XJTU-IE, 2007.9-2008.9, M: xuchen.xuchen@163.com(徐琛), lucifer_tcl@126.com(李慰祖)

第十七章 供应链管理

工作有终结的一天，
而教育永无止境。
——大仲马

注：*backorder* 在这里往往译作“延迟”、“缺货”、“迟单”或“积压单”；*work backorder* 译作“作业积压单”。*backorder*（迟单，缺货）与 *stockout*（脱销，缺货）并用时的区别应在于 *MTO* 与 *MTS* 环境。

17.1 引言

这本书的主题就是库存在生产系统运营行为中的核心作用。第一篇中，我们对库存控制及其与生产控制的关系进行了历史回顾。第二篇中，我们进一步研究了库存（具体来说是WIP）与产出、周期时间等绩效量度之间的相互作用，第三篇中，我们要将历史的和工厂物理学的见识结合起来，解决制造系统中库存管理的实际问题。我们的目标提高整个系统的库存效率（*efficiency*）。也就是说，我们不是要简单地削减库存，而是要实现以最小投资构建库存的目标。照时髦的说法，这种库存持有与流动的全系统协调称作**供应链管理（supply chain management**）。

出于这里讨论的需要，我们将供应链中的库存分成四类：

1. **原材料（Raw material）**是从工厂外部采购的用于工厂内部制造/装配作业的元件、组件、或物料。
2. **在制品（Work in process，WIP）**包括所有已投入产线但尚未完成的工件或产品。
3. **制成品库存（Finish goods inventory，FGI）**指尚未出售的完工产品。
4. **备件（Spare parts）**指用于维护或修理生产设备的部件。

持有这几类库存的原因不同，因而提高其效率的选择也不同。因此，我们在以下的讨论中按各类分别对待。（582|583）

17.2 持有库存的原因

17.2.1 原材料

如果我们能以完全准时制的方式（即，恰在生产系统需要时）从供应商处接收物料，就不需要保有任何原材料库存了。这在现实中永远不可能出现，故所有的制造系统都会储存原材料。影响原材料存量规模的因素主要有三个。

1. **成批（Batching）**。供应商给出的数量折扣、工厂采购职能的限制（如，可下达和追

踪的订单数量有限）以及运输经济批量刺激了对原材料的大量订购。[1]我们称讨论成批问题时的库存为**周期库存（cycle stock）**，因为它表示在订购周期内持有的存货。

2. **变动性（Variability）**。当生产超前于排程、供应商配送落后于排程或质量问题导致过量报废时，如果没有额外的存货可用，产线就要由于缺少原材料而关停。这种额外的存货可以直接计划为**安全库存（safety stock）**（即，通过订购使得存货的期望水平保持在安全水平之上）或作为安全提前期（safety lead time）（即，以在需要之前到达的方式订购原材料并因而以原材料库存的形式等待）的结果。两种情形中，我们都称所持的用于防范变动性的存货为安全库存。

3. **废弃（Obsolescence）**。需求或设计变更致使某些原材料不再有用，所以制造系统中的某些库存无益于上述两种目的。这类**废弃库存（obsolete inventory）**，曾作为周期库存或安全库存采购而现在却完全无用，并应在财务报告许可的情况下尽快处理及勾销。

了解持有原材料库存的这些原因有助于确定改善措施。但是，必须记得它们不是严格孤立的。比如，我们在第二章指出，安全库存和周期库存防范的变动性（即，由于以非常大的批量订购，我们就降低了库存水平低于可能出现缺货情况的频率）。还有，废弃库存的水平显然受周期库存和安全库存的影响（即，如果以大批量订购或持有大量的安全库存，就会出现由于系统变更使得大量库存废弃的风险）。充分认识到这些相互作用还可以帮助我们设计原材料管理政策。

17.2.2 WIP

除了 JIT 的零库存目标，我们永远不能以零 WIP 运行制造系统，因为如在第二篇所见零 WIP 导致零产出。在第七章，我们推导出**临界 WIP（critical WIP）**，即产线在最佳情形下实现全部产出所需的最小量 WIP 。在现实情形下，实际 WIP 水平常常大量地超出临界 WIP（如，常是 20~30 倍）。这种 WIP 将处于以下五种状态中的一种：（582|583）

1. **排队（Queueing）**等待资源（人员、机器或运输工具）。
2. 正在被一种资源**加工（Processing）**。
3. **等待成批（Waiting for batch）**，不得不等待其他加工任务的到来以形成批次。这个批次将用于补给大量制造作业（如，由于同时烧结而要一满屋子加工任务的热处理）或转运作业（如，加工任务只以满栈板移动）。注意加工或转运批次一旦形成，任何其他的等待资源的时间（如，等待热处理设备或叉车可用）都被划入排队时间。
4. **转运（Moving）**，在各种资源之间运输。
5. **等待匹配（Waiting for match）**，组件在装配作业前等待对应物的到达以使装配发生。一旦“成套”部件到达，任何其他的等待装配资源的时间都定义为排队时间。

[1] 正是这些因素引起了第二章 EOQ 模型中的固定订货成本。EOQ 模型平衡了这项固定订货成本与库存持有成本来确定一个经济订货批量。

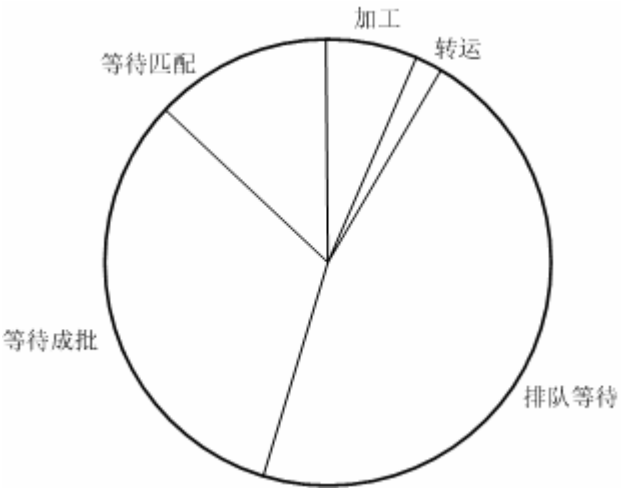

图 17.1 制造系统中 *WIP* 的典型分类

要在 WIP 管理/削减项目中使用上述分类，需要两种认识。第一，如图 17.1 所揭示的，大多数制造系统中处于转运或加工状态的 WIP 的比例很小（如，低于 10%；见 Bradt 1983 的经验资料）。WIP 的主要状态是排队、等待成批或等待匹配。显然，WIP 削减项目要想成功，必须解决后一类的问题。

第二，排队 WIP、等待成批 WIP 和等待匹配 WIP 的起因不同。如在第二篇所见，排队的主要由高利用率和变动性（流动变动性以及加工变动性）引起。等待成批 WIP 显然由成批加工或转运引起；批量越大，需要的 WIP 越多。等待匹配 WIP 由到达装配作业的工件缺乏协调引起，这种不协调有些要归因于简单的流动变动性有些要归因于生产控制过程。这些区别意味着不同种类的 WIP 将受不同管理政策的作用，如后面要讨论的。（584|585）

17.2.3 FGI

如果我们可以把生产的所有东西在加工完成后立即直接送达客户，那就不会需要 FGI 了。有些制造系统（如，生产定制化产品的重负荷加工车间）几乎能够做到这一点，而许多都不行。持有 FGI 的基本原因有五条。

1. **客户响应（Customer responsiveness）**。为了提供短于制造周期时间的配送提前期，许多企业使用**备货生产（make-to-stock）**（而非**接单生产（make-to-order）**）政策。许多产品，比如，建筑材料（如，屋顶纸板、木料）、标准电子元件（如，电阻、电容）和基本的食品（如，苏打、食用油）都是**日用品（commodity）**。就这点来说，它们的价格与规格（如，质量）都由市场决定。这样，唯一的竞争性问题就是配送了。为此，这类产品常常生产出来作为存货。支持给定备货生产系统所需的 FGI 数量依赖于客户需求变动性以及想要的客户服务水平。

一种联合备货生产与接单生产效力的方法是**接单装配（assembly-to-order）**。这种程序生产组件作为存货，然后在接到订单时组装。以第十章的术语来讲，接单生产的**推/拉界面（push/pull interface）**在原材料处，备货生产的在制成品处，而接单装配的在二者之间的某个地方。结果就是，其响应比传统的接单生产快，库存比备货生产低。

2. **成批生产（Batch production）**。无论出于什么原因，如果生产按预设的数量（批量）发生，那么产出有时候不会完全匹配订单，任何多余的都将进入制成品库存。例如，一间钢厂的批量为 250 吨（以有效利用铸造炉），但订单平均为 50 吨，则它将常常把批次中的剩余部分放入 FGI。

3. **预测误差（Forecast error）**。在没有固定的客户订单情况下投放加工任务时，无论是在备货生产体系中补充存货还是在接单生产体系中满足期望的订单，未如期望情形销售的产品将不可避免地累积。这种过量将以 FGI 的形式结束。

4. **生产变动性（Production variability）**。在订单不能及早送达（或在多早送达时有限制）的接单生产体系中，生产时机（*timing*）的变动性有时将导致居留于 FGI 以等待运送的产品。对于接单生产或备货生产体系，产量的变动性（如，由于随机的产出损失）可能引起相对于需求的过量生产（如，如果我们“过度膨胀”来补偿产出损失）。再一次地，过量部分将进入 FGI。

5. **季节性（Seasonality）**。一种处理季节性变化需求（如，剪草机、吹雪机、室内空调）的方法是在淡季建立库存来满足旺季需求高峰。这种**预建库存（build-ahead inventory）**也将成为 FGI 的一部分。

注意这些提高 FGI 的因素是相互影响的。例如，一旦建立 FGI 来提供短的提前期或满足季节性需求，也就增大了遭受预测误差的风险。因此，系统地看待 FGI 是很重要的。只有这样我们才可能考虑提供显著优势的基本结构变革。比如说，系统可能更应当以接单生产的方式运行，而非备货生产；可能应当使用额外的产能或季节性劳力来解决季节性需求，而非预建库存，或者是应当重新定位推拉界面（如，使用接单装配策略）。在对改进策略的讨论中，我们再回到这些选项上来。（585|586）

17.2.4 备件

备件并不直接用于生产最终产品，但它们确实通过保持机器运行而支持生产过程。在许多系统中，保有库存占用的资金并不大，但短缺的后果可能极其严重（如，整条产线可能由于缺少某种关键工件而关停）。然而另一些系统（如，支持全国联网机器维修的合约服务）中，备件库存占用的资金可能非常显著。对于每一种情形，存储备件的基本原因都是

1. **服务水平（Service）**。任何备件系统的主要目标都是支持维护和修理过程。如果修理人员必须等待某个部件（如，从一个集中存储地或外部供应商），那么完成这次修理的时间就会被大大延长。在其他条件都一样的情况下，较高的服务水平（即，避免由缺乏部件引起的延迟）需要较高的备件库存水平。

2. **采购/生产提前期（Purchasing/production lead times）**。如果备件可以立刻买到或产出，那么就没有必要去存储。不幸的是，这种情况永远不会出现；所以为了提供设定的服务水平，我们必须持有备件库存。一般来说，获取备件的提前期越长，我们要持有的库存就越多。

3. **成批补充（Batch replenishment）**。如果在补充备件时存在规模经济（如，采购时有数量折扣或生产时有大的固定成本），那么大量采购或生产就很有意义了。当然了，较大的补充批次意味着较高的库存水平。

从理论上讲，备件库存系统与 FGI 系统区别不大。对于两者，我们都是存储部件，可能是以批次，来以某个服务水平满足不确定的需求过程。由于这种相似性，很可能可以用差不多的工具来控制备件和 FGI。然而，也要认识到这两类库存的不同作用。例如，基于行业标杆，为 FGI 设定 90%的补给率（fill rate）可能是合理的。但考虑到某种关键部件短缺引起较长时间机器停工的物料与财务后果，90%的部件补给率可能是远远不够的。故而，当使用相似的模型来阐述这两类库存的问题时，我们必须仔细考虑所包含的成本与目标，从而为模型设定合适的参数。

回顾了持有不同种类库存的原因之后，我们现在来看提高各类库存效率（即，以较小的总投资达到同样的收益）的技术。（586|587）

17.3 管理原材料

如上面讲到的，原材料管理的目标是使它们在为生产过程所需时可用，且持有量不超过需要量。有些策略可以在所有部件上增强我们的这种能力。另一些只在某些特定类别的部件是经济可行的。因此，我们的基本策略是一种“分而治之（divide and conquer）”，也就是对不同类别的原材料应用不同的方法。在下面的小节里，我们给出一些整体改进策略，一种分类方案，以及对应于具体部件类别的聚焦控制政策。

17.3.1 能见度改进（Visibility Improvement）

很明显，如果我们知道需要什么部件而不用猜测，就可以在采购原料上做得更好。不幸的是，制造周期时间和采购提前期通常很长，以致于我们不得不在接收固定客户订单之前采购某些原材料。短期内，除了维持原材料安全库存以缓冲采购错误，我们别无选择。但长期内，我们可以通过下面的政策来改善境况：

1. **改善预测（Improve forecasting）**。如果真正糟糕的是对未来需求的预测，那么可以通过使用系统化的预测技术来达到更好的效果（见 附录 13A）。但是，这些方法仍然避免不了预测的第一定律——预测总是错误的（*forecasts are always wrong*）。因此，通过预测达到的可能改进是有限的。

2. **缩短周期时间（Reduce cycle times）**。缩短制造周期时间意味着加工任务可以更接近交期投放。因此，当客户需求固定时，部件采购可以较迟下单。对于周期时间很长的系统，缩短周期时间可以比使用精巧的预测技术更好地改善预测。我们在 17.4 节中讨论了缩短周期时间（和 WIP）的具体技术。

3. **改善排程（Improve scheduling）**。如果排程很糟糕，采购件的计划使用将与实际使用有很大不同。比如，无限产能 MRP 模型生成的排程计划出的加工任务完成期可能早于实际情形。这将导致采购的部件早于实际需要到达，并因而引起原材料库存膨胀。良好的有限产能排程方法将生成更现实的排程，并因而使采购件在更接近需要时送入。

17.3.2 ABC 分类

大多数制造系统中，小部分的采购件占用了大部分的采购支出。[2]因此，为了获取最大效益，管理的重点应聚集于这些部件。为了达到这一点，许多制造企业对采购的部件和物料采用某种**ABC分类（ABC classification）**。在对ABC分类的典型阐述中，我们按各部件的年度采购资金对其排序，并定义

A 类部件（A parts）：前 5%~10%的部件，占到 75%~80%的总年度支出。（587|588）
B 类部件（B parts）：接下来 10%~15%的部件，占到 10%~15%的年度总支出。
C 类部件（C parts）：末尾约 80%的部件，仅占约 10%的年度总支出。

因为数量相对较小而成本较高，所以值得用精巧、费时的方法来紧密地协调 A 类部件

[2] 这是**帕累托定律（Pareto's law）**的一个例子。帕累托定律一般称作“80-20 定律”，取名于以意大利经济学家 Vilfredo Pareto（1848-1923），他观察到大部分的财富趋向集中于人口的很小一部分。

的到达与生产过程对它们的需要。此类措施一般不适合 C 类部件，因为持有这类过量部件的库存费用并不大。B 类部件居中，所以它们应受的关注多余 C 类，但少于 A 类。各个系统的具体做法不一样，但 ABC 分类的要点是一致的：不同种类部件的库存应当区别对待。

我们在下面的小节中讨论一些合适的技术以及各自的适用情形。

17.3.3 JIT

库存持有成本很高的昂贵 A 类部件，以及持有库存很不方便的极端大量部件（如，包装材料），都需要紧密的库存控制。保持部件的最低库存水平的方法是，协调配送与生产过程的需用。这正是准时制（JIT）背后的观点。

与供应商的典型 JIT 合约要求以紧密匹配生产排程需用的小量进行频繁的配送（如，每周、每天或更甚，取决于具体系统）。由于生产排程变化无常，大多数 JIT 合约允许几乎在配送开始之前调整订购量（尽管大多数合约也指定了允许的变更上限）。

为了使供应商更容易地满足配送需求，管理良好的 JIT 采购体系向供应商知会其生产排程。基本目的是使供应商尽可能快地应对排程的变更。但这类可见性还有其他好处。它能消除对采购订单的需要。例如，与汽车制动器供应商的合约可能要求它们查看最终装配计划并送入合适的制动器来支持生产。该体系甚至还可以通过简单地计数生产的汽车并为使用的制动器支付供应商来消除账目清单。（隐含的也是合理的假设是每辆汽车使用一套制动器。）

从理论上讲，与供应商的 JIT 合约很有吸引力。然而，为了使其运行起来，供应商必须在配送时机和质量上可靠。如果出货晚点或有缺陷，则整条产线将因为缺少工件而停止。由于这一点，广泛依赖 JIT 配送原材料的企业一般都制定了某种形式的**供应商认证（vendor certification）**项目。良好的供应商认证项目包括评审程序，以及帮助供应商改进其系统的努力。（588|589）

由于严格监督与培养供应商是原材料 JIT 配送的先决条件，对于小企业来说这种方式或许并不可行。一家采购量只占供应商业务量很小一部分的企业可能无力说服供应商按 JIT 方式配送部件。当前朝向响应性的趋势（如，体现在基于时间的竞争（*time-based competition*）、全周期时间（*total cycle time*）、短周期制造（*short-cycle manufacturing*）这样的流行语中）可能增加了愿意为非最大客户提供 JIT 配送的企业数量，但真正的 JIT 合约对于典型的小企业仍然是不可得的。故而，它们必须寻求其他的办法来管理昂贵的原材料库存。

17.3.4 为采购件设置安全库存/提前期

即使企业不能或不愿对昂贵的A类部件使用JIT配送，仍然有必要紧密地连接这些部件的采购与生产排程（替代这样的方式，即偶尔以大批量采购并从储备充分的原料仓中供给产线）。以MRP的术语讲，这意味着昂贵的部件应当以**批对批**（或称因需定量法）**（lot-for-lot）**的方式下单。例如，如果我们计划在*n*周后生产 1,000 件高清晰度监视器（high-resolution monitor），就应当订购 1,000 件阴极射线管（cathode-ray tube）并使它们在先于排程某个固定安全提前期时到达。[3]

注意这种方法不同于 JIT，因为我们是按计划的（*planned*）排程订购部件，而不是让它们与实际的（*actual*）排程同步配送。但如果真正的 JIT 无法做到，这也就是我们的最佳选择了。当然，如果（当）排程变更，设定的产量可能由于缺少合适的原材料而无法完成。这意味着当排程中固定订单较多而预测订单较少时，短的配送提前期比长的更受欢迎，因为采购可以更靠近交期。长期来看，高价、短提前期的供应商可能比低价、长提前期的更划算。

如我们在第十二章在供应商质量的背景下所见，采购件的管理对于使用多部件进行组装的系统极其重要。在那里我们已指出，如果以每个都有 95%的服务水平这样足量的安全提

[3] 如果要考虑产出损失，我们也得维持一个安全库存的计划水平。

前期采购 10 种部件，则全部 10 种部件如期到达来满足排程的概率为 0.95^10 = 0.5987，是个非常糟糕的服务水平。使用许多采购件进行组装的系统对各个部件都要求极高的服务水平从而来可靠地满足排程。例如，若要 10 种部件以 95%的概率在组装时到齐，则每种的服务水平要达到 0.95^(1/10) = 0.9949。

最后，注意并不必要为各种批对批采购的 A 类部件设置相同的服务水平。如果某种部件尤为昂贵，就应当将它的服务水平设得相对较低（比如说，96%），并将其他部件的服务水平设得较高以作补偿。如果共有 n 种部件，并令 S_j 表示部件 j 的服务水平，则以 95%的概率实现排程就需要选择各个 S_j 使

$$S_1 \cdot S_2 \cdots S_n = 0.95$$

选择各个部件的服务水平从而以最小的平均库存投资实现整体服务水平的规范方法，见 Hopp 和 Spearman（1993）的著作。（589|590）

17.3.5 为采购件设置下单频率

以上提到的 JIT 和批对批采购法对于昂贵的 A 类部件很适用，对于中间的 B 类也可以，但一般不适于便宜的 C 类。预定螺钉、垫圈、二极管之类物品的配送与生产排程紧密同步没什么意义。由此增加的断供风险，以及额外的采购和物料搬运费用不能通过削减的库存投资来抵消。

管理廉价采购件的问题可以从**决定批量（lot sizing）**的角度来看。基本的权衡在于库存投资与采购成本。回忆起这正是经济订购批量（EOQ）所阐述的。事实上，如果愿意忽略部件之间的相互影响，就可以直接应用第 2.2 节中给出的单产品模型。即，若令

N = 系统中不同部件的种类数

D_j = 部件 j 的需求速率（件/年）

c_j = 部件 j 的单位生产成本

A = 采购任何部件一次下单的固定成本

h_j = 部件 j 的年库存持有成本系数

Q_j = 部件 j 的下单数量或批量（决策变量）

我们可以用标准的 EOQ 公式为部件 j 计算批量：

$$Q_j^* = \sqrt{\frac{2AD_j}{h_j}} \qquad (17.1)$$

这个公式中最难以估算的输入是固定订购成本[4]，A。理想地，它应当反映每次下定单时都会发生的成本，可能包括实际的运输成本、采购部门处理和追踪订单所用的时间、接收订

[4] 回忆起第一篇中我们批评生产系统的固定采购成本假设，因为它常常替代表示产能（capacity）约束，而产能约束随时间变化并不能先于排程确定。可是对于采购系统，可能不需要考虑能力（capacity），这样固定采购成本就是一个合理得多的建模假设。

单所需的时间等等。间接成本（overhead cost，或称管理费用）（如，维持一个采购部）不应摊入 A_j。

上面方法的一个潜在问题是没有考虑部件之间的相互影响，这些情况是（1）部件共享公用的配送系统以及（2）需要考虑采购部门的整体能力。例如，如果不同的部件可以共享配送卡车，这时就有动力去尽量同时地下单。第二章中，我们提到 2 的幂（power-of-two）补充政策是一种解决的办法。考虑到 EOQ 成本函数的稳健性以及输入数据的粗糙性，多部件采购问题的一个合理解法是，简单地使用 EOQ 公式计算各部件的最优订货间期（即，D_j / Q_j^*），再圆整到最近的某个订购周期 2 次幂。例如，若可以每周下单一次，则圆整 EOQ 间期到这组时间中的最近值：1 周、2 周、4 周、8 周，等等。

要考虑采购职能的整体能力，我们可以将该问题视为在平均下单频率不超过某个指定值 F 的约束下最小化所有部件的总库存持有成本。由于年度采购订单的总量等于每项的平均下单频率乘以 N，以上的方法也等效于在年度订单总量不超过 NF 的约束下最小化总库存投资。但我们发现平均下单频率更易于理解，于是就以这样的方式来陈述问题。（590|591）

要建立数学模型，我们记起如果部件 j 的订购量是 Q_j，则部件 j 的平均库存（以件为单位）为 $Q_j/2$，且年度持有成本为 $h_j\ Q_j/2$。部件 j 的下单频率是 D_j / Q_j。因此，总持有成本是 $\sum_{j=1}^{N} h_j Q_j / 2$，平均下单频率是 $1/N(\sum_{j=1}^{N} D_j / Q_j^*)$。故我们可以将在平均下单批量 F 约束下最小化总持有成本的问题表述为

$$\text{最大化} \qquad \sum_{j=1}^{N} h_j Q_j / 2 \tag{17.2}$$

$$\text{受限于:} \qquad \frac{1}{N}\sum_{j=1}^{N} \frac{D_j}{Q_j} \le F \tag{17.3}$$

注意到如果用单位成本 c_j 替换 h_j，就变成在平均下单频率约束下最小化总库存投资（*investment*）的问题。有些决策制定者认为库存投资比持有成本容易理解，但当 $h_j = ic_j$ 时二者是等效的（即，导致同样的批量），其中 i 是利率。所以选择用持有成本还是用库存投资作目标函数，就因人而异吧（just a matter of taste）。

上述表达式是**非线性规划问题（nonlinear programming problem）**的一个例子。求解这类问题的标准技术是**拉格朗日法（method of Lagrange）**，它通过为违反约束附加补偿条件并将其整合进入目标函数，将约束优化问题转化为非约束的（Bazaraa 和 Shetty 1979）。听起来很复杂，但简而言之就是为式（17.1）寻找满足约束（17.3）的固定采购成本。我们通过下面这样的迭代搜索方法做到这一点。

算法（多产品 EOQ 模型）

步骤 0. 为 A 赋初值

步骤 1. 将 A 带入式（17.1）来计算所有 $j = 1, 2, 3, \ldots, N$ 的批量 Q_j。

步骤 2. 计算下单频率：

$$F(A) = \frac{1}{N}\sum_{j=1}^{N}\frac{D_j}{Q_j}$$

步骤 3. 如果 $F(A) = F$，结束。[5]否则，

若 $F(A) > F$，减小 A

若 $F(A) < F$，增大 A

回到步骤 1.

A值的增大与减小可以用试错法或其他的一些更精巧的方法，如两分法（interval bisection）。[6]只要使用的方法在接近最优值时的计算间隔越来越小，该程序将逐渐收敛。（591|592）

当程序结束时，就得到最优订购批量 Q_j，$j = 1, 2, 3, \ldots, N$。我们还得到合适的固定订购成本 A。这项成本也可以理解为*总库存持有成本对平均下单频率的边际减量*（*decrease in total inventory holding cost per unit decrease in the average order frequency*）。如果我们知道愿意为降低平均下单频率（每项每年减少一次）支付多少年度持有成本，就可以立即使用这个值代入式（17.1）来计算最优订购量。如果像一般情况那样，这个值难以确定，我们可以对多个 F 值运行上述算法，并绘出最优持有成本（若用 c_j 替代 h_j，则是库存投资）与平均下单频率的关系图。该曲线有些像图 2.3 表示的单产品情形，但这里是多产品。

我们可以直接应用上述程序计算出的 Q_j，$j = 1, 2, 3, \ldots, N$。然而，如果同时订购会产生节约，还是有必要将与这些批量相关的订购间期圆整到 2 的幂。方法是给出部件 j 的再订购间期

$$T_j^* = \frac{Q_j^*}{D_j}$$

如果将 T_j^* 圆整为最近的 2 的幂之数值，那么如第二章中的讨论，不同的部件的订单将趋于“排成一列”（line up）。当然了，这样的圆整将会影响库存水平与平均下单频率。如果将 T_j^* 圆整为 $T_j^{'}$，订购批量将变成

---

[5] 由于 $F(A)$ 是连续值，它不会严格等于 F。所以一般当 $F(A)$ 落入某个指定的微小公差范围内就停止。

[6] 两分法基本上始于 A 的两个点，即一个过大的上限（即，使得 $F(A) < F$）和一个过小的下限（即，使得 $F(A) > F$），并尝试代入二者的中点。如果过大，则中点替代原来的上限；如果过小，则中点替代原来的下限。上下限的区间将越来越小。当足够小（即，低于某个指定的公差）时，我们就停止迭代。

$$Q_j^{'} = T_j^{'} D_j$$

由此，实际库存持有成本将是

$$\frac{\sum_{i=j}^{N} c_j Q_j^{'}}{2}$$

且实际平均下单频率将是

$$\frac{1}{N}\sum_{i=j}^{N}\frac{D_j}{Q_j^{'}}$$

如果库存投资的增量相对于最优水平过大，或平均下单频率比目标水平 F 大得多，则从 2 的幂圆整方法获得的收益可能抵不上它们的损失。如果实际解和最优解之间的差别不大，则此类圆整可能是值得的。

例子：

为了阐述上述程序，我们考虑一个非常简单的四部件例子，具体数据在表 17.1 中给出。目标是在平均年度下单频率 $F = 12$（即，每月一单）的约束下最小化平均库存投资。注意由于目标是平均库存投资，我们使用一个等于单位成本的持有成本系数 $c_j = h_j$。

表 17.1 多部件订货批量的输入数据

工件 j	D_j	c_j
1	1,000	100
2	1,000	10
3	100	100
4	100	10

表 17.2 多部件举例的计算结果

迭代次数	*A*	*Q1(A)*	*Q2(A)*	*Q3(A)*	*Q4(A)*	*F(A)*	库存投资($)
1	1.000	4.47	14.14	1.41	4.47	96.85	387.39
2	100.000	44.72	141.20	14.14	44.72	9.68	3,873.89
3	50.000	31.62	100.00	10.00	31.62	13.70	2,739.25
4	75.000	38.73	122.47	12.25	38.73	11.18	3,354.89
5	62.500	35.36	111.80	11.18	35.36	12.25	3,062.58
6	68.750	37.08	117.26	11.73	37.08	11.68	3,212.06
7	65.625	36.23	114.56	11.46	36.23	11.96	3,138.21
8	64.065	35.80	113.19	11.32	35.80	12.10	3,100.68
9	64.845	36.01	113.88	11.39	36.01	12.03	3,119.50
10	65.235	36.12	117.22	11.42	36.12	11.99	3,128.87
11	65.040	36.07	114.05	11.41	36.07	12.01	3,124.19
12	65.138	36.09	114.14	11.41	36.09	12.00	3,126.53

表 17.2 汇总了将上述程序应用于这个例子的结果。表中最右列给出对应于各个订购量组合的平均库存投资，计算式是（592|593）

$$\frac{\sum_{i=j}^{N} c_j Q_j}{2}$$

要启动这个程序，我们始于 $A = 1$。如表 17.2 所示，它导致了 96.85 的平均下单频率，显然是很高的。因此，A 值必须增大。故我们尝试 $A = 100$。正如预料，由于严重地惩罚频繁下单，它导致了高得多的订购量，且平均下单频率跌至 9.68。这个值过低，我们现在就确定了 A 的上限，也就是说 A 的最优值（实现下单频率为 12 的那个）在 1~100 之间。故再尝试 $A = 50$。它导致 13.70 的下单频率，$A = 50$ 过低了。故再尝试 $A = 75$。它使下单频率降到 11.18。按照这样的方式继续下去，该程序逐渐收敛到设定的下单频率。注意到只要部件种类不多，所有的计算都可以在电子表格中处理。事实上，用 Excel 自带的目标求解（Goal Seek）或 Solver 可以很简单地找到合适的 A 值。

表 17.2 最后一行给出多部件批量程序的求解结果。数字显示，部件 1、2、3 和 4 的最优批量分别是 36.09、114.14、11.41 和 36.09。注意到部件 2 的批量大于部件 1，部件 4 的批量大于部件 3。这是因为部件 2 比部件 1 便宜，部件 4 比部件 3 便宜。直观看来，最优批量随成本递减。

进一步地，部件 1 的批量大于部件 3，即使二者的成本相等。这是因为对部件 1 的需求较大。同样的关系对于部件 2 与 4 也成立。正如预料，批量随需求速率递增。（593|594）

最后，注意部件 1 与 4 的批量相等。这是因为

$$\frac{D_1}{c_1} = \frac{D_4}{d_4}$$

从表达式（17.1）可以明显看到，批量取决于 D_j 和 h_j（并因而还有 c_j）且仅取决于它们的比率。

该程序的输出是 $A = 65.138$，它给出对改变平均下单频率的成本（以库存投资的形式）的估值。下单频率提高一次（到每年 13 次）将降低库存投资\$65.14，而减少一次（到每年 11 次）将提高库存投资\$65.14。可是，我们必须注意由于成本函数非线性，这些成本都是近似值。实际上，下单频率提高一次的节约会小于\$65.14，而提高一次的成本将大于\$65.14。尽管这样，还是给使用者一个关于更频繁下单的库存价值的大体意思。

结果的 A 值也可用于对原始下单频率目标的检查。如果下单的实际成本低于（高于）\$65.14，则我们本应该选择一个每年大于（小于）12 次的频率。要点是如果我们已有关于 A 和 F 应当是怎样的想法，而对两者都不完全确定，那么就可以通过对两者交互校验（cross-check）并调整到都很合理，从而得到更好的解。

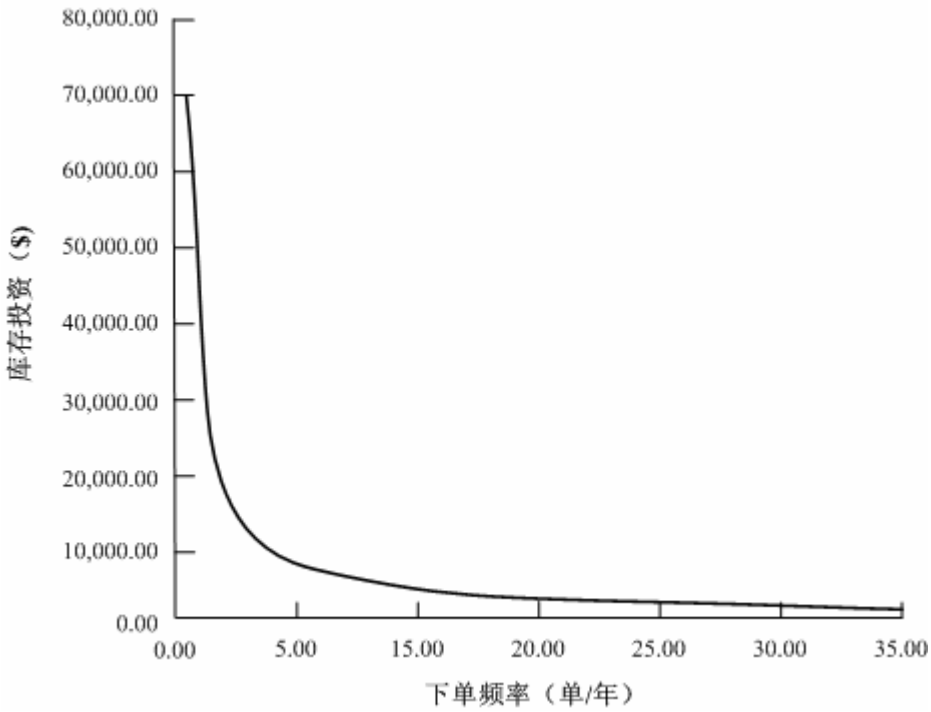

图 17.2 库存投资/下单频率

对于库存投资与下单频率之间的权衡，我们还可以做到更精确。注意到如果像表 17.2 那样追踪库存投资，则各个 *A* 值给出一个库存投资/下单频率组合。因此，通过在足够大的范围内改变 *A* 值，就可以生成库存投资与平均下单频率的关系图，如图 17.2 所示。注意到随着年度订单数目从零增到五时库存投资急剧下降。可是，在此之上提高下单频率，特别是在十之上时，效果就微弱得多了。这种收益递减（*diminishing return*）正好像图 2.3 所示的单产品情形。

最后，如果合并订单有利可图，我们可能会希望圆整采购间期到某个 2 的幂值。如果这样，首先计算采购间期：（594|595）

$$T_1^* = \frac{Q_1^*}{D_1} = \frac{36.09}{1{,}000} = 0.03609 \text{ 年} = 13.17 \text{ 天}$$

$$T_2^* = \frac{Q_2^*}{D_2} = \frac{114.14}{1{,}000} = 0.11414 \text{ 年} = 41.66 \text{ 天}$$

$$T_2^* = \frac{Q_2^*}{D_2} = \frac{11.414}{100} = 0.11414 \text{ 年} = 41.66 \text{ 天}$$

$$T_1^* = \frac{Q_1^*}{D_1} = \frac{36.09}{100} = 0.3609 \text{ 年} = 131.7 \text{ 天}$$

用天作为基本时间单位，我们选择 T_1' 到最接近 13.17 的 2 的幂值，即 $2^4 = 16$。选择 T_2' 和 T_3' 到最接近 41.66 的 2 的幂值，即 $2^5 = 32$。选择 T_4' 到最接近 131.73 的 2 的幂值，即 $2^7 = 128$。

这些订购间期产生如下的采购量：

$$Q_1 = \frac{D_1 T_1^{'}}{365} = 1{,}000 \times \frac{16}{365} = 43.84 \text{ 件}$$

$$Q_2 = \frac{D_2 T_2^{'}}{365} = 1{,}000 \times \frac{32}{365} = 87.67 \text{ 件}$$

$$Q_3 = \frac{D_3 T_3^{'}}{365} = 100 \times \frac{32}{365} = 8.77 \text{ 件}$$

$$Q_4 = \frac{D_4 T_4^{'}}{365} = 100 \times \frac{128}{365} = 35.07 \text{ 件}$$

把这些值代入计算库存投资和下单频率的式子，得

$$\text{库存投资} = \frac{\sum_{j=1}^{4} c_j Q_j^{'}}{2} = \$3{,}243.84$$

$$\text{平均下单频率} = \frac{1}{4}\sum_{j=1}^{4} \frac{D_j}{Q_j^{'}} = 12.12$$

由于我们已假定按 2 的幂订购间期合并订单会带来一些节约，这时的平均下单频率比原初 12 次的水平稍高一些也是可以接收的。但是要注意，库存投资从\$3,126.53 增加到\$3,243.84。这些增加的成本必须通过联合订单的某种好处来抵消（如，要处理的单独订单减少、分享卡车空间）以说明 2 的幂政策是值得的。

17.4 管理 WIP

管理 WIP 时要注意的第一件事就是，里特定律（Little's Law）

$$CT = \frac{WIP}{TH}$$

指出对于固定的产出，削减 WIP 与削减周期时间是直接相连的。因此，我们建议来提高 WIP 效率的办法正是那些用于压缩周期时间的。（595|596）

管理 WIP 的第二要点是，如之前指出的，大多数生产系统（即，间断的流水线（disconnected flow line））中大量 WIP 的状态是排队等待（由变动性和高利用率引起）、等待成批（由批次引起）或等待匹配（由缺乏同步引起）。因此，WIP 削减项目应当（明智地）指向降低利用率、熨平变动性或提高协同。

在下面的小节中，我们来回顾削减排队等待、等待移动以及等待匹配 WIP 的技术。

17.4.1 减少排队

回忆起对于平均加工时间为 t_e、加工时间变异系数为 c_e、到达变异系数为 c_a 以及利用率为 u 的单机工站，周期时间可以近似为

$$CT \approx \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u}{1-u}\right)t_e + t_e \qquad (17.4)$$

因此根据里特定律和关系式$u = r_a t_e$，其中r_a是到工站的平均到达速率，有

$$WIP = CT \cdot r_a \approx \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u}{1-u}\right)u + u \qquad (17.5)$$

故而，为了削减工站处的 WIP 和 CT，我们可以降低到达工站的变动性（c_a^2）、有效加工时间的变异系数（c_e^2）或利用率（u）。

达到这些目标的一般选择如下：

1. **设备变更/增加（Equipment changes/additions）**。工站处提升产能并因而降低利用率最简单的办法就是以更快的型号替换现有机器，或通过引进并联机器增加现有产能。这个方案几乎不需要什么想象力，但它很有效。可是，为了选择好的新增设备，我们必须考虑采购成本、对工站处产能和变动性的影响以及对下游（流动）变动性的影响。我们将在第十八章讨论一个关于此的框架。

2. **拉式体系（Pull systems）**。正如在第十章中所见，拉式体系将以较低的平均 WIP 水平达到同样的产出。原因在于向产线投料与产线的状态相协调（即，作业仅在产线有容纳空间时才被允许进入）。这有点像在线首削减c_a，但不确切。拉式体系真正做到的是将向产线的投料与线内作业的完成连接起来。更重要的是，它们建立了 WIP 上限，从而阻止了产线的 WIP 水平超越某个给定值。因此，拉式体系要求（*mandate*）对 WIP 进行削减。而挑战是，以不损失产出的方式实现这种 WIP 削减。这就需要采取某些其他的变动性削减或产能提升活动。

3. **有限产能排程（Finite-capacity scheduling）**。如果向产线投料时没有充分考虑产能（如，像在 MRP 中），则瓶颈资源处可能发生 WIP 爆炸。如第十五章所描述的，有限产能排程系统可以帮助规制投料合乎系统产能。尽管它连接投料与产出不像拉式体系那样强烈，（拉式体系连接投料与实际的（*actual*）产出，而有限产能排程连接投料与期望的（*expected*）产出），有限产能排程能通过防止向产线系统性地过量投料而显著地降低 WIP。理想情况是并用有限产能排程体系与拉式体系，从而在实际情况偏离排程时还能保持系统处于受控状态。（596|597）

4. **缩短换模时间（Setup reduction）**。在其他条件一样的情况下，缩短换模时间可以增加工站的有效产能，从而降低利用率。可是，一般来说缩短换模时间后，我们采用较小的批量并因而要进行更多次的换模。即使换模次数的增加抵消了产能增加的效果，如我们在第二篇中的讨论，较短且较多次的换模将降低工站处的有效变动性（c_e）。这将有助于减少该工站以及下游的排队等待（即，因为流动变动性也被降低了）。还有，如早些的讨论，如果可以按较小的批次生产，将过量产品存储为 FGI 的需要就变小了。

5. **提高可靠性/可维护性（Improve reliability/maintainability）**。提高平均失效间期或平均恢复间期都提高了机器的可用性并因而增加了它的产能。还有，降低平均恢复间期可以显著地削减有效变动性（c_e）。故而，这些类型的改善措施可以减少工站处的排队，并通过

降低下游的流动变动性，还可以减少后续工站处的队列。

6. **提高质量（Enhanced quality）**。如第十二章中讲到的，减少重工或产出损失可以从本质上提升产能以及削减有效变动性。出于这一点，质量改进的努力可以是 WIP/周期时间削减项目的主要部分。

7. **漂移作业（Floating work）**。经过交叉培训的工人可以移动到需要产能的地方，从饿提高了产线的有效产能。交叉培训也趋于使工人产生对产线的整体图景，并更多地思考产线中各工站面临的问题。对于人工装配系统，不管是同步的（paced）还是非同步的，漂移作业可以通过标示某些任务为“共享的”来达到效果。例如，某个组件可能被指定系于工人 A（上游）或工人 B（下游）。只要工人 A 跟得上产线，她将加工该共享组件。然而，一旦工人 A 落后于产线（如，一个质量问题使她慢下来），她可以将该组件交给工人 B 来继续。一般来说，漂移作业只在激励体系促进向产线整体目标努力的情况下对作业有效。

最后，我们得出与外购件 ABC 分类时相同的观点：*并非所有的 WIP 都应同等对待*（*Not all WIP need to be treated equally*）。很有必要按需用量来对部件分类。大量而族类较少的部件可以分配给流水线（High volume parts could be assigned to lines with few part families），并因而有较少的换模，且物流的平稳有利于使用高效的拉式体系。小量部件可以在加工车间环境中生产，因而以低效为代价的高柔性将仅仅影响整个业务的一小部分。这种类型的**聚焦的工厂（focused factory）**策略可以极大地简化有着多种不同部件的工厂的管理。

17.4.2 削减等待成批 WIP

出于加工原因的成批不可避免（如，一项需要 24 小时的成批烧结作业可能在大批量同时加工时才能提供足够的产能）。而出于转运原因的成批就是另一回事了。任何使加工任务以较小批次，进而有较少等待，从一个工站移动到另一个的措施，都将明显地降低 WIP 和周期时间。具体途径有：（597|598）

1. **批次分离（Lot splitting）**。回忆起加工批次与转运批次不必要一致。即使某个一次加工一件的工站由于换模时间很长而出于产能原因迫使（加工）批次很大，也没有必要等到该批次全部完成才移动某些加工任务到下个工站。例如，一个机加中心以 10,000 件的批量（即，在换模生产不同种类的曲轴之前）加工曲轴，却以 100 件的批量将产出品送入后续的涂饰工站。从理论上说，曲轴甚至可以一次一件地从机加移动到涂饰工站。限制因素是移动物料所需的时间量。

2. **流程导向的布局（Flow-oriented layout）**。工厂的布局可以促进更频繁的转运。单元布置（cellular layout）的一种好处是工站紧邻从而使物料可以在其间易于移动。物料搬运系统（如，传送带、AGV）也能促进工站之间的小批量转运，即使这些工站之间的并不靠近。

3. **共享手推车（Cart sharing）**。对于多个并联机器生产同种产品的工站，共享送入和/或送出的手推车（或用于在工站之间转运加工任务的任何容器）可以削减在工站前后等待的 WIP 数量。例如，图 17.3 显示了 12 台机器供给不同数量的送出手推车的情形（我们没有明确地表示出送入手推车）。平均来说，关于等待送入下个工站的完工部件数量，1 辆车系统是 12 辆车系统的十二分之一。可是要注意，它假设两个系统中机器操作员花费同样的时间将完工部件搬到手推车中。如果由于地理原因，1 辆车系统的操作员要比 12 辆车系统的走得远，则共享手推车可能延长了有效加工时间。取决于不同的系统，手推车共享带来的周期时间减少可能会抵消产能下降带来的负面影响。可是一般来说，手推车共享仅适用于搬运时间和不方便性都很微弱的情况。对于图 17.3 中的 12 台机器，三辆车或四辆车的安排可能是最实际的选择。（598|599）

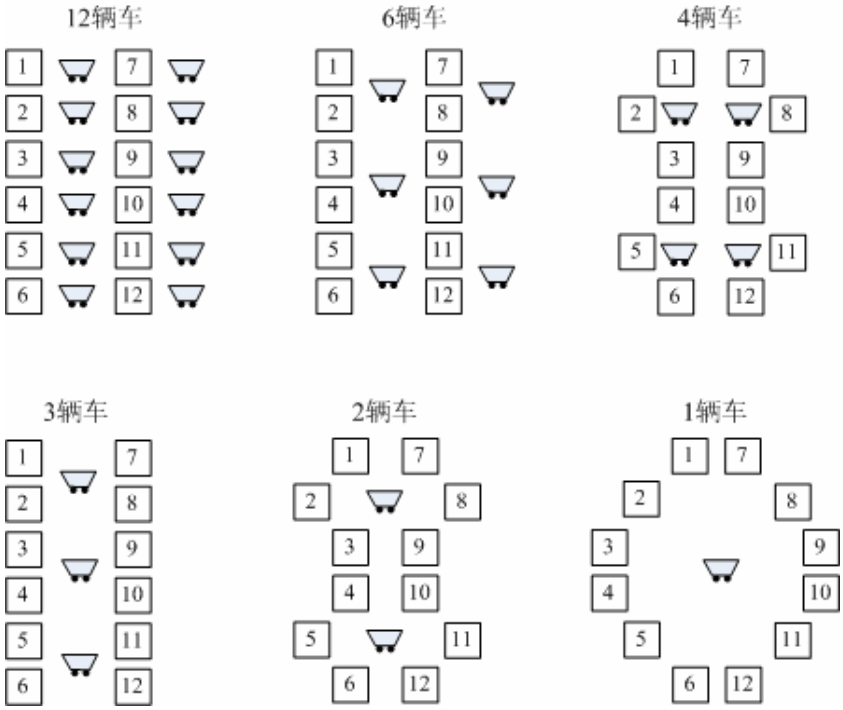

图 17.3 共用车辆安排

17.4.3 削减等待匹配 WIP

在组装工站处，只有所有子组件（subcomponent）都齐备时才能开始作业。我们已经在本章及第十二章中讨论了管理供给装配作业的采购件的问题，所以在这里只考虑子组件由工厂的不同产线自行生产的情况。

理想地，我们希望投放各子组件的加工单并在产线进行加工从而以与最终装配计划非常一致的方式，恰好同时到达装配工站。变动性一般会使这种情况无法发生，但我们可以做些事情来提高同步性。

1. **拉式体系（Pull systems）**。从第十四章可知，拉式体系特别是 CONWIP 可以自然地协调向产线投料与最终装配。如果产线的长度不同（即，穿越所需的时间不同），则需要不同的 WIP 水平（卡片数量）。它意味着同时投入产线的物料并不必要对应相同的最终产品。可是，如果产线 WIP 水平设置地合理，子组件到达装配作业将能同步。

2. **公共的作业积压单（Common work backlog）**。上述用于协调投料与最终装配的 CONWIP 方案只能在投料顺序不被产线作业打乱时让子组件同步到达装配工站。如果，比如说应用局部的分派准则如最短加工时间（SPT）于工站，加工任务可能互相超越并使得同步计划落空。即使我们在制造产线的工站处应用先进先出（FIFO）准则，多机工站处仍可能发生互超。因此，维持与最终装配计划同步的途径是在制造产线的各工站处遵循公共的**作业积压单（work backlog）**。这份清单简单地按最终装配计划序列列示加工任务。只要制造工站按积压单指定的顺序执行作业，加工任务就将同步到达装配工站。如果必须常常违反积压单（如，由于成批或质量问题），就不得不在装配工站前维持 WIP 缓冲来避免 “不同步（out-of-sync）”引起的停工。

3. **平衡分批（Balanced batching）**。如果制造产线由于长的换模而使用大的加工批量，则它可能无法与最终装配计划协调一致。解决这个问题的办法有三种。（1）在这条产线上超

前于最终装配计划作业，并在该产线与最终装配工站之间保持大量缓冲。（2）生成与制造产线成批需求一致的最终装配计划。（3）在制造产线处压缩换模时间或增加产能使较小的批量可行，并能与设定的最终装配计划协调。前两种是短期选项，第三种需要较多的时间来实施。（599|600）

17.5 管理 FGI

制成品库存在生产与需求之间的起缓冲作用。如前面提到的，这种缓冲被用于（1）将使客户与制造周期时间隔离开来，或许能提供“及时”交付，（2）吸收生产或需求过程中的变动性，或者（3）平滑产能负载（如，由于季节性）。这些意味着任何更紧密连接生产与需求的措施都可以使持有的 FGI 减少。可能的选择有：

1. **改善预测（Improved forecasting）**。尽管我们不会有将预测作为万能药这样不切实际的期望，但预测误差确实会膨胀 FGI。如果用更好的技术，如第十三章中的时间序列模型，预测需求可以降低生产与需求之间的差异，FGI 就将被削减。除了这一点，由于我们预测未来的能力有限，所以下面的其他选择可能对于大多数系统更有效。

2. **动态提前期提报（Dynamic lead time quoting）**。许多系统向客户提报固定的提前期。然而，由于工厂的负载随时间发生变化，实际的制造周期时间也随时间而变。因此，如果设定一个使按时交付的比例很合理的固定提前期，则高百分比的加工任务将早于它完成。如果不允许提起配送，这些加工任务将作为 FGI 等待。我们可以通过动态地提报敏感于工厂负载的客户提前期，来消除这个问题。

例如，我们供职于一家金属柜制造商，它在产品目录中公布了 10 周的固定提前期。如果使用了动态提前期提报系统，则在工厂几乎空闲时下单的客户将可能得到两周的提前期，而在工厂几乎爆满时下单的客户将可能得到 12 周的提前期。整体来说，平均提前期将缩短，并且实现同样的按时交付绩效将需要较少的产品作为 FGI 等待发货。

3. **缩短周期时间（Cycle time reduction）**。减少预测误差的一个非常有效的方法就是降低对预测的依赖。如果周期时间（包括接收订单、编码、工程设计、排程、制造产品、配送产品等等的整个增值链）能被压缩，作业投放就能更接近其交期。由于预测随时间向前越久越糟糕，较晚的投料能使主生产计划更为可靠。如果周期时间变得足够短，所有的投料都能结合固定的客户订单，因而预测误差导致的 FGI 可以被一体消除。值得高兴的是，前面列出的所有 WIP 削减技术也都是周期时间压缩技术（里特定律）并因而非常适合于这个目的。

4. **降低周期时间变动性（Cycle time variability reduction）**。第十二章指出，如果想确保某个服务水平，提报给客户的提前期受到周期时间均值与标准差的共同影响（见 图 12.9 和 12.10）。周期时间的变动性越高，提报中要加入的确保高百分比按时交付的安全提前期就越长。较高的安全提前期意味着除非允许提起配送，否则产品作为 FGI 等待的时间越长。幸运的是，我们可用于压缩平均周期时间的技术（减少换模时间、提高可靠性/可维护性、实施拉式机制、减少重工与报废）也有助于降低周期时间变动性。

5. **延迟差异化（Late customization）**。即使为了提供短的客户提前期而有必要持有库存，也不一定非得以 FGI 的形式。对于某些情形，有可能以半成品的方式存储产品并接单组装或进行定制。如果半成品可用于生产一种以上的成批，则半成品库存更有柔性，并有可能减少总库存持有量。600|601）

例如，一家水龙头支架（faucet fixture）制造商能提供 20 种不同的样品，而它们是五种基座与四种把手的组合。通过存储基座与把手，该制造商只需维持九种不同的物品，而不是

20 种。由于变动性汇聚，九种部件的需求预测比 20 件制成品容易，故而所需的总库存减少了。

再举一个例子，一家家电制造商生产一族有着不同附件（有或者没有面团钩（dough hook））、零售商店（标签和包装可能表示不同的商店品牌）和目标市场（说明书的语言可能不同）的电动搅拌机。通过储存通用的族类，并由塑料部件的颜色来区别，制造商能迅速贴标与包装以供应对不同制成品的需求。使用这种策略，预测只要在族的水平精确就行了，故预测误差引起的 FGI 能被显著削减。

这种策略的潜在缺点是（1）客户提前期没有以制成品形式存储时减少得多，当竞争对手那样做时就会带来问题，并且（2）半成品可能难以存贮；例如，如果搅拌器没有装箱，污垢和破损也许是个问题。

能以半成品形式存储可能也是产品设计的职责之一。例如，前面提到的金属柜制造商大部分部件有 10 周的提前期，因为产线很长而每件产品都必须从零开始（即，金属板）。通过在一小套有着不同油漆颜色、面框与特征（旋塞、电气连接、玻璃门，等等）的标准模块（存货）的基础上建立较短的产线，竞争者可以提供四周的提前期以满足客户需求。由于客户常常是建筑师，而他们又常常落后于进度，响应性在这个市场上就非常重要。竞争者由于精湛的产品设计策略而明显占了上风。

6. **平衡人力、产能与库存（Balancing labor, capacity, and inventory）**。对于许多市场，产品生产于需求的低谷期并作为 FGI 来满足高峰期的需求。这对某些情形可能是最佳选择，却绝不是解决季节性需求的唯一办法。另一种办法是改变劳力的规模，或者通过在高峰期使用零时工，或者将季节性偏移的产品（如，吹雪机和割草机）配对并在产线之间转移工人。还有一个对多数传统管理者来说比较异端的选择是，维持足够的超额产能以满足高峰需求而不建立库存。当考虑到预测误差带来的持有 FGI、废弃与低客户服务水平成本时，这些其他的选择可能比存储大量 FGI 更经济。至少，可以组合使用这些方法，如建立有限的存库，同时维持一些过量库存和一些浮动劳力。

17.6 管理备件

管理备件是整个维护政策（maintenance policy）的重要组成部分，而维护政策可能是制造系统运营效率的主要决定因素。由于它的重要性和复杂性，产业中存在着多样的备件管理实践（见 Cohen、Zheng 和 Agrawal 在 1994 年的标杆性研究）。我们并不试图对这些实践进行调查。相反，这一节中，我们要建立评估备件库存的框架，并在第二章模型的基础上开发合适的工具。（601|602）

17.6.1 需求分层

部件有两种不同的类型，一种用于计划内的**预防性维护（preventive maintenance）**，另一种用于计划外的**紧急修理（emergency repairs）**。例如，过滤器可能在月度保养中更换，而保险丝仅在损坏时才换。这两种零部件应当区别管理。

计划的维护针对的是可预测的需求来源。事实上，如果认真对待维护程序，这种需求比客户对成品的需求稳定得多。故而，标准的 MRP 逻辑适用于这些部件。也就是，始于计划的需求，减去当前库存（与计划接收量）并使用批量准则（批对批、固定数量等），来生成计划的产出量，并依据采购提前期向前偏移以生成采购订单。如果这些部件由内部生产，我们可用一些计划方法替代固定采购提前期以生成生产排程。对于任意一种情形，需求过程稳定、可预测是属性使得这些预防性维护部件相对易于管理。

计划外的紧急维修从定义上看就是不可预测的。因此，MRP 逻辑对这些部件的效果不好。我们在接下来的小节中给出维持足量安全库存以支持及时维修的办法。

17.6.2 为紧急修理存储备件

对于需求不可预测的部件，难题是以成本经济的方式提供高的服务水平。由于需求不确定，我们在第二章中讨论的（Q, r）模型成为检视这种权衡的潜在工具。为应用这个模型，我们必须确定如何表示多部件环境的服务水平。

备件系统中，服务水平与其支持的机器的可用性相关。此外，由于缺少\$2 的保险丝与缺少\$3,000 的计算机单元都会同样地使机器停机而变得不可用，所有部件的缺货成本都相等的假设常常是合理的。因此，如果知道了迟单成本（backorder cost）或脱销成本（stockout cost），就可以使用 2.4.3 小节中的模型单独地分析各部件。

但是如之前所见，迟单与脱销成本常常难易估计。对于备件系统，部件短缺的成本取决于它引起机器停机的成本，并进而取决于这次停机引起的客户服务延迟成本。由于这一点，我们常常以服务水平约束而非服务水平成本的形式考虑这个问题。幸运的是，成本与约束公式之间有着紧密的联系。

为使（Q, r）模型适应于多产品情形，我们使用与 2.4.3 小节同样的记号，但用下标 j 来表示部件 j 的参数，$j = 1, \ldots, N$。故有

N = 系统中不同部件类型的总数

D_j = 部件 j 每年的需求（件/年）

ℓ_j = 部件 j 的补货提前期（天）（602|603）

θ_j = 部件 j 在补货提前期内的期望需求（$\theta_j = D_j \ell_j / 365$）

σ_j = 部件 j 在补货提前期内需求的标准差

$p_j(x)$ = 部件 j 在补货提前期内需求为 x 的概率（概率质量函数（probability mass function））

$G_j(x)$ = $\sum_{y=0}^{x} p_j(x)$，部件 j 在补货提前期内需求少于或等于 x 的概率（累积分布函数（cumulative distribution function））

A = 任何部件每次换模或采购成本（美元）

c_j = 部件 j 单位生产或采购成本（\$/件）

h_j = 部件 j 年度单位持有成本（\$/（件·年））

k = 任意部件的脱销（缺货）成本（\$）

b = 部件的年度单位缺货成本（\$/（件·年））。注意惩罚未能用库存补给需求时用 k_j 或 b_j，而不同时使用。

B = 设定的总迟单水平

S = 设定的平均服务水平

F = 设定的平均下单频率

Q_j = 部件 j 的订购量（决策变量）

r_j = 部件 j 的再订购点（决策变量）

$F_j(Q_j)$ = 部件 j 的下单频率（每年的补货订单），为 Q_j 的函数

$S_j(Q_j, r_j)$ = 部件 j 的补给率（由库存补充的订单比例），为 Q_j 和 r_j 的函数

$B_j(Q_j, r_j)$ = 部件 j 的严重延迟（outstanding backorder）的平均数量，为 Q_j 和 r_j 的函数

$I_j(Q_j, r_j)$ = 部件 j 的平均持有库存水平（件），为 Q_j 和 r_j 的函数

通过这种标注，我们可以用两种方式表达总成本。我们在下面开发这两种模型及其附带的约束式。

迟单模型（Backorder Model）。我们始于以平均缺货水平刻画服务水平。可以定义成本函数为换模、缺货与持有成本之和

$$Y_b(Q,r)=\sum_{j=1}^{N}\left[AF_j(Q_j)+bB_j(Q_j,r_j)+h_jI_j(Q_j,r_j)\right] \qquad (17.6)$$

其中 **Q** =（Q_j, j = 1, …, N），**r** =（r_j, j = 1, …, N）表示订购数量和再订购点向量。由于成本函数 Y_b 是依赖于（Q_j, r_j）对的独立项的简单加和，我们可以通过最小化每个 j 的独立项来使其最小化。并且我们已在第二章这么做了。因此，利用在那里用过的同样的近似方法（即，通过基准库存缺货式子 $B_j(r_j)$ 来近似估计（Q, r）的缺货式子 $B_j(Q_j, r_j)$）来得到最优订购量和再订购点的相同表达式：（603|604）

$$Q_j^* = \sqrt{\frac{2AD_j}{h_j}} \qquad (17.7)$$

$$G(r_j^*) = \frac{b}{b+h_j} \qquad (17.8)$$

注意这些就是我们所熟悉的 EOQ 和基准库存公式。此外，如果我们假定部件 j 的提前期需求服从均值 θ_j 和标准差 σ_j 的正态分布，就可以将式（17.8）简化为

$$r_j^* = \theta_j + z_j\sigma_j \qquad (17.9)$$

其中 z_j 是标准正态分布表中对应于 $\Phi(z_j)=b/(b+h_j)$ 的值。

注意到这些 Q_j 和 r_j 表达式对部件之间的差别很敏感。例如，其他条件都相同时，高成

本部件（有较高的h_j系数）会比低成本部件有更小的订购量Q_j与再订购点r_j。此外，如可以预期的，Q_j和r_j随需求速率D_j增大。[7]对于需求服从正态分布的情形，如果z_j>0，再订购点r_j随提前期需求的标准差而增大；正如我们在第二章所见，只要$b>h_j$就是这样。最后，我们注意到提高固定订购成本A会增大所有的订购量Q_j，且提高缺货成本b会增大所有的再订购点r_j。

如果能为固定换模（订购）成本 A 和单位缺货惩罚 b 确定合理的值，就能利用（17.7）和（17.9）式来计算多产品（Q, r）系统的库存参数。但是，如在第二章所见，这在实践中常常难以做到。在生产环境中，由于成批采购的动力是避免频繁换模引起的产能损失，A 常用来代表产能。在采购环境中，产能不需要直接考虑，直接估计 A 就简单得多。但即使在这种情况下，要估计缺货成本 b 也很困难，因为它涉及给损失的客户喜好以及其他无形关系赋值。出于这个原因，使用约束模型常常更为直观。当服务水平合适地由严重延迟的总数（对于所有部件类型）刻画时，我们就可以为这个问题构模：

最小化　库存持有成本

约　束：　平均下单频率 $\le F$

总体缺货水平 $\le B$

我们可以使用像早些描述的用于多产品 EOQ 模型中的迭代程序来解决这个约束问题。基本思路是先调整固定订购成本 A 直至满足下单频率约束，然后调整缺货成本 b 直满足缺货水平约束。注意，当检验给定的（Q_j, r_j）值是否满足缺货水平约束时，我们用准确的（*exact*）公式来计算缺货水平，而不用那个用于推导（17.8）式的近似表达式。还有，因为缺货水平 $B_j(Q_j, r_j)$ 依赖于 Q_j 和 r_j，而下单频率 $F_j(Q_j) = D_j / Q_j$ 仅依赖于 Q_j，所以要先调整 A 再调整 b。我们会在下一页里正式地陈述该程序。（604|605）

算法（多产品（Q, r）迟单（backorder）模型）

步骤 0. 为 A 和 b 赋初值。

步骤 1. 代入 A 到（17.7）式计算批量 Q_j，$j = 1, \ldots, N$。

步骤 2. 计算这时的下单频率

$$F(A) = \frac{1}{N} \sum_{j=1}^{N} \frac{D_j}{Q_j}$$

步骤 3. 如果 $F(A) = F$，转到步骤 4。否则，

若 $F(A) < F$，减小 A

若 $F(A) > F$，增大 A

[7] 为了说明 r_j 随 D_j 增大，注意到增大的 D_j 增大了 θ_j，而由（17.9）式可知 r_j 随 θ_j 增大。

并转到步骤 1。

步骤 4. 代入 b 到（17.9）式计算再订购点 r_j，$j = 1, \ldots, N$。

步骤 5. 计算这时的总缺货水平

$$B(b) = \sum_{j=1}^{N} B_j(Q_j, r_j)$$

步骤 6. 如果 $B(b) = B$，结束。否则

若 $B(b) < B$，减小 b

若 $B(b) > B$，增大 b

并转到步骤 4。

脱销模型（Stockout model）。如果服务水平的刻画用平均补给率比用总缺货水平好，则可以定义另一个成本函数为换模、缺货与持有成本之和

$$Y_s(Q,r) = \sum_{j=1}^{N} \{AF_j(Q_j) + k[1 - S_j(Q_j, r_j)] + h_j I_j(Q_j, r_j)\} \qquad (17.10)$$

其中 $\mathbf{Q}$ =（$Q_j, j = 1, \ldots, N$），$\mathbf{r}$ =（$r_j, j = 1, \ldots, N$）是订购量和再订购点向量。如缺货成本模型那样，我们可以分别按每种部件 j 进行优化。使用我们在第二章中用过的相同近似法（即，用 EOQ 模型计算 Q_j，用第二类近似（type Ⅱ approximation）即 $S_j(Q_j, r_j) \approx 1 - B_j(r_j)/Q_j$ 估计补给率，用基准库存缺货表达式 $B_j(r_j)$ 估计缺货水平 $B_j(Q_r, r_j)$）得到相同的最优订购量与再订购点表达式：

$$Q_j^* = \sqrt{\frac{2AD_j}{h_j}} \qquad (17.11)$$

$$G(r_j^*) = \frac{kD_j}{kD_j + h_j Q_j} \qquad (17.12)$$

如果进一步假设产品 j 的提前期需求服从均值 θ_j 标准差 σ_j 的正态分布，就可将（17.12）式简化为

$$r_j^* = \theta_j + z_j \sigma_j \qquad (17.13)$$

其中 z_j 为标准正态表中对应于 $\Phi(z_j) = kD_j/(kD_j + h_j Q_j)$ 的值。（605|606）

正如在缺货模型中，这些 Q_j 和 r_j 表达式对部件之间的差别很敏感。再一次地，其他条件都相同时，高成本部件将比低成本部件有更小的订购量 Q_j 与再订购点 r_j。还有，Q_j 和 r_j

再次随需求速率 D_j 增大；以及对于正态分布情形，如果 z_j >0，再订购点 r_j 将随提前期需求的标准差而增大。最后，如可以预期的，提高固定订购成本 A 将增大所有的订购数量 Q_j，且提高缺货成本 k 将增大所有的再订购点 r_j。与缺货模型的一条区别是，r_j^* 的值依赖于 Q_j。

如果我们可以为固定换模(订购)成本 A 和单位缺货惩罚 k 确定合适的值，就能用(17.11)和（17.13）式来计算多产品（Q, r）系统的库存参数。如果出于第二章讨论过的原因，我们无法做到这一点，则可以使用约束优化方法。当服务水平由平均补给率合适地刻画，我们就可以为这个问题构模：

最小化　库存持有成本
约　束：　平均下单频率≤ F
　　平均补给率≥ S

我们可以使用类似于之前在缺货模型中用过的迭代程序。和那一样，我们用精确的（*exact*）式子表示补给率以检查补给率约束。再一次地，重要的是在调整 k 达到补给率约束之前，先调整 A 来满足下单频率约束。规范程序陈述如下：

算法（多产品（Q, r）脱销（stockout）模型）

步骤 0. 为 A 和 k 赋初值。

步骤 1. 代入 A 到（17.11）式计算批量 Q_j，$j = 1, \ldots, N$。

$$F(A) = \frac{1}{N}\sum_{j=1}^{N}\frac{D_j}{Q_j}$$

步骤 3. 如果 $F(A) = F$，转到步骤 4。否则，

若 $F(A) < F$，减小 A

若 $F(A) > F$，增大 A

并转到步骤 1。

步骤 4. 代入 k 到（17.13）式计算再订购点 r_j，$j = 1, \ldots, N$。

步骤 5. 计算此时的平均补给率

$$S(k) = \frac{\sum_{j=1}^{N} D_j S_j(Q_j, r_j)}{\sum_{j=1}^{N} D_j}$$

步骤 6. 如果 $S(k) = S$，结束。否则

若 $S(k) < S$，减小 k

若 $S(k) > S$，增大 k

并转到步骤 4。（606|607）

多产品（Q, r）例子。为说明迟单（backorder）与脱销（stockout）模型在多产品（Q, r）问题上的应用，以及二者的区别，我们来考虑表 17.3 给出的例子。该表分别给出了单位成本c_j、年度需求D_j、补给提前期ℓ_j以及提前期需求的均值θ_j和标准差σ_j。我们的目标是在平均下单频率以及平均补给率或平均缺货水平的约束下最小化平均库存投资。注意到由于用库存投资为目标，我们设定持有成本等于单位成本：$h_j = c_j$。

表 17.3 多部件（Q, r）例子的成本和需求数据

j	c_j （$/件）	D_j （件/年）	l_j （天）	θ_j （件）	σ_j （件）
1	100	1,000	60	164.4	12.8
2	10	1,000	30	82.2	9.1
3	100	100	100	27.4	5.2
4	10	100	15	4.1	2.0

首先我们来解决设定订购量Q_j的问题。为此，假设目标平均下单频率 F =12 次/年。注意单位成本与年度需求数据同表 17.1 一样。这样我们就解决了这个问题，因为多部件算法中计算Q_j的部分与多部件 EOQ 算法相同。从先前的例子中可知，选择固定订购成本 A = 65.138 可以得到能使平均下单频率达到 12 次/年的Q_j值。这些Q_j值记在表 17.4、17.5 中。

表 17.4 多部件脱销模型（Q, r）计算结果

j	Q_j （件）	$kD_j/(kD_j + h_j Q_j)$ （无量纲）	r_j （件）	F_j (下单频率)	S_j (补给率)	B_j （缺货水平）	I_j (库存水平,$)
1	36.1	0.666	169.9	27.7	0.922	0.544	2,410.66
2	114.1	0.863	92.1	8.8	0.995	0.022	670.24
3	11.4	0.387	25.9	8.8	0.749	0.918	512.52
4	36.1	0.666	5.0	2.8	0.988	0.014	189.33
				12.0	0.950	1.497	3,782.75

表 17.5 多部件迟单模型（Q, r）计算结果

j	$b_j/(b_j+h)$ （无量纲）	Q_j （件）	r_j （件）	F_j (下单频率)	S_j (补给率)	B_j （缺货水平）	I_j (库存水平,$)
1	0.538	36.1	165.6	27.7	0.875	0.974	2,024.77
2	0.921	114.1	95.0	8.8	0.997	0.010	698.76
3	0.538	11.4	27.9	8.8	0.840	0.511	671.85
4	0.921	36.1	7.0	2.8	0.998	0.002	209.10
				12.0	0.934	1.497	3,604.48

这样就只剩下计算再订购点 r_j 的问题了。我们始于目标平均补给率 S=0.95 的脱销模型。使用上述脱销模型算法，我们找到使平均补给率等于 95%的惩罚成本 k =7.213。表 17.4 记录了此时各部件的临界比率、再订购点、补给率、缺货水平以及库存水平。同时也计算出平均补给率（95%）、总缺货水平（1.497 单位）以及总库存投资（$3,782.75）。

注意到该算法为廉价、高需求量的部件 2 生成了非常高的补给率（99.5%），而为昂贵、需求量低的部件 3 生产了较低的补给率（74.9%）。直观地看来，该算法试图以尽可能廉价的方式实现 95%的平均补给率，所以它在能廉价地这么做或对整体均值有较大影响时提高服务水平。（607|608）

另一种用补给率刻画服务水平的方法是用迟单水平替代脱销水平。用迟单模型算法来调整缺货成本 b，直到总缺货水平达到既定目标。为比较脱销和迟单模型，我们使迟单的目标总缺货水平达到脱销模型时的水平，即 B=1.497 件。

在继续进行之前，我们要停下来说明：建立目标缺货水平并不总是一件容易的事。它不同于以无量纲的百分比表示的补给率，总缺货水平测度任何时候的平均显著延迟数量。因此，不能轻易地将一个系统的缺货水平移用到另一个系统（如，五件的平均缺货水平对部件少需求低的系统很可怕，却对部件多需求大的系统不算什么）。要更直观地理解这种缺货水平，可以从客户需求因迟单而产生的平均等待时间角度来考虑。若令 W 表示需求的平均等待时间，D 表示年需求总量，则由里特定律

$$B = D \times W$$

或

$$W = \frac{B}{D}$$

本例中，D=2,200 件/年，所以 1.497 件的缺货水平转化为

$$W = \frac{1.497}{2200} = 6.8045 \times 10^{-4} \text{年} = 5.96 \text{ 小时}$$

这意味着平均每种部件（任何部件，而不仅仅是遭遇缺货的那种）都将经历库存短缺导致的 5.96 小时的延迟。当然，它真正的意思是大多数部件不会遭遇延迟，而其他的会经历显著长于 5.96 小时的延迟。但关注各部件的平均延迟，可以使决策者了解到给定的缺货水平意味着多大的破坏（disruption）。事实上，延迟小时数与缺货水平作为算法的绩效目标是完全等效的——我们所做的，不过是是将缺货水平除以需求速率，再乘以一年的小时数。

现在，假设 1.497 件的目标缺货水平是合理的，我们可以用迟单算法来找出使总缺货水平达到这个值的缺货惩罚系数。结果表明 b=116.50。表 17.5 记录了此时各部件的临界比率、再订购点、补给率、缺货水平以及库存水平。同时也计算出平均补给率（93.4%）、总缺货水平（1.497 单位）以及总库存投资（$3,604.48）。（608|609）

注意到该算法的结果是廉价部件 2 和 4 的缺货水平低，而贵重部件 1 和 3 的缺货水平高。

此外，它趋于使需求较大部件的缺货水平较高（即，部件 1 的缺货水平高于部件 3，部件 2 高于部件 4），因为当其他条件相同时，需求较大的部件往往产生较高的缺货水平。如脱销模型那样，迟单模型将大量的库存投资置于贵重、高需求的部件 1。

但这两种解法之间有些关键的区别。注意到当总缺货水平一样时，这是我们强制的结果，补给率和库存水平都不同。迟单模型以较小的库存投资达到给定的缺货水平（$3,604.48 对 $3,782.75）。但它的代价是较低的补给率（93.4%对 95%）。如果用迟单模型来调整缺货成本 b 以使补给率等于 95%，将导致比脱销模型高的库存投资。结论就是，脱销模型找到一种高效使用库存来达到给定补给率的政策，而迟单模型找到一种高效使用库存来达到给定总缺货水平的政策。谢天谢地，这正是我们期待的结果。但由于这两种方法说明的是不同的权衡，我们一定要为特定情况选一个合适的。如果缺货水平（或时间延迟）较能代表服务水平，则迟单模型就更好。

最后，我们观察到可以用脱销或基准库存模型，生成在库存投资与补给率或缺货水平的权衡曲线。为此，需要简单地变更脱销成本 k 或缺货成本 b，并绘出各种库存投资与补给率（或缺货水平）对曲线。图 17.4 按一系列下单频率为前面的例子绘制了曲线。注意到，如所预期的，库存投资在补给率接近 100%时呈指数增长。此外，可以看到每年增加六次下单带来的收益随下单数的增加而减少。这些曲线表示各个下单频率/补给率对所需的最低库存投资，称为**效率边界（efficient frontier）**。管理者可以使用这类的图，来了解为实现各种服务水平都需要对库存投资多少。利用这个信息，他或她可以选择一个明智的补给率目标。类似的库存投资与补给率的曲线可以通过迟单模型来生成。（609|610）

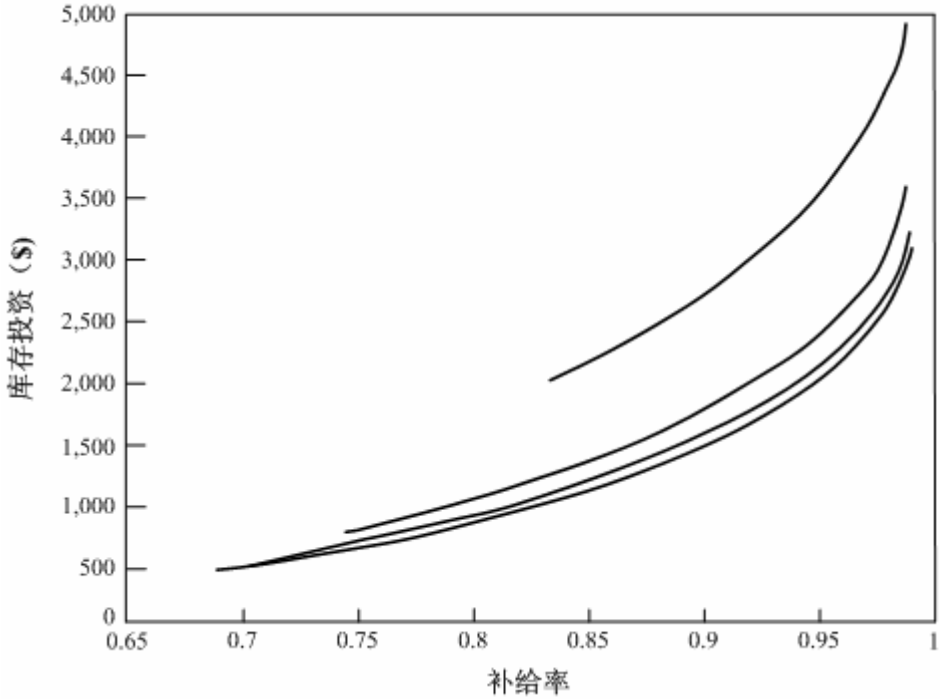

图 17.4 多部件（Q, r）模型中下单频率、补给率以及库存投资之间的权衡

17.7 多级供应链

许多供应链，包括供应备件的，除多部件外还包括多层级。例如，零售商可能将货物存储于区域性仓库，仓库供应商店，商店最终供应消费者。或者，对其产品提供服务合约的设

备制造商可能将备件存储于一个主要的配送中心，该中心供应区域性设施，而这些设施最终为客户设备的维护提供部件。由于变动性汇聚，在仓库或配送中心这样的中心地点存储，比在各个需求地点独立存储，所需持有的安全库存较少。然而，以分散的方式存储库存（如，在零售商店或服务设施）能因地缘接近而较快地对需求做出反应。多级供应链的基本挑战在于，平衡集中库存的效率与分散库存的响应性，从而在没有过量库存投资时实现系统高绩效。研究表明将单级方法直接应用于多级问题的效果很差（Hausman 和 Erkip 1994，Muckstadt 和 Thomas 1980）。这使我们将多级系统特别对待。

从分析角度看，多级供应链的复杂性和多样性使其非常具有挑战性。对这类系统的认真研究始于 Clark 和 Scarf（1960）的经典著作，并延续至今（见 Federgruen 1993，Axsater 1993，Nahmias 和 Smith 1992 的杰出调查，以及 Schwartz 1981 关于该主题的选集）。更近期的研究将多级库存管理置于供应链管理的背景中（见，如，Lee 和 Billington 1992，Fisher 1997，Simchi-Lev、Kaminsky 和 Simchi-Levi 1999）。由于在此无法给出全面的对策，我们将聚焦于定义问题，并指出某些较早的单级结果如何应用于多级系统的设定。

17.7.1 系统布局

多级供应链的特征是低级节点由高级节点供给。不过，这个框架之内还有许多可能的变种，并且如果我们允许同级节点之间的转运（如，区域性仓库可以彼此供应），则绝对的层级定义模糊起来。简而言之，多级系统可能非常复杂。

出于讨论的目的，我们将主要集中于各个库存点由单一源供给的**树状（arborescent）**系统（见图 17.5）。特别地，我们将考虑单一中心仓库（仓库、配送中心）供给多个零售商店（设施、需求地）的两级树状系统。这么做是因为（1）此类系统在现实中很普遍；（2）存在着对其行为进行近似表示的良好模型（见 Deuermeyer 和 Schwartz 1981，Sherbrooke 1992，Svoronos 和 Zipkin 1988）；（3）二级问题的解法可成为开发更复杂的多级系统解法的构建模块。（610|611）

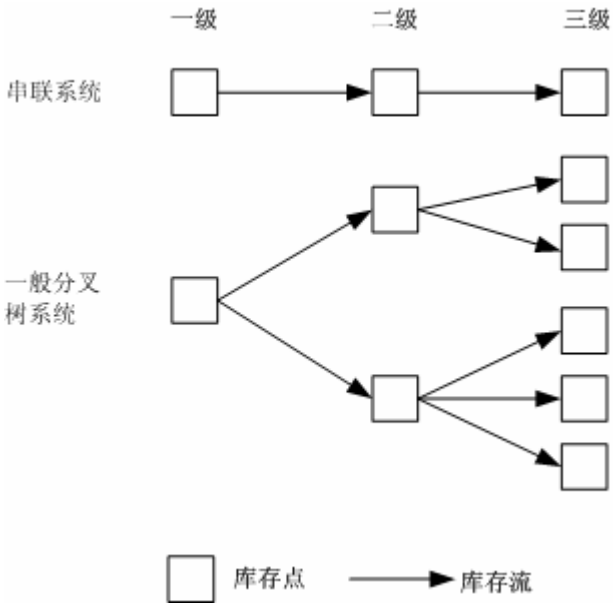

图 17.5 树状多级供应链

不过，在开始分析之前，有必要指出系统布局本身就是个决策变量。某个系统当前按三级树状结构布局，但它并不一定非要如此。事实上，确定库存层级数目、仓库位置以及互联

政策，可能是企业关于其配送系统要做的最重要的物流决策。即使这些系统提出了挑战性的问题，公开解决也比将现状视为不变而忽视显著机会的好。

作为这种反思系统布局的例子，我们给出所熟悉的一家设备制造商的案例。这家企业对其设备提供服务合约（如，保证机器每月的最长停机时间低于某个值），并存储备件以支持维护过程。这些部件存储于三个层级中：一个主要的配送中心，多个区域性设施以及多个客户地（针对有服务合约的客户）。差不多所有从配送中心到各设施的运输都是隔夜邮递（除了那些离配送中心非常近的设施，它们可以派维修人员直接去取部件）。维修人员从设施为作业现场补充库存。系统中大约一半的库存保持在配送中心，其余的在外地（即，设施与现场）。

这样的布局提出一个明显的问题。为什么要在配送中心存储部件？[8]某个设施可以隔夜收到部件，而从配送中心或其他设施都是等效的。（事实上，我们发现设施经理在配送中心缺货时，可以通过信息系统从其他设施获取隔夜邮递的部件。）因此，配送中心就可能将其库存划入各个设施。这样库存将更接近需求点，并因而使停机的客户等待关键部件隔夜到达的可能性降低。还有，当某个设施缺少部件时，假如系统中另个设施有这种部件存货，它就可以从其他设施而非配送中心隔夜获得。配送中心将停止作为物理存货点，转而成为逻辑采购机构（即，从供应商处采购部件或内部制造）并发挥协调职能（即，维护追踪系统中库存位置的信息系统）。净效果就是系统中的库存总量相同，而客户受到更好的维修服务。这种大胆的重构比对现有系统的细节优化产生更大的整体收益。（611|612）

17.7.2 绩效量度

为制定设计决策或开发模型，都要用具体项目表示想要考察的绩效。可以使用一系列的量度指标，包括：

1. **补给率（Fill rate）**是使用存货满足的需求的比例。这个概念可用于系统的任何层级。然而，必须记得应用于高层级（如，中心仓库）的量度仅仅是达到目标的手段。真正服务客户的低层级量度，才决定着系统的最终绩效。

2. **缺货水平（Backorder level）**是等待满足的订单的平均数量。这个量度用于会发生缺货的系统（如，不论需求发生时是否有部件存货但最终必须满足需求的一些备件系统）。如早先提到的，缺货水平与平均缺货延迟紧密相关，通过里特定律可得

$$\text{平均缺货延迟} = \text{平均缺货水平} \ / \ \text{平均需求速率}$$

例如，如果某部件的用量是 100 件/年，平均缺货水平是 1 件/年，则每个部件（所有部件，而不仅仅是那些发生缺货的）的平均延迟是 $\frac{1}{100}$ 年，即 3.65 天。

3. **损失的销售（Lost sales）**是由于缺货而损失的潜在订单的数量。这个量度用于客户不等待缺货品目而转向他处的系统（如，零售商店）。如果每个遇到缺货情况的需求都损失了，则每年的期望损失销售通过下式与补给率相连

$$\text{损失的销售} = (1 - \text{补给率}) \times \text{平均需求速率}$$

例如，若某给定部件的补给率是 95%，而需求是 100 件/年，则每年因缺货造成的损失为（1－0.95）（100）= 5 件。

4. **延迟的概率（Probability of delay）**指任务（如，机器修理、多部件订单的配送）由于缺乏库存而发生延迟的可能性。这个量度常用于需要高度可靠性的系统（如，飞机维护）。一般来说，多部件多层级系统的延迟概率是各个部件补给率的函数，尽管部件联合需求（如，用于同次修理或客户订单）的多样使得这种依赖性可能很复杂（见 Sherbrooke 1992 中更完整的讨论）。

[8] 我们要感谢 Yehuda Bassok 教授向我们指出这个“明显”的问题。

从以上讨论中可以得出结论，补给率与平均缺货水平是关键的量度，因为其他量度可以通过它们算出。出于这个原因，大多数数学模型要么直接使用这两个量度，要么使用一些依赖于它们的成本函数。（612|613）

17.7.3 牛鞭效应

多级供应链中出现的一个重要问题是**渠道合作（channel alignment）**。这是指各个层级之间的策略协调机制，涉及信息共享，库存控制，运输，以及其他管理决策。（612|613）因为存在那么多可能的决策变量，即使一家公司能控制供应链中的所有层级，渠道协调也很有挑战性。当这些层级由不同的公司组成时，问题会变得更让人望而却步。

面对复杂的多级供应链，自然的反应是将各种不同的层级看成相互独立的。即让每个层级使用本地的信息来实现本地的“最优”策略。事实上，当每个层级由独立的公司组成时，这种策略就是传统的默认状况。但真正要实施时，这种将各层级独立的方法会导致整条供应链的非常低的绩效。差劲的渠道协调导致的最明显的结果就是效率低下（如，库存会保持在一个低效率的数量和存储位置）。但更微妙、也同样具破坏力的结果就是**牛鞭效应（bullwhip effect）**，它指的是需求波动从供应链底部到顶部的扩增。

图 17.6 描述了牛鞭效应。即使供应链底部的需求（如，零售层）在时间轴上相对稳定，在供应链顶部（如，制造层）也是非常不稳定的。这种现象是 Forrester（1961）在一个工业动力学模型的案例研究中发现的。它也在麻省理工学院 20 世纪 60 年代开发的广为人知的啤酒游戏的一部分中作为行为背景而被注意。近来它也在实践中被观察到。例如，宝洁公司指出，帮宝适尿布的零售需求是非常稳定的，而分销商向制造商的订货却具有高度的变动性。类似的行为也在惠普打印机和礼来（Eli Lilly）生产的胰岛素的需求中被观察到。正如我们所知道的，变动性必须被缓冲——通过库存、产能或时间。因此，牛鞭效应会导致负面的效果，如过多的在制品，产能的不良利用，大量的客户订单积压，以及快速增长的成本。

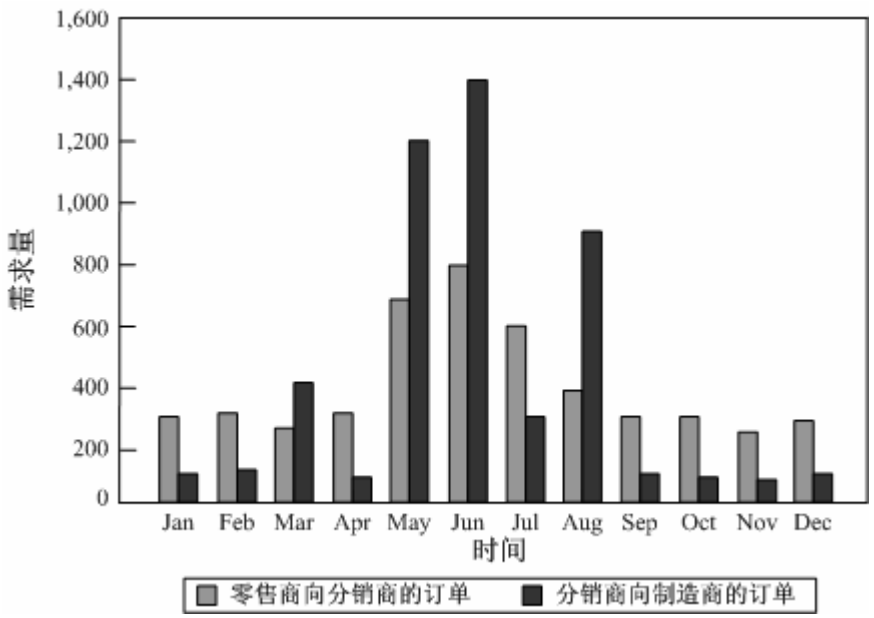

图 17.6 供应链不同层级的需求

既然牛鞭效应是客观存在的，那么关键的问题就是，它是由什么引起的？我们能对此做些什么？Lee、Padmanabhan 与 Whang（1997a，1997b）将引起牛鞭效应的原因分为四种。

在介绍他们的理论结构之后，我们将把这些进行总结并给出可能的补救措施。

批量（Batching）。在供应链的最低层（如，零售层）需求常常是稳定的，或至少是可预见的，因为采购总是以较少的数量进行。（613|614）举个例子，单个糖尿病患者通常会购买少量的胰岛素，足够应付几个星期或数月的需要。由于糖尿病患者独立做出自己的决策，总的零售需求在时间轴上是水平的。如果零售商在补充其库存时直接批对批地将订单传给分销商，分销商也同样将其订单传给制造商，这种平滑的趋势就将在供应链中得到保持。但是，如果零售商和分销商使用某种批量订货规则（如，他们遵照（Q, r）策略，就会一直等到他们的需求与批量Q的补充订单相符），那么他们的需求就比在零售层的需求具有更大的波动性。进一步地，如果某个给定层级上的决策者有数据同步（如，他们都在月初重新生成他们的MRP系统[9]），那么这种波动将会被进一步放大。

由于需求变动性的扩大是由批量订货引起的，推行以较少的数量补充库存的策略就会减小牛鞭效应。一些可能的选择是这样的：

1. *减少补充订单的费用*（*Reduce the cost of the replenishment order*）。我们从第二章中知道，采取批量订购的一个主要原因就是下采购订单的费用。一种降低该费用的方式是使用**电子数据交换（EDI）**来减少或取消采购订单。通过大幅度减少相关的文书工作数量，这种“无纸化”订购系统可以使少量多次的库存补充更容易地实现。

2. *合并订单以装满卡车*（*Consolidate the orders to fill the trucks*）。批量订购的另一个原因在于运输费用。批发商或分销商下的订单数刚好等于一辆卡车的满载量的情况并非罕见，这是因为满载运输的费用明显少于非满载运输。但是，一辆卡车并不需要都装同样的产品。所以减少订购数量而保持满载费用优势的一种方式，就是向同一个供应商订购多种产品。另外，货物补充过程也可以移交给第三方物流公司，这可以将多个供应商和/或多个客户的货物合并。任何一种情况都会使更频繁的运输都会变得容易。

预测（Forecasting）。在层级由独立的决策者管理的供应链中（如，他们由多个独立的公司组成），需求预测会改变订购的变动性。为了了解这种情况，我们假设零售商看到了需求的一个小峰值。因为订单必须满足预期的需求以及安全库存，这会导致一个比需求高峰更大的订单高峰。以零售商订单为基础预测需求的分销商看到这个峰值时，会将自己的安全库存增加到预期需求，并向制造商下一个更大的订单高峰。当零售商看到需求下降时，相反的情况产生了。因此，当我们在供应链上推进时，需求的波动性将增加。

加剧牛鞭效应的根本原因，是每个层级以它看到的需求，而不是以实际的客户需求为基础来更新它的预测。因此，以合并需求为基础的预测策略会减少牛鞭效应。一些可能的选择是这样的：

1. *共享需求数据*（*Share demand data*）。在多层级状况下，减少单独预测的扩大效果的简单补救措施，是使用一套通用的需求数据。在由一家企业控制的供应链中，从最低层共享需求数据，这从概念上来说是直截了当的（尽管与普遍实践相差甚远）。（614|615）在包括了许多家公司的供应链中，这就需要明确的合作。举个例子，IBM、惠普和苹果公司都把从经销商获得销售数据作为合同的一部分。在合作伙伴使用 EDI 的供应链中，信息共享在原则上是比较简单的；挑战在于获取合作伙伴的同意以实现信息共享。

2. *供应商管理库存*（*Vendor-managed inventory*）。确保利用低层级需求数据来预测的更积极主动的方式，是依靠一个单独的实体来做这件事。在**供应商管理库存（VMI）**系统中，

[9] 同步的 MRP 系统引起总需求在某些时候凸起的现象称为 **MRP 抖动（MRP jitter）**。

制造商在整个供应链上控制商品供应。例如，宝洁在从供应商（3M）到客户（沃尔玛）的所有途径上控制帮宝适的库存。使用 VMI 的联盟可以集中所有层级的库存，这使得他们在不协调的供应链中，可以用比所需数量少很多的库存来运作。

3. 提前期压缩（*Lead time reduction*）。订单预测的放大效应是由进入系统的需求峰值对安全库存数量产生效果而导致的。但如我们在第二章看到的，安全库存量随着补货提前期增加而增加。因此，一个明显、但潜在意义重大的可以降低预测带来的需求波动的方法，就是缩短提前期。我们在 17.4 节中讨论的任何关于减少在制品/周期时间的效率提升方式，都可以运用在各种不同的层级上，来达到这一目的。

定价（Pricing）。另一个导致需求在供应链的更高层级“凝结”成峰值的因素是价格折扣。每当一种产品的价格较低，客户往往会由于促销价格而提前购买（例如，购买比实际需求更多的数量）。当价格恢复正常，客户便消费过剩的存货，因此订单会比正常的少。这样的结果就是一个变动的需求过程。

由于导致需求变动的是价格变量，显著的补救办法就是稳定价格。具体的能够支持更稳定价格的策略有

1. 每日低价（*Everyday low pricing*）。最直接的稳定价格的方式，是简单地减少或消除依赖于折扣的促销活动。在食品杂货业，几家生产商已建立统一的批发价格策略，并通过围绕“天天低价”或“超值价格”的营销活动来促进策略的实施。

2. 作业成本法（*Activity-based costing*）。传统的会计制度可能不会显示一些促销价格所产生的活动成本，例如，区域折扣导致零售商在一个地区大量购买，又将产品运输到别的地区去消费。作业成本法（ABC）系统可以解决库存、运输、处理等问题，因而在论证和实施每日低价策略上很有用。

博弈行为（Gaming Behavior）。最后的一个导致牛鞭效应的因素是一种行为模式，即客户以博弈的方式使用他们的订单。例如，假设一个供应商在短期内按照订单数量的比例分配一种产品，那么客户便有一个清晰的动机来夸大他们的订单以得到更多的产品。当供应的数量赶上需求数量时，客户会取消过多的订单，把满载的库存留给供应商。这种情况在 20 世纪 80 年代的计算机存储芯片市场发生了不止一次，缺货鼓励计算机制造商向几家供应商订购存储芯片，从第一家购买并交付订单，同时取消其余的订单。

在这里，根本的问题是当博弈行为存在时，客户的订单会给供应商提供关于实际需求的极坏的信息。减少博弈订单动机的可能选择包括以下这些：（615|616）

1. 根据过去的销量分配短缺商品（*Allocate the shortages according to past sales*）。如果面对一种短缺商品的供应商以历史需求而不是现在的订单为基础来分配其供应的商品，那么客户在短缺的情况下就没有夸大订单的动机。

2. 使用更严格的时栅（*Use more stringent time fencing*）。回顾第三章的内容，冻结区和时栅是一种用来限制或惩罚改变订单的客户的工具。如果客户不能随意取消订单，那么博弈策略会带来更高的成本。当然，一个供应商必须在响应客户服务和稳定需求之间决定一个合理平衡点。

3. 压缩提前期（*Reduce lead time*）。另一种会导致博弈行为产生的情况，是当产品涉及较长的提前期的部件时。举例来说，我们在印刷电路板（PCB）工厂工作，该厂供应计算机组装（装箱）设备。为组装电路板，印刷电路板工厂需要采购未加工的卡和元件来装配到电路板上。一些元件有非常长的获取提前期，长达一年或更久。为鼓励其客户早一些传达他们

的需求，印刷电路板工厂有一系列的时栅，在所需的到期日之前以不同的提前期限制订单数量和类型的改变。但是，由于企业知道提前期长的部件在需求增长时会很难获取，客户就会有强烈的过高估计他们需求的动机。确实，当我们检查数据时，每次都发现需求在每个时栅处显著下降（如，如果一个时栅允许减少 15%的订单数量而没有成本惩罚，那么当他们达到这个时栅时，许多订单就刚好减少了这个量）。结果是使得过多数量的长提前期部件进入到印刷电路板生产厂的库存中。如上讨论，一个补救办法是限制客户更改订单的能力。举个例子，如果印刷电路板生产厂有一个比所有部件提前期要长的冻结区，这样的博弈行为就不会发生。但当然，给客户强加一个一年的冻结区也是不合理的。因此，替代的办法就是缩短部件的提前期，这样客户就有更少的动机欺骗系统过多地订购这些部件。

最后，我们观察到减少所有导致牛鞭效应的因素的根本性策略，是消除供应链的所有层级。这正是戴尔电脑直销系统的做法，在系统中计算机由制造商销售给客户，而不通过经销商。这不但使戴尔能够直接接触客户需求的数据，还消除了整体库存水平，从而减少了库存成本。在 20 世纪 90 年代，戴尔成为美国最成功的企业之一，这个战略起到了主要的作用。

17.7.4 二级系统的近似表示

我们现在转向具体的供应链问题。考虑一个两级库存系统，它有一个仓库供给数家设施，各设施最终满足客户需求。假设仓库和设施都使用连续检查的库存控制策略，且仓库使用（Q, r）模型而设施使用基准库存模型（即，一次补充一件库存，从效果上相当于 $r = 1$ 的（Q, r）模型）。这种类型的系统适合于备件，因为配送的速度很重要而物品数量相对较小。因此，设施很可能频繁地从仓库接收部件，一次一件的补给方式是个实用的选择。这个假设不太适合零售系统，因为商店的补给不太频繁而数量较大要采用批量运输。有兴趣的读者可以参阅 Nahmias 和 Smith（1992）的著作来了解零售系统建模的细节。（616|617）

设施处一次一件的补给假设暗示设施处的需求直接传递到了仓库。这意味着如果各设施处的需求服从泊松分布，则仓库处的总需求也服从泊松分布。（回忆起第二章中所见，泊松分布常常是表示需求过程的合理建模假设。）这使我们可以采取下面的方法。首先用单级（Q, r）模型分析仓库，固定服务水平（补给率）后计算各部件的订购量与再订购点。然后计算各部件各时点的期望延迟数目，并用它估计来自设施的订单将经历的延迟。用这个值，我们估计设施的提前期，也就是从仓库的实际运送时间与这个延迟之和的期望值。然后，用这些修正的提前期，我们对各设施应用基准库存模型来计算各部件的再订购点。

为建立模型，我们使用以下的记号。它们类似于前面多产品（Q, r）模型中的，并增加了下标 m 来表示设施：

N = 系统中不同部件类型的总数

M = 由仓库服务的设施的数目

$D_j = \sum_{m=1}^{M} D_{jm}$，仓库处部件 j 的年度需求（件/年）

ℓ_j = 仓库处部件 j 的补货提前期（天），假设为常量

θ_j = 部件 j 在补货提前期里的期望需求（$\theta_j = D_j \ell_j / 365$）

$p_j(x)$ = 仓库处部件 j 在补货提前期内需求小于或等于 x 的概率（概率质量函数）

$G_j(x)$ = $\sum_{y=0}^{x} p_j(y)$，仓库处部件 j 的需求在补货提前期里小于或等于 x 的概率（累计分布函数）

W_j = 仓库处部件 j 因延迟而产生的期望等待时间

D_{jm} = 设施 m 处部件 j 年度的需求（件/年）

ℓ_{jm} = 设施 m 从仓库获取部件 j 的提前期（天），假设为常量

θ_{jm} = 部件 j 在补货提前期内的期望需求（$\theta_j = D_j \ell_j / 365$）

$p_{jm}(x)$ = 设施 m 处部件 j 在补货提前期内的需求恰等于 x 的概率（概率质量函数）

$G_{jm}(x)$ = $\sum_{y=0}^{x} p_j(y)$，设施 m 处部件 j 在补货提前期内的需求小于或等于 x 的概率（累积分布函数）

L_{jm} = 设施 m 对部件 j 的订单由仓库满足的提前期（包括缺货延迟），是个随机变量

c_j = 部件 j 的单位成本（美元）

Q_j = 仓库对部件 j 的订购量（决策变量）

r_j = 仓库对部件 j 的再订购点（决策变量）

r_{jm} = 设施 m 对部件 j 的再订购点（决策变量）（617|618）

R_{jm} = $r_{jm}+1$，设施 m 处部件 j 的基准库存水平（决策变量，等价于 r_{jm}）

$F_j(Q_j)$ = 仓库对部件 j 的下单频率（每年的补货订单数），为 Q_j 的函数

$S_j(Q_j, r_j)$ = 仓库处部件 j 的补给率（由存货满足的订单的比例），为 Q_j 和 r_j 的函数

$B_j(Q_j, r_j)$ = 仓库处部件 j 显著缺货（outstanding backorder）的均值，为 Q_j 和 r_j 的函数

$I_j(Q_j, r_j)$ = 仓库处部件 j 的平均持有库存水平（件），为 Q_j 和 r_j 的函数

仓库层级（Warehouse Level）。我们可用早先单级问题提供的任何一种方法求解仓库问题（即，计算部件的 Q_j 和 r_j）。也就是，我们可以用成本模型，指定固定订购成本 A 以及缺货成本 b 或脱销成本 k。还可以用约束模型，设定对每年下单数 F 以及补给率 S 或平均缺货水平 B 的约束。通常情况下，由于仓库持有库存的目的在于最小化设施处（进而还有客户）的缺货，使用基于缺货成本或约束的模型比基于补给率的模型要好。

不论用什么模型，输出都将是一系列Q_j和r_j值，再用它们以及第二章中开发的函数计算各部件的F_j、S_j、B_j和I_j，$j=1,\ldots,N$。然后用这些值作为设施水平计算的输入。

设施层级（Facility Level）。注意到从设施处来的订单由于缺货而要在仓库处等待的时间（以天计）为

$$W_j = \frac{365B_k(Q_j, r_j)}{D_j} \qquad (17.14)$$

注意，这只是里特定律在缺货上的应用（即，等待类似于周期时间，缺货水平类似于 WIP，需求速率类似于产出）。因此我们可以估计设施 m 处部件 j 的平均有效提前期（天）为

$$E\left[L_{jm}\right] = \ell_{jm} + W_j \qquad (17.15)$$

我们可以把这个平均提前期视为常量，并将其代入基准库存模型计算设施的绩效量度。事实上，研究者已经证明，将这些提前期按其均值处理（即，L_j），结果也是合理的（见 Sherbrooke 1992）。然而，很明显L_{jm}是个能表现出大量变动性的随机变量。从设施处来的订单发现仓库有存货可用时，$L_{jm}=\ell_{jm}$。但是仓库缺货时，L_{jm}就比ℓ_{jm}长得多了。设施处有效提前期的计算很复杂（见 de Kok 1993）。但我们可以用一种近似方法将提前期变动性的影响纳入考虑。（618|619）

技术性注释

为近似表示从设施到仓库的订单的有效提前期方差，假设仅有以下两种可能性：订单没有延迟且提前期为ℓ_j，或遭遇缺货延迟且提前期为$\ell_{jm}+y$，其中 y 是一个确定的延迟常量。由于缺货的概率为$1-S_j$（为表示的方便，我们忽略S_j和B_j对Q_j和r_j的相关性）

$$E\left[L_{jm}\right] = S_j\ell_{jm} + \left(1-S_j\right)\left(\ell_{jm}+y\right) = \ell_{jm} + \left(1-S_j\right)\left(y\right) \qquad (17.16)$$

但为了使该式与（17.15）式相一致，须有

$$y = \frac{W_j}{1-S_j} \qquad (17.17)$$

为计算L_{jm}的方差，首先计算

$$E\left[L_{jm}^2\right] = S_j\ell_{jm}^2 + \left(1-S_j\right)\left(\ell_{jm}+y\right)^2 \qquad (17.18)$$

则

$$Var\left(L_{jm}\right) = E\left[L_{jm}^2\right] - E\left[L_{jm}\right]^2$$

$$= S_j\left(1-S_j\right)y^2$$

$$= \frac{S_j}{1-S_j}W_j^2 \qquad (17.19)$$

所以，设施处有效提前期的标准差（天）近似等于

$$\sigma\left(L_{jm}\right)=\sqrt{\frac{S_j}{1-S_j}}W_j \qquad (17.20)$$

我们可以在基准库存模型中使用$E\left[L_{jm}\right]$和$\sigma\left(L_{jm}\right)$，来计算部件 j 在设施 m 处的基准库存水平R_{jm}。

层级整合（Integrating Levels）。要协调两个层级，必须解决两个问题：仓库处选用的模型以及该模型中的参数。一旦选定了这些值，以上的设施层级建模方法可相应地用于调整设施处的基准库存水平。

在多级备件供应链中，仓库层级最自然的模型是缺货模型。原因在于，向客户的服务水平与部件短缺引起的延迟紧密相关。因此，仓库处服务水平的关键指标是时间延迟，我们已看到它与缺货水平成正比。这样，仓库处合乎逻辑的选择是有缺货水平约束的缺货（Q, r）模型。我们可以用之前描述的算法来计算仓库处的订购量Q_j与再订购点r_j。等效地，我们也可选用以缺货成本 b 替代对缺货水平的约束的缺货模型。但是，设定目标缺货水平约束（或时间延迟）常常比指定缺货成本更为直观。

在其他多级供应链中，如零售系统，客户服务水平由补给率来测度将更合适。例如，若不能由仓库即使满足的订单要么损失要么分流到（较昂贵的）第三方，则补给率作为仓库服务水平的量度就非常合理。但是，我们需要修改模型，来应对损失的销售或提前期对仓库服务水平的不同依赖性问题。

一旦有了应用于仓库层级的模型，我们还需要为它指定参数。如果使用约束的缺货模型，则关键决策就是以什么作为目标下单批量 F 和目标缺货水平 B。可以通过考虑仓库采购系统的能力进而有它每年能处理的补给订单数目，来直接选取目标下单频率。或者，我们可以指定固定下单成本 A，并将其代入多级 EOQ 的（17.7）式来计算订购量。（619|620）

选择目标缺货水平则较为困难。仓库处允许多少缺货，取决于这对设施绩效的影响。因此，几乎不可能先验地指定目标缺货水平。相反，我们应该做的是将目标缺货水平作为一个变量，可以调整它来找到系统的最佳整体绩效。具体来说，我们利用一个给定的目标缺货水平来对仓库层级求解。然后对设施层级求解，以在各设施处实现上述目标缺货水平或补给率，并观察库存持有成本（或投资）。最后返回在仓库层级尝试一个不同的目标缺货水平，再对两个层级求解从而观察是否能以较低的库存成本实现设施处相同的绩效。变更目标缺货水平将移动仓库库存与设施库存之间的平衡。对实现最优平衡的目标缺货水平的寻找，可以用电子表格或其他程序自动完成。

例子：

我们以一个二级供应链例子来结束本节。因为目的是强调层级之间的关系，我们将以单一部件为例，对情况进行简化。

假设我们为维修部经理 Jack 解决的问题（第二章，表 2.6）实际上代表着一个两级供应链中的仓库。Jack 在仓库中存储备件以供应各个地区性设施，再由设施向实际的机器修理作业提供部件。因为这是个单部件例子，我们省略下标 j。仓库的关键数据为 $D = 14$ 件/年，$Q = 4$，$r = 3$。回忆起第二章中我们使用缺货成本模型（假定固定采购成本 $A = \$15$，缺货成本 $b = \$100$）计算出订购量 $Q = 4$，再订购点 $r = 3$。我们也可以很容易地构建一个对下单频率 F 和缺货水平 B 进行限制的约束模型。

现在让我们通过一个 $D_m = 7$ 的设施来扩展这个例子（即，从仓库角度来看，该设施占其年度需求的一半）。从第二章中的计算，可知 $B(4,3) = 0.0142$ 件，故由于库存短缺造成的补充订单平均等待时间为

$$W = \frac{365B(4,3)}{D} = \frac{365(0.0142)}{14} = 0.3702 \text{ 天}$$

假设从仓库接收部件的实际运输时间是一天，则部件的期望提前期为

$$E[L_m] = 1 + 0.3702 = 1.3702 \text{ 天}$$

故补货提前期内对设施的需求期望为

$$\theta_m = \frac{1.3702 \times 7}{365} = 0.0263 \text{ 件}$$

同样从第二章的计算，可知补给率 $S(4,3) = 0.965$。因此，补货提前期的标准差为（620|621）

$$\sigma(L_m) = \sqrt{\frac{S}{1-S}}W = \sqrt{\frac{0.965}{1-0.965}}(1.3702) = 1.944 \text{ 天}$$

假定设施处的需求服从泊松分布，我们可以用（2.58）式来计算提前期需求的标准差

$$\sigma_m = \sqrt{\theta_m + \left(\frac{D_m}{365}\right)^2 \sigma(L_m)^2} = \sqrt{0.0263 + \left(\frac{7}{365}\right)^2 (1.944)^2} = 0.166 \text{ 件}$$

注意到该例中 $\sigma_m = 0.166$ 非常接近 $\sqrt{\theta_m} = \sqrt{0.0263} = 0.162$。原因在于，（2.58）式中的膨胀因子（inflation factor）相对较小。这表明提前期需求非常接近泊松分布。因此，我们可用泊松分布的公式来近似估计各种基准库存水平引起的服务水平。[10]例如，若令设施处的再订购点等于 $r_m = 0$，则补给率是

$$G_m(r_m) = \sum_{y=0}^{r_m} p(y) = p(0)$$

$$= \frac{\theta_m^0 e^{-\theta_m}}{0!} = e^{-0.0263}$$

[10] 由于实际变动性比泊松分布稍大些，实际服务水平将比泊松分布方法预测的稍低些。

$= 0.974$

若将再订购点增至 $r_m = 1$，则服务水平将增至 0.997。所以，取决于该部件在设施中的临界性，再订购点为零或一都很合适（Depending on the criticality of the part at the facility, it looks as if a reorder point of zero or one will be appropriate.）。

17.8 结论

库存管理如制造业本身一样久远。库存控制的分析方法可追溯至科学管理时代（即，二十世纪早期）且是运筹学/管理科学最早的例子之一。尽管如此，该领域仍在不断发展。就连如 EOQ 和（Q, r）模型这般古老的技术，也在经历着突破（如，新算法及其在多级供应链中的应用）。因此，对库存与供应链管理下结论还为时尚早。本章提出的模型为某些情形提供了解决方法，但适用于新情形的更好方法与扩展无疑会继续出现。这意味着库存将成为不断发生持续改善的领域，并且制造管理者需要不断学习该领域内的新技术。

同时，以下提示值得一记：

1. *理解为何持有库存*（*Understand why inventory is being held*）。不同类型的库存因不同的原因而持有，有些是有意识的而有些无意识。严格追问为何某个系统内持有各种库存，可以揭示出那些被当作理所当然的低效。

2. *寻求结构变革*（*Look for structural changes*）。通过复杂的模型精调供应链也很好。但是，真正巨大的改进很可能需要结构变革。例如，策略从存储 FGI 转向存储半成品并接单生产，可能会对总库存投资产生巨大影响。类似地，取消中央仓库并将所有备件存储于区域性设施，可以在不增加库存的情况下带来客户服务中心的显著上升。可能进行的具体变更取决于系统本身。识别的关键在于，尽量少地将现状视为理所当然。（621|622）

3. *使用经验评估方法*（*Use empirical evaluation procedures*）。任何模型都建立在简化的假设（如，稳态、泊松型需求）之上，且输入数据充其量是近似估计。因此，分析过程能做的不过是帮助我们寻找一个合理的政策（想找“最优解”，办不到）并检视权衡。知道了这点，我们就应当为分析补上经验观察与反馈。我们应当监控的参数的例子有（1）服务水平，与我们模型预测的值比较，以确定是否需要政策变更；（2）原材料与 FGI 的最低库存水平和缺货频率，以确定持有的安全库存是不足还是过量；（3）关键工站处的队列长度和饥饿时间，以检测 WIP 是不足还是过量。重要的是，确定一些关键量度并为之建立良好的数据收集与解释系统。

4. *压缩周期时间是至关重要的*（*Cycle time reduction is crucial*）。里特定律告诉我们，有 WIP 就会有周期时间。所以削减 WIP 和压缩周期时间实际上是同义的。但更重要的是，压缩周期时间可能降低部件采购与作业排程对远期预测的依赖。净效果就是较低的 WIP 水平，以及较低的原材料和 FGI 水平。

5. *协调多级供应链的各层级*（*Coordinate levels in multiechelon supply chain*）。当存货保持于多层级时，库存管理变得很复杂。除了有效管理各层级，一定是要保证各层级绩效支持整个系统的效率。牛鞭效应就是各独立层级的短视控制导致整个系统出现巨大问题的鲜明例子。为避免这种现象，一定要将供应链当作整体而非单独的部分来分析，尽可能分享公共数据（如，零售需求数据），以及简化供应链来避免不必要的复杂性。

6. *协调激励系统与运营目标*（*Coordinate incentive systems with objectives*）。建立一个有着具体绩效目标的库存管理系统是很好的。然而，任何此类系统都依赖于使它运行的人。因

此，如果薪酬结构不支持系统目标，它就不可能运行。（回忆起人力定律：人，而非组织，是追求自身利益最大化的。）例如，我们最近为一家有多级供应链的企业工作。其中的设施主要按客户服务水平来评估，但以库存效率的名义也会每月审计一次各自的库存水平。可以预见，设施管理者有整月都囤积库存（即，比建议的数量多）的趋势。就在月底审查之前，他们会把过量库存送回配送中心。一旦审查结束，他们又会下单补到原先的“过多”水平。其效果是破坏了库存与服务水平之间的所有平衡。显然，没有任何模型或分析能纠正这个问题。只有修改设施评价方法（如，使用结合服务水平与库存水平的评级制，其中库存以美元为单位连续或随机地测评）才能合理化设施处的库存水平。