transparentes" aos olhos de qualquer outra ordem MTA, deixando de causar qualquer tipo de perturbação às ordens MTA - como se não estivesse utilizando, de fato, a capacidade protetiva necessária (GOLDRATT, 2009a).

A sugestão do citado autor é dar para essas ordens não prioritárias uma cor diferente, como a azul-claro, que denote ausência de prioridade. O pessoal do chão de fábrica deverá ser orientado para trabalhar nas ordens azul-claras somente quando não houver outro trabalho regular aguardando processamento, e parar imediatamente de processá-las sempre que uma ordem MTA estiver disponível.

Goldratt (2009a) também faz algumas outras recomendações em relação a esse procedimento. Caso um produto seja ao mesmo tempo MTA e azul-claro, deverão ser designados a ele nomes e códigos diferentes, de acordo com os diferentes usos. Para evitar erros de interpretação, às ordens azul-claras nunca será dada prioridade em relação a qualquer cor regular, mesmo quando seu estoque estiver totalmente esgotado. A atitude de vendas deverá ser: "ou nós temos isso no armazém de produto acabado - ou nós não podemos aceitar o pedido".

6 Considerações finais

Existem algumas importantes diferenças entre produzir sob encomenda e produzir para estoque, como, por exemplo, o fato de que produzir em antecipação à demanda envolve maiores riscos. Porém, alguns sistemas não fazem algumas necessárias distinções na forma de planejar e controlar a produção desses dois tipos de ambientes.

Quando a abordagem TPC foi inicialmente desenvolvida no início dos anos 1980 (GOLDRATT; COX, 1984), ela não desafiava a premissa que sustentava não haver diferenças na forma de se fazer PCP em ambientes baseados em ordens firmes ou em ordens antecipadas de uma demanda futura, tratando cada ordem de produção, em ambos os casos, como possuindo uma data final de conclusão. Da mesma forma, o GP não diferenciava uma ordem MTO de uma MTS (SCHRAGENHEIM, 2012) . Infelizmente, a maioria dos trabalhos científicos que abordam o método TPC desconhece, ainda hoje, a existência de sua forma simplificada (TPC-S) e afirma que os pulmões para a TOC devem sempre ser estabelecidos na forma de tempo. Se isso é válido para ambientes MTO, deixou de sê-lo com o modo MTA de produção.

Este artigo mostrou a necessidade de se ressaltar essas diferenças, mostrando suas decorrências e alguns detalhes operacionais de uma nova forma de gerenciar ambientes MTS, voltada a garantir, com elevada probabilidade, a disponibilidade imediata de seus produtos. Das discussões anteriores, pode-se concluir que a abordagem MTA possui alguns elementos que guardam semelhanças com outros sistemas de gestão de operações, mas também importantes diferenças, como as que seguem:

- Ao controlar a liberação de ordens de produção com base na carga de trabalho sobre RRC, o sistema MTA está em consonância com a abordagem Controle de Carga (*Workload Control – WLC*). Porém, além deste controle de carga ser feito de uma maneira bastante particular, é aqui aplicada em ambientes que produzem para estoque, enquanto a abordagem WLC é especialmente projetada para atender às especificidades de ambientes MTO (THURER; GODINHO FILHO, 2012);
- Ao produzir em resposta ao consumo real, o modo MTA de produção se assemelha aos sistemas puxados, como aqueles baseados na manufatura enxuta ou no sistema CONWIP. Contudo, o uso de controles de carga de trabalho, do status do pulmão para definir prioridades no chão de fábrica, do gerenciamento dinâmico dos pulmões para evolutivamente ajustar os níveis alvo de estoque, entre outras especificidades operacionais, e suas implicações estratégico-mercadológicas, faz da abordagem MTA uma forma bastante particular de puxar a produção; e
- Pelas razões já expostas, se o fundamento do modo MTA é responder rápido ao consumo real, parece haver importantes elementos que o distinguem das técnicas geralmente enquadradas como QRM.

O modo MTA de produção traz também importantes implicações em termos de oportunidades de mercado, pois permite adicionar mais valor a seus clientes de uma forma que seus competidores poderiam ter dificuldades em emular. Oferecer e ter condições de garantir entrega imediata de seus produtos acabados permitem que, do tempo de reposição aos clientes – como distribuidores, atacadistas e varejistas –, seja eliminado o *lead time* de produção, possibilitando a eles níveis significativamente mais baixos de estoque, incluindo reduções significativas na parcela referente ao estoque de segurança. Outra oportunidade relevante e particularmente permitida a uma empresa praticante da abordagem MTA é que ela pode oferecer a seus clientes – especialmente àqueles de grande porte – a responsabilidade pela gestão de seus produtos na planta dos clientes, como é o caso

da prática de VMI. Ainda que essa já seja uma prática consolidada na Gestão da Cadeia de Suprimentos, o modo MTA abre uma nova e particular oportunidade de executá-la (SCHRAGENHEIM; DETTMER; PATTERSON, 2009).

Outra consequência esperada do uso da abordagem MTA é que alguns efeitos indesejáveis típicos de ambientes MTS podem ser eliminados, como constantes excessos de estoques (que reduzem o espaço físico disponível, limitam o fluxo de caixa e geram pressão sobre vendas para se eliminar esses estoques) e faltas de produtos.

Apesar das relevantes vantagens permitidas por esse modo de produção, a identificada inexistência de trabalhos publicados em periódicos científicos sobre o tema indica uma clara necessidade de maior divulgação e discussão sobre suas premissas e técnicas, além de melhor compreensão de seu desempenho em casos reais. No Brasil, tem-se o conhecimento de três empresas – todas de grande porte – que vêm implantando a abordagem MTA, duas produtoras de bens de consumo e uma terceira do ramo varejista. De qualquer forma, a ausência de trabalhos científicos voltados a avaliar casos reais não permite uma avaliação crítica mais efetiva de seu desempenho.

Este artigo se delimitou a apresentar alguns conceitos básicos sobre a aplicação do TPC-S em ambientes MTS, mas alguns detalhes técnicos não foram aqui incorporados e que deveriam merecer atenção especial em futuras pesquisas. Destacam-se, por exemplo, as formas pelas quais sistemas híbridos MTO e MTA devem gerenciar prioridades de produção e capacidade de produção; como lidar com sazonalidades na demanda; como implantar o modo MTA para o nível de componentes (aplicável em ambientes como *assemble to order* - ATO); como implantá-lo em sistemas produtivos que apresentam elevados tempos de *setup* ou mesmo *setups* dependentes; como determinar para que itens e para quais mercados o modo MTA de produção faz sentido; como migrar de um ambiente MTS ou MTO para MTA; e como identificar as funcionalidades essenciais que um *software* deve possuir para garantir a operacionalização desse modo de produção.

Referências

GOLDRATT, E. M.; COX, J. **The Goal**. Great Barrington: North River Press, 1984.

GOLDRATT, E. M. **Light blue 1, or**: a way to exploit the protective capacity - Have your cake and eat it too. Roelofarendsveen: Goldratt Marketing Group, 2009a. (Série Goldratt's TOC Golden Nugget – parte 11). Disponível em: <www.goldrattconsulting.com>.

GOLDRATT, E. M. **Moving from Make to Stock (MTS) to Make to Availability (MTA) - GST MTA**. Roelofarendsveen: Goldratt Marketing Group, 2009b. Série Goldratt Webcast. Disponível em vídeo.

GOLDRATT, E. M. Standing on the Shoulders of Giants - Production concepts versus production applications: The Hitachi Tool Engineering example. **Gestão & Produção**, v. 16, n. 3, p. 333-343, 2009c. http://dx.doi.org/10.1590/S0104-530X2009000300002

HOPP, W. J.; SPEARMAN, M. L. **Factory Physics**. Foundations of Manufacturing Management International Edition. Burr Ridge: Irwin McGraw-Hill, 2000.

LEE, J. H. et al. Research on enhancements of TOC Simplified Drum-Buffer-Rope system using novel generic procedures. **Expert Systems with Applications**, v. 37, p. 3747-3754, 2010. http://dx.doi.org/10.1016/j.eswa.2009.11.049

PIRES, S. **Gestão da Cadeia de Suprimentos (Supply Chain Management)**: conceitos, estratégias, práticas e casos. São Paulo: Atlas, 2004. 310 p.

ROTHER, M.; HARRIS, R. **Criando Fluxo Contínuo.** São Paulo: Lean Institute Brasil, 2002.

SCHRAGENHEIM, E. Make to stock under Drum-Buffer-Rope and Buffer Management Methodology. In: APICS INTERNATIONAL CONFERENCE, 2002, Nashville. **Proceedings**... Nashville: APICS, 2002. Session I-09, 5 p.

SCHRAGENHEIM, E. Managing Make-to-Stock and the concept of Make-to-Availability. In: COX III, J. F.; SCHLEIER, J. G. (Orgs.). **Theory of Constraints Handbook**. New York: McGraw-Hill, 2010. p. 239-264.

SCHRAGENHEIM, E. Supply Chain Management: The Production part, the TOC way so far and what lies ahead. In: THEORY OF CONSTRAINTS INTERNATIONAL CERTIFICATION ORGANIZATION CONFERENCE, 2012, Chicago. **Proceedings**...Chicago: TOCICO, 2012.

SCHRAGENHEIM, E.; DETTMER, H. W. **Manufacturing at Warp Speed**. Boca Raton: St. Lucie Press, 2001.

SCHRAGENHEIM, E.; DETTMER, H. W.; PATTERSON, J. W. **Supply Chain Management at Warp Speed**. Boca Raton: Taylor & Francis, 2009. http://dx.doi.org/10.1201/9781420073362

SOUZA, F. B.; BAPTISTA, H. R. Proposta de avanço para o método Tambor-Pulmão-Corda Simplificado aplicado em ambientes de produção sob encomenda. **Gestão & Produção**, v. 17, n. 4, p. 735-746, 2010. http://dx.doi.org/10.1590/S0104-530X2010000400008

SURI, R. **Quick response manufacturing**: a companywide approach to reducing lead times. Productive Press, 1998.

THURER, M.; GODINHO FILHO, M. Redução do lead time e entregas no prazo em pequenas e médias empresas que fabricam sob encomenda: a abordagem Worload Control (WLC) para o Planejamento e Controle da Produção (PCP). **Revista Produção**, v. 19, n. 1, p. 43-58, 2012.

VOLLMANN. T. et al. **Manufacturing Planning & Control Systems for Supply Chain Management**. New York McGraw Hill, 2005. 598 p.

WEBSTER, J.; WATSON, R. Analyzing the past to prepare for the future: writing a literature review. **MIS Quarterly**, v. 26, n. 2, p. xiii-xxiii, 2002.

WHETTEN, D. A. What Constitutes a Theoretical Contribution? **Academy of Management Review**, v. 14, n. 4, p. 490-495, 1989.