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

第九章 变动性的腐蚀性作用

当幸运女神降临在身边时，不需思考就可以成功。
——乔尔丹诺•布鲁诺，于 1600 年被烧死在火刑柱上

知道得越多，就会变得越幸运。
——达拉斯的 J. R. Ewing

9.1 引言

前一章里我们开发了刻画、评估加工时间变动性与流动变动性的工具。本章中，我们将使用这些工具描述有变动性的制造系统的基本行为。

像在第七章那样，我们将主要结论陈述为工厂物理学定律。这些定律有些是一直成立（如，物料守恒定律），而其他的则在大多数情况下（*most of the time*）成立。从表面上看，似乎不够科学。然而，我们指出物理学定律，如牛顿第二定律 $F = ma$ 和能量守恒定律，也只是近似地成立。尽管已经被更为深奥的量子力学和相对论代替，它们依然很有用。工厂物理学定律也是如此。

9.1.1 变动性可能是好的吗？

第七章和第八章（与这一章的标题）可能会给人留下变动性是有害的印象。利用精益制造（Womack 和 Jones 1996）的行话，一个人可能会受到吸引将变动性和*muda*（浪费）[1] 等同起来并做出变动性应该被清除的结论。

但是我们必须仔细而不至于忘记公司的基本目标。就像我们在第一章里所观察到的，亨利·福特是一个对于减少变动性极度极度狂热的人。顾客可以得到他想要的任何颜色的车，只要是黑色（*black*）的。（287|288）车型罕有变化，型号也很少。通过稳定产品和保持生产过程的简单有效，福特开创了一次重大的革命——使汽车量产成为可行的。但是，在二十世纪三四十年代，当阿尔弗雷德·P·斯隆的通用汽车提供了更多的产品型号时，福特汽车公司失去了大量市场占有率，濒临破产。当然，更多的产品型号意味着通用汽车的生产系统中更大的变动性，更大的变动性意味着通用汽车的系统不能运行得像福特那样有效率。然而，通用汽车做得比福特好。为什么？

答案很简单。无论通用还是福特都没有着力去减少变动性甚至 *muda*。他们着力于获取长期投资的高回报（*make a good return on investment over the long term*）。增多产品型号会增加变动性因而增加 *muda*，但这也在一定程度上增加了收入；如果收入的增加多于需要抵消的额外成本，那么它也可以是一种合理的策略。

9.1.2 好的变动性与坏的变动性的例子

为了强调变动性可能是好的（一种商业策略上的必要调整）或坏的（一种恶劣的运营方

[1] Muda 在日语中是“浪费”的意思，并且它被定义为“任何吸收资源却不创造价值的人类活动”。Ohno 举了七个 *muda* 的例子:产品中的缺陷、产品的过量生产、等待进一步处理或消耗的产品库存、不必要的处理、不必要的移动、不必要的运输以及等待。

针的不合需要的副作用）方式，我们来看一些例子。

表 9.1 列示了产生不合需要的变动性的几种原因。举例而言，在第八章中我们看到，计划外的储运损耗，例如机器崩溃，可以导致系统中巨大的变动性。当这样的变动性成为不可避免的，它就不是我们能够着意引入系统的了。

作为对比，表 9.2 给出了一些有意将变动性引入系统的有效整体策略的例子。像我们上面提到的那样，二十世纪三四十年代通用汽车的变动性是更多的产品类型所导致的结果。在八九十年代的英特尔，变动性是技术不断变化的环境中产品迅速革新的结果。通过大胆地在上一代微处理器的生产过程稳定之前就推出下一代产品，英特尔刺激了对新计算机的需求，并为竞争者的进入创造了有利的壁垒。在以提供立等可取（while-you-wait）换油服务为企业核心商业策略的壳牌润滑油，需求的变动性是不可避免的结果。壳牌润滑油可以像在传统的汽车厂里一样通过制定计划减少这个变动性，但是这样做会使公司丧失竞争优势。

表 9.1 坏的变动性的例子

原因	例子
计划的断供	换模
计划外断供	机器故障
质量问题	产出损失和重工
工人变动	技能差别
设计不合格	工程变更

表 9.2 （潜在）好的变动性的例子

原因	例子
产品多样性	20世纪30至40年代的通用汽车（GM）
技术变革	20世纪80至90年代的英特尔（INTEL）
需求不确定	壳牌润滑油（Jiffy Lube）

无论变动性在商业策略中是好是坏，它都会导致运营中的问题并且必须被管理。处理变动性的特定策略应取决于系统的结构和公司的战略目标。（288|289）

在本章中，我们呈现的定律管理了多样性影响生产制造系统行为的方式。这些确定的要点权衡那些在开发有效操作必须要面对的定律。

9.2 绩效与变动性

由第六章系统分析的术语来讲，管理任何一个系统都始于一个**目标（objective）**。为了实现这个目标，决策制定者实施**控制（control）**并用各种**量度（measure）**来评估绩效。例如，一趟航班的目标是安全和适时地将乘客从 A 地送到 B 地。为了实现这一点，飞行员在监控飞机表现的各种量度的同时还采取了许多控制措施。控制与量度的联系可以通过航空工程学来建立。类似地，工厂经理的目标是通过高效地将原材料转化成将售的货物从而增加企业的长期获利性。就像飞行员，工厂经理也有许多控制措施和量度要考虑。工厂物理学的首要目标就是理解制造经理们面对的各种控制与量度之间的关系。

生产系统中关乎控制如何影响量度的一个核心概念是变动性。如在第七章所见，产线的最佳情形发生于无变动性时，而最差情形发生于最大量的变动性时。在第八章我们观察到，

工站的几种重要量度，如周期时间和在制品（WIP），都是变动性的增函数。

要理解变动性如何影响比第七章的理想产线或第八章的单工站更一般化的生产系统的绩效，我们需要在绩效的定义方式上更加精确。我们将进行这个主题，首先讨论生产系统的完美（*perfect*）绩效。然后，通过观察这个绩效可以分解出的维度，我们定义一组量度。最后，依据制造环境和企业业务战略来讨论这些量度的相对权重。

9.2.1 制造绩效的量度

任何人如果看过飞机的驾驶舱就会知道，飞机的性能并不是通过单一的指标来评价的。一大堆给人留下了深刻的印象测量仪、刻度盘、仪表、LED 读数器等都证明，尽管目标并不困难（从 A 点飞至 B 点），度量工作却不简单。高度、方向、推力、风速、对地速度、升降舵设置、引擎温度都必须小心地监视以保证目标的实现。

依照同样的方式，一个生产企业也有一个相当简单的基本目标（赚钱）。但却有许多潜在绩效衡量指标如产出、库存、客户服务水平及质量等等（见图 9.1）。对绩效衡量指标的适当的数字化定义取决于工作所处的环境。例如，一家生产聚苯乙烯的工厂能够从每磅的直接产出测量每天的产量。

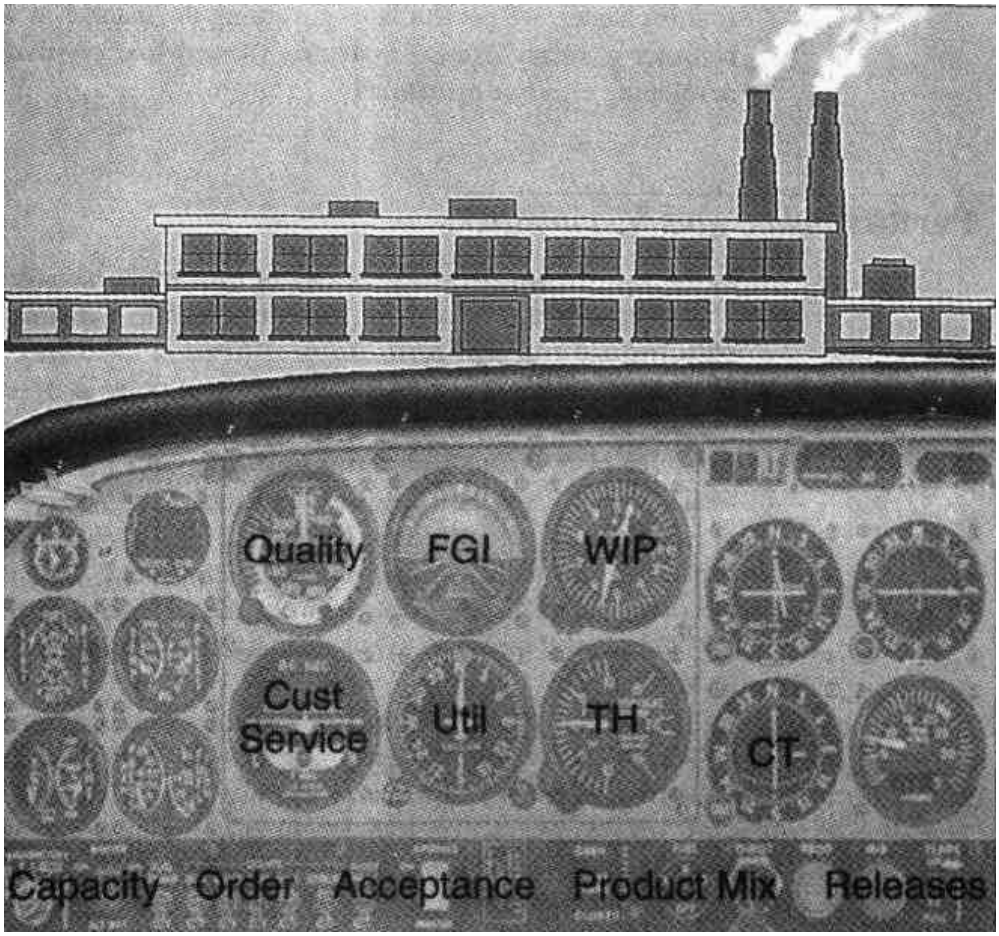

图 9.1 制造控制的仪表盘

一家播种机制造商（播种机是一种由拖拉机牵引能够种植植物并对其进行施肥的设备，每次播种至少 4 行至多 30 行）可能并不想以播种机每天播种几磅作为单位测量产出，原因在于不同尺寸的播种机之间存在着显著的多样性。用每天播种的行数为单位衡量产出可能是

一种能更好合计产出的方法。事实上，在多种产品复杂物流的系统中，产出是用美元/天来计算的，以便将产出汇总到一个简单的数字。（289|290）

绩效衡量指标的相对重要性也依赖于特定系统和它的经营战略。例如，联邦快递，它的竞争优势就在于运输速度和定位能力。因此它就把很大的权重放在反应时间（提前期）和客户服务水平（及时送达）上。相反，美国邮政总局主要在价格上与之竞争，所以强调如设备的利用率和材料的承载量等成本控制方法。即使这两个企业都是承担包裹运输业务，而他们有不同的经营战略，针对不同的细分市场，所以需要不同的指标。

由于生产环境与竞争战略的广泛性，我们不可能制定一套适用于所有制造系统的绩效衡量指标。但是了解指标的类型和它们是如何与多样性联系是有可能的，考虑简单单一产品生产线绩效是很有用的。原则上，更为复杂的多种产品生产线的衡量指标能够由单一产品生产线延伸出来，由多生产线组成的系统的指标可以由每条生产线方法的集合得到。

第七章中使用了产出、周期时间和在制品去刻画一个简单的流水线的工作情况。很明显，这些都是重要的衡量指标，但是他们还不够全面。由于成本的原因，我们也必须考虑到设备的利用率。既然生产线由采购流程供给，另一项重要指标是原材料库存。（290|291）当我们考虑到顾客，提前期、服务水平和最终产品库存就是相互关联起来。最后，由于产出损失和重工经常出现，质量也是一个关键的绩效指标。一条完美的单一产品流水线就是产出恰好等于需求，所有设备充分利用，平均周期时间和提前期都尽可能短。完美的客户服务（不延迟或积压客户订单工作），完美的质量（无废料或返工），零原材料原料或最终产品库存，并且有最小的在制品数量。

我们能更准确的用定量的有效值来刻画这些方法。对于一个完美生产表现的有效值，零却代表可能的最坏的生产状况。为了避免这个，我们要利用一下符号，为了更加精确起见，我们将以件为单位来计量库存和以天为单位来计量时间：

$r_e(i)$ = 工站 i 的有效速率，包含了停机、换模与作业员效率等扰动因素（件/天）

$r^*(i)$ = 工站 i 无扰动时的理想速率（件/天）

r_b = 产线包含扰动时的瓶颈速率（件/天）

r_b^* = 产线不包含扰动时的瓶颈速率（件/天）

T_0 = 包含扰动时的原始加工时间（天）

T_0^* = 不包含扰动时的原始加工时间（天）

W_0 = $r_b\, T_0$ = 包含扰动时的临界 WIP（件）

W_0^* = $r_b^*\, T_0^*$ = 不包含扰动时的临界 WIP（件）

D = 平均需求速率（件/天）

WIP = 产线平均在制品水平（件）

FGI = 平均制成品库存水平（件）

RMI = 平均原材料库存水平（件）

CT = 从制成品或线间缓冲的存货点投料开始计的平均周期时间（天）

LT = 提报给客户的提前期（天）；在提前期固定的系统，$LT = 0$；在向客户分别提报的系统，这个符号代表均值

TH = 由产线产出速率给出的平均产出（件/天）

$TH(i)$ = 工站 *i* 的平均产出（件/天），其中可能包括某些路由选择（routing）或重工部件的多次造访

注意到带星号的参数，$r^*(i)$、r_b^*、T_0^* 和 W_0^* 正是 $r_e(i)$、r_b、T_0 和 W_0 的理想情况。需要它们的原因是以瓶颈速率和原始加工时间运行的产线，由于 r_b 和 T_0 引起许多低效，而实际上不能表现出完美绩效。完美绩效，也因此有两个水平。首先，产线必须达到给定参数下可能的最佳绩效；这就是第七章的最佳情形。其次，它的参数必须尽可能地好。故而，完美绩效意味着最好中的最好（*best of best*）。

使用上面的参数，我们可以定义七项效率指标来衡量单产品产线的绩效。

产出（Throughput）被定义为所用的（*used*）产线生产部件的速率。理想情况下，它应与需求精确地匹配。产出太少，将损失销售；产出太多，又将建立不必要的制成品库存（FGI）。（291|292）因为我们还将有另一个量度来惩罚（penalize）过量的库存，所以将**产出效率（throughput efficiency）**定义为产出是否足够满足需求，即

$$E_{TH} = \frac{\min\{TH, D\}}{D}$$

如果产出大于或等于需求，产出效率等于 1。任何短缺都将降低这个量度。

工站的**利用率（Utilization）**是它处于运转状态的时间的比例。未加利用的产能意味着额外的成本，所以理想产线中各工站都百分之百地被利用。[2]进一步地，因为完美产线不会受扰动（detractor）的侵袭，在最佳可能（无扰动）速率下利用率将达到 100%。因此，对于一条有*n*个工站的产线，**利用效率（utilization efficiency）**定义为

$$E_u = \frac{1}{n}\sum_{t=1}^{n}\frac{TH(i)}{r^*(i)}$$

库存（Inventory）包括 RMI、FGI 和 WIP。完美产线中没有 RMI（供应商恰如其时地配送），没有 FGI（产品也将恰如其时地送达客户），只有为达到设定的产出（根据里特定律，其值为 $\sum_i TH(i)/r^*(i)$）而需要的最少量 WIP。因此**库存效率（inventory efficiency）**为

$$E_{inv} = \frac{\sum_i TH(i)/r^*(i)}{RMI + WIP + FGI}$$

周期时间（Cycle time）对于成本和收益都很重要。较短的周期时间意味着较少的 WIP，较高的质量，较好的预测和较少的报废（scrap）——所有这些都将降低成本。它也意味着较快的响应，从而提高销售收入。根据里特定律，平均周期时间完全由产出和 WIP 决定。因此，有着完美的产出效率和库存效率的产线理所当然有完美的周期时间效率。然而对于完美产线，WIP 不完全决定于库存效率（由于它包含了 RMI 与 WIP），所以周期时间成为一

[2] 注意到 100%的利用率只对完美（*perfect*）产线是可能的。在实际有变动性的产线，强迫利用率接近一将严重地降低其他量度。务必要记得，系统绩效由所有的效率指标来衡量，而不仅是其中一个。

项独立的量度。我们定义**周期时间效率（cycle time efficiency）**为最佳可能周期时间（无扰动是的原始加工时间）与实际周期时间的比值：

$$E_{CT} = \frac{T_0^*}{CT}$$

提前期（Lead time）是预报（quote）给客户的时间期限；出于竞争的原因，它越短越好。事实上，对于备货生产体系，提前期为零，显然是最短的了。然而对于接单生产体系，零并不是一个合理的目标。因此，给定提前期（LT）至少与理想的原始加工时间一样大，我们定义**提前期效率（lead time efficiency）**为理想的原始加工时间与实际提前期的比值。如果提前期小于原始加工时间，则定义提前期效率为一。其形式如下：

$$E_{LT} = \frac{T_0^*}{\max\{LT, T_0^*\}}$$

注意，在接单生产体系中我们可以预报不尽合理地短的提前期（小于T_0^*）并确保这一项量度为一。但如果产线不能如此快速地送出产品，客户服务的量度将受损失。（292|293）

客户服务水平（Customer service）是被及时满足的需求的比例。对于备货生产，它是补给率（由存货补给需求，而非延期交付的比例）。对于接单生产，客户服务水平是在提前期内完成（即，周期时间小于或等于提前期）的订单的比例。因此，我们以客户服务水平自身来定义**服务效率（service efficiency）：**

$$E_S = \begin{cases} \text{备货生产中由存货补给的需求的比例} \\ \text{接单生产中在提前期内完成的订单的比例} \end{cases}$$

质量（Quality）是关于产品、过程和客户的一项复杂特性。出于运营的目的，质量的本质是第一次就合格地从产线流出的部件的比例。任何的报废或重工都降低这个值。因此，我们衡量**质量效率（quality efficiency）**

$$E_Q = \text{备货生产中由存货补给的需求的比例}$$

这些效率指标以单产品产线的形态描述。然而，可以通过综合物流（flow）与库存（如，以美元计）或单个地衡量周期时间、提前期与服务水平（见问题 1）来将这些量度扩展到多产品产线。

对于完美的单产品产线，上述七项效率指标都等于 1。例如，第七章 Penny Fab One 中无扰动，故$r_0 = r_0^*$、$T_0 = T_0^*$。如果原材料的配送恰如其时（每两小时一件毛坯），客户订单的接受（与输出）两小时一次，CONWIP 水平设在$WIP = WIP^*$，则库存、提前期和服务效率都将等于 1。最后由于没有质量问题，质量效率也等于 1。显然我们不能期盼在现实世界中看到如此完美的产线。所有现实的生产系统都有一些效率指标小于 1。

对于不完美的产线，绩效是这些效率指标（或适用于产线特殊环境的类似指标）的组合。理论上，我们可以构造一个这些指标加权平均形式的单一效率量度。然而，如我们所见，每项的权重高度依赖于产线及其业务的本性。例如，使用贵重的资产型设备的商品生产者强调利用和服务效率，不看重库存效率；而专业的加工车间会强调提前期效率，不惜以利用效率

为代价。

考虑图 9.2 所示的例子。它是一间生产个人电脑的“box plant”内补给装配作业的一条卡片控制填料产线。在这种情形下，制成品库存对于由看板控制的最终装配作业来说，实际上是中间存货。最后一个工站处 5%的重工表示要被修正的。已重工的卡片不会再次重工。

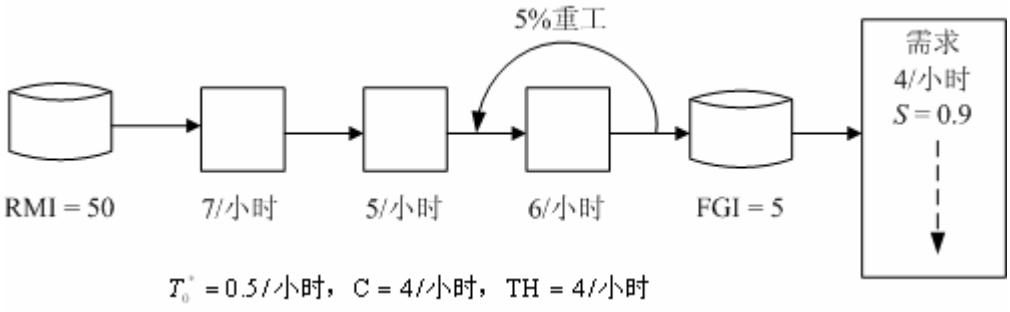

图 9.2 运营效率示例

因为 TH 等于需求，产出效率 E_{TH} 等于 1。周期时间效率 $E_{CT} = T_0^* / CT$ = 0.5 / 4 = 0.125. 利用率效率为各个工站的利用率的均值。要得到这个值，必须先计算出各工站的产出。因为工站 3 处有 5%的重工，

$$TH(3) = TH + 0.05TH = 1.05(4) = 4.2$$

工站 1 和工站 2 处无重工，TH(1) = TH(2) = 4。因此利用效率

$$E_u = \frac{1}{3}\sum_{t=1}^{3}\frac{TH(i)}{r^*(i)} = \frac{\frac{4}{7}+\frac{4}{5}+\frac{4.2}{6}}{3} = 0.6905$$

该问题的资料给出，服务效率 E_S 为 0.9。生产受控于看板系统，提前期为零故而有 E_{LT} = 1.0。质量效率 E_Q 也给出，为 0.95。要得到库存效率，必须先根据里特定律计算出 $WIP = TH \times CT$ = 4×4 = 16 件。理想 WIP 为 $\sum_i TH(i)/r^*(i) = \frac{4}{7}+\frac{4}{5}+\frac{4.2}{6}$ = 2.071。然后有

$$E_{inv} = \frac{\sum_i TH(i)/r^*(i)}{RMI + WIP + FGI} = \frac{2.071}{50+16+5} = 0.0292$$

表 9.3 系统效率的比较

指标	卡片填料系统	修正的卡片填料系统
周期时间	0.1250	0.1250
利用率	0.6905	0.6905
服务水平	0.9000	0.9990
质量	0.9500	0.9500
库存	0.0292	0.0256

假设现在我么提高看板水平，使 FGI 平均有 15 个卡片的数量；并且假设这个变化引起服务水平提高到 0.999。这样，其他的绩效指标不变，E_S 变成 0.999，E_{inv} 下降到 0.0256。

表 9.3 比较了前后两个系统。

哪个系统较好？这取决于企业的业务策略更看重高的服务水平还是低的库存。这种环境中，很可能是修正的系统较好。因为投料产线（stuffing line）的客户是组装线，因而压缩10%的周期时间很困难引起最终客户所不能承受的服务水平。

9.2.2 变动性定律

既然已经用合理而具体的术语定义了绩效，我们就可以刻画变动性对绩效的影响了。变动性可能影响供应商的配送、制造加工时间或客户需求。详细检视它们就会发现，任何来源的变动性增加都会降低上述效率量度中的至少一个。例如，若保持产出恒定时提高加工时间变动性，由第八章的 *VUT* 方程可知 WIP 将增加，故而降低了库存效率。（294|295）若对 WIP 设置限制（通过看板或 CONWIP），则根据我们对有阻塞的排队系统的分析，一般而言，产出将下降（因为瓶颈会饥饿），故而降低了产出效率。

这些观察的结构时下述工厂物理学定律的特例。

定律（变动性）：变动性的增加总是降低生产系统的绩效。

这是一个极为有力的观念，因为它指出提高任何类型的变动性都将伤害某些绩效量度。因而，不管企业给各个绩效量度设定多大的其中，变动性削减都是绩效改进的中心议题。事实上，JIT 方法的大部分成功都是认识到变动性削减的威力并开发出实现方法（如，生产平滑、缩短换模时间、全面质量管理与全面预防性维护）的结果。

通过观测变动性的增加将沿库存、产能和时间三个维度影响系统，我们能加深对变动性定律的认识。显而易见地，库存效率衡量了库存的影响，产出与利用效率衡量了产能的影响，周期时间与提前期效率衡量了时间的影响。服务效率也衡量了时间的影响，因为客户必须等待尚未就绪的部件。最后，质量效率在全部三个维度上影响系统：报废或重工需要额外的产能，重新执行操作需要额外的时间，正在（或等待）维修或重工的部件增加了系统的库存。

另一种看待的方法是，将这三种影响视为控制该系统的缓冲。较差的绩效相应于较多的缓冲。我们可以将其总结为下述的工厂物理学定律。

定律（变动性缓冲）：生产系统中的变动性可以通过

1. 库存
2. 产能
3. 时间

的某种组合来缓冲。

这个定律是变动性定律极为重要的延伸，因为它列举了变动性影响系统的方式。变动性会减低系统绩效已是毫无疑问的了，我们又有了对它如何影响的选择权。不同的业务环境中，有着对付变动性的不同策略。例如，在前面电路板零件组装（board-stuffing）例子中，修正的系统采用较大的库存缓冲来实现较小的时间（服务）缓冲，在那个环境中是个很好的变革。我们在列举一些缓冲变动性的不同方式的例子。

9.2.3 缓冲的例子

下面的例子显示（1）变动性必须被缓冲以及（2）最佳的缓冲策略取决于生产环境与业务策略。我们有意选取了一些非制造业的例子，以强调变动性定律不仅适用于生产系统，它同样适用于服务系统。（295|296）

圆珠笔（Ballpoint pens）。设想有个零售商销售便宜的圆珠笔。需求不可预测（变动的）。顾客如果在此处买不到就会到别处去买（谁会等待一支廉价圆珠笔的延期交付？），零售商不能用时间来缓冲这种变动性。类似地，顾客即刻带走的需求也排除了接单生产的可能性，产能不能用于缓冲。这样就只剩下库存了。而事实上，这正是零售商的做法——保有圆珠笔存货。

紧急服务（Emergency service）。火警或急救服务的需求必定是变动的，因为人们显然不能计划出这类紧急事件。我们不能用库存（消防队伍或医院库存？）来缓冲这种变动性，业不能用时间来缓冲，因为响应时间就是该系统的关键绩效量度。因此，唯一可用的缓冲是能力。而事实上，消防车与急救车的利用率很低。“过量的”能力对于满足需求的高峰是必需的。

肾脏移植（Organ transplants）。肾脏移植的供给和需求都是变动的，因为二者都无法计划。供给速率取决于捐赠者的死亡速率，所以不能（伦理道德上）提升产能。又由于捐赠者死后肾脏只有很短一段存活期，不能使用库存作为缓冲。这样就只剩下时间。而事实上，大多数肾脏移植的等待时间都很长。医疗系统也得遵守工厂物理学的定律。

丰田生产方式（The Toyota Production System）。丰田生产方式是 JIT 的诞生地，并一直是精益制造的典范。依赖于这种方式的成功，丰田从一个相对无名的企业成为世界领先的汽车制造商之一。他们是怎么做到的？

第一，丰田不遗余力地削减变动性，尤其是：

1. 需求变动性。丰田的产品设计和市场营销非常成功，以至于对它的需求一直超过供给（1970s 晚期美国的三大汽车巨头 also did their part by building particularly shoddy cars）。这有几个好处。首先，丰田可用限制选择方案（options）的数目。栗色丰田车总是由栗色的内饰。许多选择方案，如镀铬的行李箱与收音机，由经销商安装。其次，丰田可用提前几个月建立生产计划。这事实上消除了制造设施可见的所有需求变动性。

2. 制造变动性。通过关注缩短换模时间、标准化作业程序、全面质量管理、防错法（error proofing）、全面预防性维护以及其他的物流平滑技术，丰田在消除工厂内部变动性方面做了许多工作。

3. 供应商变动性。1980s 早期丰田与供应商的关系多少有些像封建制度。丰田的需求占它的供应商产品的庞大份额，以至于它有着非凡的影响力。事实上，丰田的主管常常作为董事参与其供应商的董事会。这样就确保了（1）当需要时丰田能找到供应商，（2）供应商采用丰田“建议”的变动性削减技术，（3）供应商承担所有必需的库存。

第二，丰田使用产能缓冲来对付残余的制造变动性。厂内作业每天少于三班，并在每班末预防性维护的时间里补足落后于生产定额的部分。结果就是每天的生产速率高度可预见。

第三，与美国 JIT 作者“零库存”、“罪恶的库存”的倾向不同，丰田确实在系统中保有 WIP 和 FGI。但由于它的强有力的变动性削减努力以及采用产能来缓冲的意愿，所需库存的数量远低于 1980s 汽车制造商的一般水平。

9.2.4 现在支付还是稍后支付

缓冲定律也可以被视为“现在支付还是稍后支付”，原因是如果不在削减变动性上花费，就会（*will*）在以下的一种或几种上花费：

- 损失的产出
- 浪费的产能

• 膨胀的周期时间
• 较高的库存水平
• 长的提前期和/或差的客户服务水平

为了检视缓冲定律在更加具体的制造术语（item）中的体现，我们考虑图 9.3 所示的简单的两工站产线。工站 1 从无限的原材料供给处拉进包含 50 件的加工任务，执行作业，再收入工站 2 前面的缓冲区。工站 2 从缓冲区拉进加工任务，执行作业，再送入下游。再这个例子中，我们假定工站 1 完成一个加工任务要 20 分钟，，并且它是瓶颈。这意味着理论产能为 3,600 件/天（24 小时/天出×60 分钟/小时×1 加工任务/20 分钟×50 件/加工任务）。[3]

作为开始，我们假定工站 2 的平均加工时间也是 20 分钟，故产线是平衡的。因而，理论最小周期时间为 40 分钟，最低 WIP 水平为 100 件（每个工站处有一个加工任务）。然而，由于变动性，系统达不到这个理想的绩效。下面我们讨论此系统在多种状况下的计算机仿真模型的结果，来说明产能、变动性与缓存区容量数值改变的影响。这些结果汇总在表 9.4 中。

表 9.4 现在支付还是稍后支付模拟结果的汇总

情形	缓冲区（工件数）	$t_e(2)$ 分钟	CV	TH（每天）		CT（分钟）	WIP（件）
				E_{TH}	E_u	E_{CT}	E_{inv}
1	10	20	1	3,321		150	347
				0.9225	0.9225	0.2667	0.2659
2	1	20	1	2,712		60	113
				0.7533	0.7533	0.6667	0.6667
3	1	10	1	3,367		36	83
				0.9353	0.7015	0.8333	0.8451
4	1	20	0.25	3,443		51	123
				0.9564	0.9564	0.7843	0.7776

平衡的产线、中度变动性、大的缓冲区（Balanced, Moderate Variability, Large Buffer）。作为开始，我们考虑两台机器平均加工时间均为 20 分钟、中度变动性（即，加工CVs等于 1，故 $c_e(1)=c_e(2)=1$）的平衡产线，工站之间的缓冲区可保持 10 个加工任务（500 件）[4]对此系统 1,000,000 分钟（694 天，每天 24 小时运行）的模拟结果显示，产出为 3,321 件/天，平均周期时间为 150 分钟，平均WIP为 347 件。注意到产出可以改写为 3,321 件/天÷1,440 分钟/天 = 2.3 件/分钟，由此可以核对里特定律（$WIP = TH \times CT$）

347 件≈2.3 件/分钟×150 分钟 = 345 件

因为我们模拟的是有变动性的系统，TH、CT 和 WIP 的估值不可避免有偏差。然而，由于采用了长期仿真，系统能达到稳态，并非常符合里特定律。

注意到这种配置达到了合理的产出（即，只比 3,600 件/天的理论值低 7.7%），却以高的 WIP 和长的周期时间为代价。原因在于，两工站速度的波动导致工站之间缓冲区有规律地充满，从而抬高了 WIP 和周期时间。所以说，这个系统使用 WIP 作为应付变动性的主要缓

[3] 它就是第八章中的问题 10。

[4] 注意，由于产线平衡，前端无限供给，如果工站之间的缓冲区无限大时两台机器的利用率都将是 100%。但这将导致系统不稳定，WIP 增长到无限大。有限容量的缓冲区会不时满溢并阻塞工站 1，堵住工站 1 的释放并阻止 WIP 无限增长。它有利于稳定系统并使其更能表现实际的生产系统，因为实际系统的 WIP 水平绝不允许变成无限大。

冲。

平衡的产线、中度变动性、小的缓冲区（Balanced, Moderate Variability, Little Buffer）。一种降低上述情形的高 WIP 与长周期时间的方法是命令（fiat）。即，简单地压缩缓冲区容量。它实际上是在没有结构性改变的情况下实施低 WIP 看板体系的做法。为了充分地说明这种做法的影响，我们将缓存区容量从 10 个加工任务减少到 1 个。如果工站 1 完成作业是工站 2 前面的队列中还有一个加工任务，它就以非生产性的阻塞（*blocked*）状态等待，直到工站 2 完成。（298|299）

仿真模型证实了小的缓冲区如期望地压缩了周期时间和 WIP，具体值是周期时间降低到约 60 分钟而 WIP 降低到约 113 件。然而，产出降低到约 2,712 件/天（相对于第一种情形下降了 18%）。没有了缓冲区的高 WIP 水平来保护工站 2 免受工站 1 速度波动的影响，工站 2 频繁地因缺少可用的加工任务而饥饿。所以，产出与收益严重衰退了。工站 2 的利用率降低了，故系统现在使用产能作为应付变动性的主要缓冲。可是在大多数环境中，这并不是使周期时间和 WIP 下降的可接受的代价。

不平衡的产线、中度变动性、小的缓冲区（Unbalanced, Moderate Variability, Little Buffer）。上一种情形中工站 1 和工站 2 易于相互阻塞或饥饿，部分原因是它们的产能相等。若缓冲区有一个加工任务且工站 1 又在工站 2 之前完成一个加工任务，在工站 1 阻塞；若缓冲区空虚且工站 2 又在工站 1 之前完成加工任务，则工站 2 饥饿。两种状况都经常发生，故没有一个工站能以接近产能的速率运行。

不平衡的产线是一种解决方法。如果有个机器明显快于另一个，它就几乎总是先完成其加工任务，从而确保另一个以接近其产能的速率运行。为了说明这一点，我们假设工站 2 处的机器换为两倍速（即，每个加工任务的平均加工时间 $t_e(2) = 10$ 分钟），CV 保持原值（即，$c_e(2) = 1$）。缓存区容量保持在一个加工任务。

仿真模型预测产出戏剧性地升高到 3,367 件/天，而周期时间和 WIP 水平依然很低，分别是 36 分钟、83 件。当然，绩效改进的代价是浪费的产能——工站 2 的利用率不到 50%——所以该系统仍然使用产能作为应付变动性的主要缓冲。如果较快的机器不贵，这有是有吸引力的。然而，如果费用不菲，这种选择几乎肯定是不可接受的。

平衡的产线、低度变动性、小的缓冲区（Balanced, Low Variability, Little Buffer）。最后，为了以短的周期时间和低的 WIP 达到高的产出，却不凭借浪费产能，我们考虑削减变动性的选择。在这种情形下，回到平衡的产线，两个工站平均加工时间均为 20 分钟。不同的是，假定加工 CVs 从 1.0 削减到 0.25（即，从中度变动性类型到低度变动性类型）。

在这些条件之下，我们的仿真模型显示，产出很高，3,443 件/天；周期时间很短，51 分钟；WIP 很低，123 件。因此，如果这种变动性削减可以做到并且费用可以承担，它就提供了最好的方法。如我们在第八章所见，有许多措施可以削减加工变动性，像提高机器可靠性、加速设备修复、缩短换模时间、减少作业员缺勤等等。

对比（Comparison）。从表 9.4 的汇总结果可见，上述四种情形是对变动性缓冲定律现在支付还是稍后支付性质的直观说明。在第一种情形，用长的周期时间和高的 WIP 水平“支付”产出；在第二种情形，以损失的产出“支付”短的周期时间和低的 WIP 水平；在第三种情形，以浪费产能支付它们；在第四种情形，以变动性削减支付高的产出、短的周期时间

和低的 WIP。（299|300）变动性缓冲定律不能区别哪种支付形式是最佳的，但它却实在地警告要采取某种形式。

9.2.5 柔性

尽管变动性总需要某种形式的缓冲，它的影响也可以由**柔性（flexibility）**来减轻。柔性的缓冲指可以有不止一种用途的缓冲。由于柔性的缓冲往往比固定的缓冲更有效，我们给出缓冲定律的如下推论。

推论（缓冲的柔性）：柔性可以降低生产系统所需的变动性缓冲的数量。

柔性产能的一个例子是交叉培训的人力（a cross-trained workforce）。通过流动到需要能力的作业，相比于要求固定到具体任务的作业员，柔性作业员能以较少的总体能力完成同样地工作负荷。

柔性库存的一个例子是产品延迟差异化系统中保持的通用（generic）WIP。例如，惠普生产通用打印机供给欧洲市场，先空出针对具体国家的电源连接装置。通用打印机可以接单组装，从而满足来自任何欧洲国家需求。结果是，比起固定的（针对具体国家的）库存，只需少得多的通用（柔性）库存即可确保同样地客户服务水平。

柔性时间的一个例子是根据当前作业积压的情况给客户预报变动的提前期（即，积压越多，预报值越大）。如果分别预报给客户各个变动的提前期，而不是一个一致的、固定的提前期，就可以用较短的平均提前期实现给定的客户服务水平。

柔性加入生产系统的方法很多，如产品设计、设施规划、工艺设备、劳工政策、供应商管理等等。发现创造性的新方法来使资源更富柔性，是以大规模生产的成本制造多批次产品的大规模定制方式的核心挑战。

9.2.6 组织学习

现在支付还是稍后支付的例子显示，提升产能和削减变动性在某种意义上可以互换。两者都可用于压缩给定产出的周期时间或提高给定周期时间的产出。然而，还是有一些模糊的问题需要考虑。首先是实施的难易程度。提升产能通常是个简单的方案——多买些机器就行了——而削减变动性往往就更为困难（有风险），它要求识别过量变动性的源头并执行专门设计的方针来消除。从这个视角看，如果产能扩大与变动性削减对产线的成本与影响相同的话，产能提升是个更吸引人的选择。

但是还有一个重要而模糊的问题要考虑——学习。成功地变动性削减项目将会产生可以移植到其他业务的能力。进行系统分析研究的经验（在第六章中讨论的）、产生的特殊制程改进（如，减少换模时间或重工）、员工对变动性作用结果的认识提升都是变动性削减项目收益的例子。（300|301）变动性削减的心态创造了过程能力持续改进的环境。这可能是一种显著竞争优势的源泉——谁都可以多买机器，但并不是每个人都可以持续提升使用它的能力。出于这个原因，我们认为变动性削减往往是更好的改进选择，在诉诸产能提升之前应该认真地考虑它。

9.3 流动定律

变动性影响物料流经系统的方式以及多大的产能被实际利用。这一节里，我们描述与物料流动、产能、利用率和变动性传递相关的定律。

9.3.1 产品流动

我们开始于一条直接源于（自然）物理学的重要定律，即物料守恒。以制造的术语，陈述如下：

定律（物料守恒）：在一个稳定的系统中，长时期内，减去产出损失，加上产品分离后，出入系统的物料相等。（*In a stable system, over the long run, the rate out of a system will equal the rate in, less any yield loss, plus any parts production within the system.*）

短语“稳定的系统”要求系统的输入不超过（或甚至等于）其产能。下一个短语“长时期内”意味着系统经过相当长时期的观测。短期内它就很容易违反。例如，输出物料多于输入物料——暂时性地。当然了，如果这有，工厂的 WIP 将下降并逐渐变成零，导致输出停止。因而，这条定律也不会被无限期违反。最后一个短语“减去产出损失加上产品分离”是对“输入必须等于输出”这个简单陈述的重要补充说明。产出损失指不是由产出引起的系统内的部件减少（如，报废或损坏）。产品分离指一个部件变成多个部件。例如，一块金属板可能会经剪切作业形成几块较小的板。

这条定律通过产出将产线中单个工站的利用率联系起来。例如，在一条按 MRP（推式）规则、无产出损失的串联产线，任何工站 *i* 的产出 TH(i)加上产线本身的产出 TH，等于产线的投料速率 r_a。原因当然是输入的必须输出（前提条件是投料速率低于产线产能，因而是稳定的）。那么各工站的利用率就是产出与工站产能的比值（如，工站 *i* 处 $u(i) = TH(i)/r_e(i) = r_a/r_e(i)$）。

最后，这条定律支持我们的瓶颈是最忙而非最慢的工站的定义。例如，如果有产出损失，产线下游的慢速工站可能比上游的快速工站利用率低（即，因为上游工站加工的部件有些后来报废了）。因为上游工站限制了整条产线的绩效，我们认为它是瓶颈。

9.3.2 产能

物料守恒定律暗示产线的产能必须至少等于系统的到达速率。否则，WIP 水平持续上升而不能稳定。（301|302）然而，考虑到变动性的存在，这种状况就不那么充分了。考察其原因，让我们回想起第八章给出的排队模型。它指出如果系统不为 WIP 设置上限，当利用率接近一时，WIP 和周期时间都会增长至无穷大。因此，为了达到稳定，系统中所有工站的加工速率都必须严格大于其到达速率。这种行为并非什么数学怪事，它实际上是工厂物理学的一条基本原则。

为了明白这一点，注意到如果生产系统有变动性（所有的现实系统都有），那么不管 WIP 水平的高低，我们总能找到一系列引发系统瓶颈饥饿（用尽 WIP）的可能事件。确保瓶颈工站不会饥饿的唯一办法是队列中总是保有 WIP。然而，不管开始时有多少 WIP，都会有一个加工时间与达到间隔时间的组合将其逐渐耗尽。总是保有 WIP 的唯一办法是起始的 WIP 数量无穷大（*infinite*）。因而，对于 r_a（到达速率）等于 r_e（加工速率）的情形，队列中必须有无穷多的 WIP。但根据里特定律，这意味着周期时间也是无穷大。

这种行为有一个例外。当 c_a^2 和 c_e^2 都等于零时，系统是完全确定性的。在这种情形下，到达间隔时间和加工时间完全没有随机性，到达速率准确地（*exactly*）等于服务速率。然而，

现代物理学（“自然”，而非“工厂”）告诉我们随机性总是存在的，所以这种情形永远不会在实际中出现。

在这一点上，有实际经验的读者可能会怀疑，它们可能会想，“等一等。我在许多工厂呆过，它们中有许多都是尽最大努力设定与产能相等的投料速率，并且我也看到了有限数量 WIP 的例子。”这是一种正当的观点，它提出了**稳态（steady state）**的重要概念。

稳态对应于物料守恒定律中讨论过的“稳定的系统”和“长期的”绩效。处于稳态的系统，其参数必须永远不变，它还必须运行地足够长以至于初始状态无涉。[5]我们的公式都是在稳态的假定下推导出来的，所以我们的分析（它是正确的）与实际所见（它也是正确的）的差异必须置于制造系统稳态的视野中来看待。

超时的邪恶循环（The Overtime Vicious Cycle）。稳态的实情是工厂以一系列的“循环”运行，系统参数随时间重复变化。一种普遍的行为类型是“超时的邪恶循环”，其轨迹如下：

1. 将随机断供、重工、换模、作业员不可用、休息与午餐等扰动考虑近来，计算工厂的产能。

2. 根据这个有效产能制定主生产计划。投料速率等于产能。[6]

3. 或迟或早地，由于加工任务的到达，或加工时间，或二者同时出现随机性，瓶颈制程开始饥饿。（302|303）

4. 进来的比出去的多，故 WIP 上升。

5. 系统以满产能运行，所以产出保持相对恒定。根据里特定律，WIP 的上升引发周期时间近乎等比例地上升。

6. 加工任务延迟。

7. 客户开始抱怨。

8. WIP、周期时间上升得够多了，客户的抱怨声也够大了，管理层觉得采取措施。

9. 超时的“一次性（one-time）”授权、增加一个班次、转包、拒绝新的订单等等，都被许可了。

10. 作为第 9 步的结果，有效产能显著大于投料速率了。例如，如果增加了第三班，利用率从 100%降到约 67%。

11. WIP 水平下降了，周期时间缩短了，客户服务水平提高了。每个人都长出了一口气，惊异于事情就这么了解，并许诺说再也不会让这种事发生了。

12. 回到第 1 步！

超时的邪恶循环的道德意义在于，尽管管理层想要以瓶颈速率投料，然而稳态下，做不到。当超时，或增加班次，或周末加班，或转包等被授权许可，工厂的产能一下子跃升到显著高于投料速率的水平。（类似地，拒绝订单引起投料速率一下子落到产能之下。）因而，从长期来看，平均投料速率总是低于平均产能。我们以下面的工厂物理学定律来概括这项制造生活的实情。

定律（产能）：在稳定状态，所有工厂投入物料的平均速率严格小于其平均产能。

这条定律有着深邃的启示。既然工厂的资源不可能达到 100%的利用率，涉及过量产能、

[5] 回忆起第七章的 Penny Fab 例子。产线必须运行一会儿来消除所有硬币都在第一个工站处的短暂初始状态。当产线开始以相同行为一次次重复时，就达到了稳态。在有变动性的产线，实际行为不会重复，但系统处于给定状态的概率是稳定的。

[6] 请注意如果在计算产能时有一厢情愿的想法时，投料速率将大于产能。

超时或转包的现实管理决策都将是计划的策略的一部分，或都将用于对失控状态做出响应。不幸的是，许多制造经理并不欣赏这条工厂物理学定律，他们无意识地选择以“救火”模式运营自己的工厂。

9.3.3 利用率

缓冲定律和 *VUT* 方程显示，排队时间有两项驱动因素：利用率和变动性。当然，变动性有着最剧烈的影响。原因是 *VUT* 方程（对单机或多机工站）在分母上有 $1-u$ 项。故当利用率 u 接近一时，周期时间趋于无穷大。我们将其陈述为下面的定律。

定律（利用率）：如果工站不做任何其他改变就提高利用率，平均 *WIP* 和周期时间将以高度非线性方式上升。

短语“以高度非线性方式”常常反映了实际问题。探究其原因，假设利用率 $u = 99\%$，周期时间为两天，加工时间和到达间隔时间的 CVs 都等于 1。如果将利用率提高一个百分点到 $u = 0.9797$，周期时间变成 2.96 天，上升了 48%。（303|304）显而易见地，周期时间对利用率很敏感。更进一步，如图 9.4，当 u 接近一时，这种效应就更显著了。这个图揭示了 $V = 1.0$ 与 $V = 0.25$（$V = \dfrac{c_a^2 + c_e^2}{2}$）是周期时间与 u 的关系。我们看到当 u 接近 1.0 时，两条曲线都发生“爆炸”，但表示较高变动性系统（$V = 1.0$）的曲线爆炸得更快。根据里特定律，我们推断当 u 接近一时，WIP 也将会有类似的爆炸。

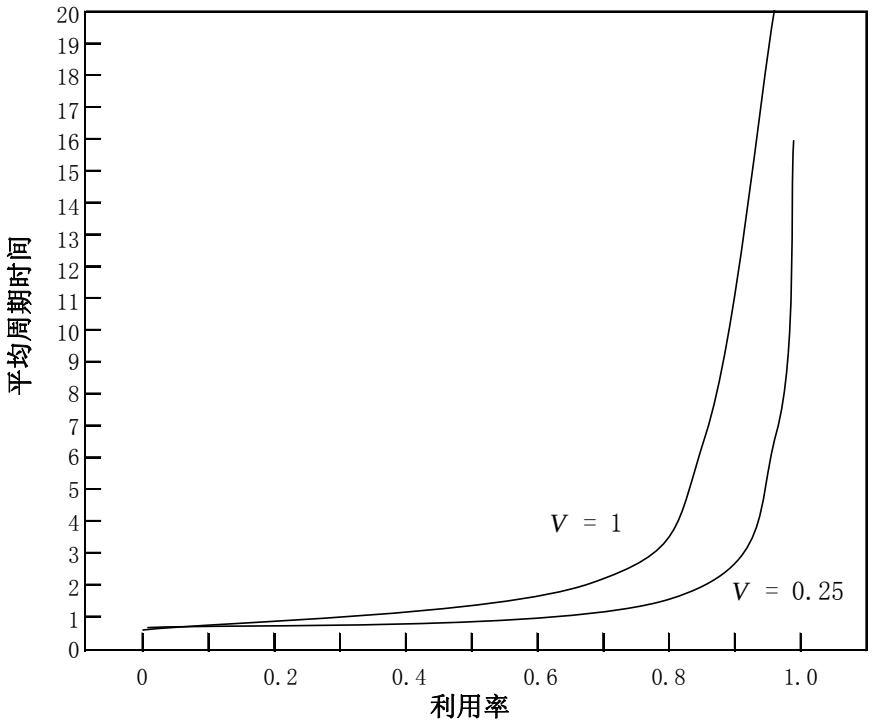

图 9.4 周期时间与利用率的关系

下面补充两条技术性说明。第一，若 $V = 0$，周期时间在利用率小于或等于 100%时保持不变，而在利用率大于 100%时变成无穷大（不可行）。以类似于我们在第七章中研究过的

最佳情形产线的方式，完全无变动性的工站能以 100%的利用率运行而不建立队列。可是所有的实际工站都有某种变动性，上述情形不会在实际中发生。

第二，现实中所有的工站都没有空间来建立无限长的队列。空间、时间或政策会将 WIP 限制在某个有限的水平。如我们在第八章的阻塞模型所见，若不做任何其他改变就对 WIP 设定限制，产出（进而还有利用率）将下降。因而，图 9.4 的定量关系依然成立，但对队列长度的限制会使曲线中高利用率、长周期时间的部分无法达到。

系统绩效对利用率的极端敏感性，是的选取一个同时达到高的工站利用率和短的周期时间的投料速率非常困难。任何偏差，尤其是偏大（很可能因对系统产能的乐观与最大化产出的意愿而发生），可能会导致平均周期时间的极大增长。我们将在第十章推式与拉式生产系统的背景下探讨接近这个问题的结构性变革。

9.3.4 变动性与物流

变动性定律讲述了变动性降低所有生产系统的绩效。但降低了多少，则依赖于变动性生成于产线的何处。在没有 WIP 控制的产线，任何工站处加工变动性都会（1）延长该工站处的周期时间并且（2）向下游工站传递更多的变动性，因而也延长了它们的周期时间。这个观察结果激发了变动性定律与第八章中的变动性传递性质的如下推论。（304|305）

推论（变动性位置）：在投料独立于产出的产线中，前端的变动性比同等但在后端的变动性导致更长的周期时间。

这条推论的含义是，削减变动性的努力应首先直接作用于线首，因为它们在那里有可能产生最大的效果（见问题 10）。

请注意这条推论只适用于投料独立于产出的情形。在 CONWIP 产线，投料与产出有直接关联，第一个工站处物流受最后一个工站处物流影响的强度与工站 i 受工站 $i+1$ 影响的强度相同。因而，产线前端与后端就没什么区别了，也就没什么动机优先在线首削减变动性。变动性位置推论适用于推式系统，而非拉式系统。

9.4 成批定律

成批是变动性的一个特别富于戏剧性的原因。如我们在第七章中最差情形绩效所见，甚至在加工时间恒定时，以大的批量运输产品都会带来最大值的变动性。那个例子中的原因是第一个部件的有效到达间隔时间很大，而其他的为零（它们同时到达）。结果就是各工站“见到”高度变动的到达；对于给定的瓶颈速率的原始加工时间，平均周期时间糟糕到极点。由于成批能对变动性，进而是绩效，有如此大的影响，制造系统中批量的设定就是一项非常重要的控制了。然而，在计算“最优”批量（留在第十章作为排配的一部分）之前，我们需要理解成批对系统的影响。

9.4.1 批次的类型

有一个常常混淆关于成批的讨论的问题，就是实际上有两种批次。考虑一条只生产一种类型产品的专用产线。每个单位产出后，运送到喷漆工序。这个批量是多大呢？

一方面，可以说是一个，因为每个单位完成后都运送到喷漆工序；另一方面，也可以说是无限，因为没有发生换产（changeover）（即，两次换产之间的部件数目无限多）。一不等于无限，哪一个对呢？

答案是两个都对。然而是有两种不同的批次的：**加工批次（process batch）**与**转运批次（transfer batch）**。

加工批次。有两种类型的加工批次。**串联批次（serial batch）**的容量是工站换产另一个族之前，加工的本族加工任务的数量。成为串联批次是因为部件顺序地生产（一次一件）。**并联批次（parallel batch）**的容量是一个真正的批处理工站，如熔炉或热处理，同时生产的部件的数量。串联与并联批次本质上很不相同，但我们将看到二者对运营有着相似的影响。（305|306）

串联加工批次的容量与换产或者换模的时间长度有关。换模时间越长，就要在两次换模时间之间生产越多的部件来实现设定的产能。并联加工批次的容量取决于对工站的要求。为了最小化（？——译者注）利用率，这类机器应以满批运行。可是，若它不是瓶颈，最小化利用率可能不时很重要，故以低于满载运行可能会很好地压缩周期时间。

转运批次。它是转运到下一个工站之前累积的部件的数量。转运批次越小，周期时间就越短，因为等待成批的时间缩短了。然而，较小的转运批次也引发较多的物料搬运，故其中有个权衡。例如，如果产线中相邻工站件的物料运输是每批 3,000 单位，叉车每班可能只需用一次；而如果每批 100 单位，叉车每班就需开动 30 趟了。

严格说来，如果将工站之间的物料搬运作业看作一道加工，转运批次也就是简单的并联加工批次了。叉车可以像运一件那样快地运 10 件，就像熔炉可以像烘一件那样快地烘 10 件。不过，直觉上看物料搬运还是与加工不同，所以我们将分别讨论转运与加工批次。

加工批次与转运批次的区别常常被忽略。事实上，从 1915 年 Ford Harris 首先推导出 EOQ 到现在，大多数生产计划这简单地假定这两种批次应当相等。当未必要如此。换模时间长而制程邻近时，就应当使用大的加工批次和小的转运批次。这种做法称为**批次分离（lot splitting）**，并可以显著地缩短周期时间（我们将在 9.5.3 小节对此做更详细的讨论）。

9.4.2 加工批次

回忆起在第四章，JIT 的拥护者热切呼唤批量等于 1。原因在于，如果一次加工一件，没有花费在等待成批的时间了，也就有较少的时间花费在队列中等待大的批次。然而在大多数现实世界的系统，将批量设定为一并非这么简单。原因是批量能影响产能。这也正是批量为一将引发工站过载（由于过多的换模时间或过多的并联批次加工时间）的情形。因而，挑战就是平衡这些产能方面的考虑与成批引发的延迟（见 Karmarkar（1987）的更完整的讨论）。我们可以在下面的工厂物理学定律总结串联与并联加工批次的关键动力学原理。

定律（加工批次）：对于进行批量作业或有着较长换产时间的工站:

1. 形成稳定系统的最小加工批量可能大于 1。
2. 随着加工批次的增大，周期时间响应成比例地增长。
3. 周期时间可以在某些加工批量处达到最小值，而该加工批量可能大于 1。

我们可以用下面的例子来说明这条定律描述的产能与加工批次之间的关系。（306|307）

例子：串联加工批次

考虑一个加工几种部件族的机加工站。部件成批到达，同一批次的部件都属同一族，但批次属不同的族。批次的到达速率已被设定，使得部件以 0.4 件/小时的速率到达。不管族类，每个部件都要一小时的加工时间。然而，机器在批次之间需要一次五小时的换模（因为已假

定要切换族类)。因而，批量的选择将影响所需的换模次数（进而是利用率）和各个批次所需的加工时间。进一步地，周期时间还受整批加工完成后部件成批离开还是使用批次分离而一个一个地离开。

注意到如果我们要采用 1 的批量，就只能每六小时加工一件（五小时的换模加上一小时的加工），可是它跟不上到达速率。我们考虑的最小批量为四件，这样就确保了没九小时四件的产能（五小时的换模加上四小时的加工），或是 0.44 件/小时的速率。

图 9.5 显示了使用与不使用批次分离是，一系列批量对应的工站周期时间。注意到最小的可行批量使周期时间在无批次分离是达到约 70 小时，有批次分离时达到约 68 小时。无批次分离时，最小周期时间约为 31 小时，对应的批量是八件；有批次分离时，最小周期时间约为 22 小时，对应的批量是九件。在这些最小水平之上，周期时间以近似直线的方式增长，并且批次分离情形以一个逐渐增大的边际优于（达到较小的周期时间）非批次分离的情形。

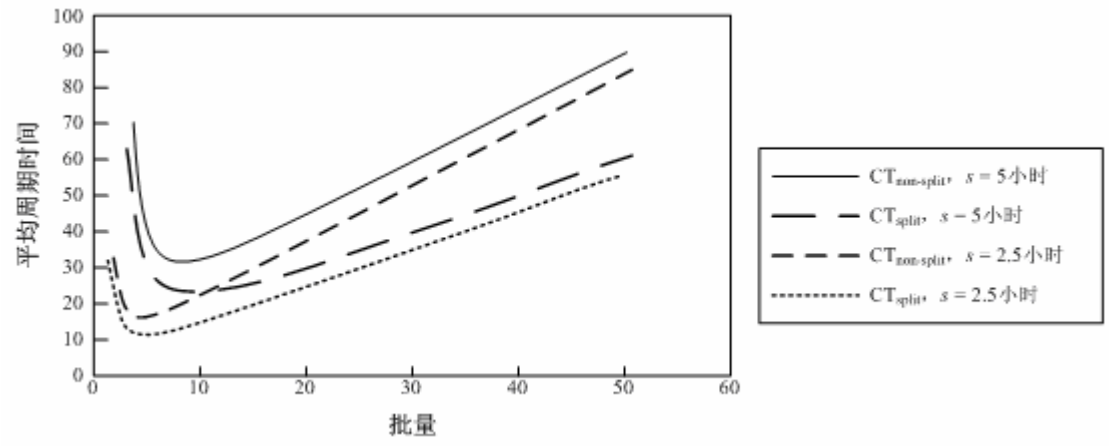

图 9.5 换模时间为 2.5 小时与 5 小时下周期时间与并联批量之间的关系

加工批次定律指出，有必要，甚至是意愿使用大的加工批量来维持利用率，从而有周期时间和 WIP，与控制之中。但应当慎重地接受这个结论。大的串联加工批量的需要源于长的换模时间。因此，第一优先级应是在经济可行性允许的范围内尽力缩短换模时间。例如，图 9.5 显示了机加工站行为的例子，同时给出了平均换模时间为 2.5 小时和 5 小时的情形。注意到有了较短的换模时间，最小的周期时间约比原来小 50%（无批次分离是约为 16 小时，有批次分离时约为 11 小时），并在较小的批量处达到（无批次分离时为四件，有批次分离时为五件）。故上述定律的全部含义时，成批和换模时间压缩应一起使用，来达到高的产出、有效率的周期时间水平。（307|308）

例子：并联加工批次

考虑一家医疗诊断设备生产企业的可靠性测试工序。该工序要在一间温控室测试成批的设备，不管里面有多少件都要进行 24 小时。该测试室一次最多容纳 100 件。假设到达速率为 1 件/小时（24 件/天）。显然地，如果一次测试一件，产能就只有 1/24 件/小时，远低于到达速率。事实上，如果以 24 件的批量送入，产能就达到 1 件/小时，测试室的利用率就达到 100%。由于利用率必须小于 100%以实现系统稳定，最小可行批量为 25 件。

图 9.6 绘制了周期时间作为批量函数的曲线。它显示出当批量为 32 件时周期时间最小，且最小值为 43 小时。其中 24 小时是加工时间，剩下的 19 小时是排队和等待成批时间。我们将在稍后开发计算这些量值的公式。

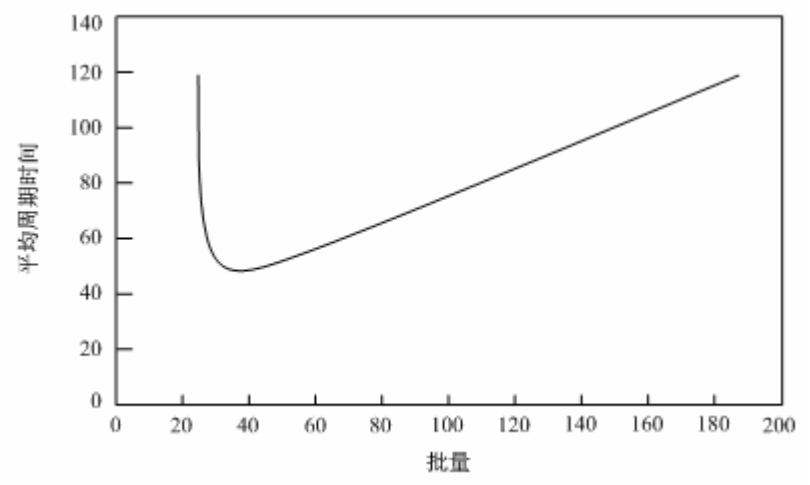

图 9.6 批处理中的周期时间–批量关系

串联批次（serial batching）。通过检视上述例子背后的模型，我们可以给出加工批次定律含蓄给出的批量-周期时间交互作用的深入解释。首先在下面的技术性注释中分析图 9.5 的串联批次情形。

技术性注释——串联批次的交互作用

为了给成批的部件到达单一机器，并在批次之间有换模地被加工的串联批次建模，我们使用下面的记号：

c = 串联批量

r_a = 到达速率（件/小时）

t = 加工单个部件的时间（小时）

s = 一次换模的时间（小时）

c_a^2 = 一个批次加工时间，包括加工时间与换模时间，的有效 SCV

进一步地，我们使用这些简化的假定：（1）有效加工时间的SCV $c_e^2 = 0.5$，而无论批量是多大，[7]（2）（批次的）到达SCV总是一。

r_a 是部件的到达速率，故批次的到达速率是 r_a / k 。批次的有效加工时间由批次中 k 件的加工时间加上换模时间给出

$$t_e = kt + s \qquad (9.1)$$

故机器利用率为（308|309）

[7] 可以固定单个过死人的 CV 并将批次的 CV 作为批量的函数来计算。然而，模型假定恒定的批量一到达 CV，显示出同样的重要行为——若是小批，周期时间急剧增长；若是大批，周期时间线性增长——并易于分析。

$$u = \frac{r_a}{k}(kt + s) = r_a\left(t + \frac{s}{k}\right) \tag{9.2}$$

注意到为了系统的稳定，必须有 $u < 1$，它要求

$$k > \frac{sr_a}{1 - tr_a}$$

VUT 方程给出在队列中的平均时间 CT_q

$$CT_q = \left(\frac{1 + c_a^2}{2}\right)\left(\frac{u}{1 - u}\right)t_e \tag{9.3}$$

其中的 t_e 和 u 由（9.1）式和（9.2）式给出。

工站处的总平均周期时间包含排队时间、换模时间、批内等待时间（wait-in-batch-time，WIBT）与加工时间。WIBT 取决于批量是否处于运输到下游的目的而分离。如果没有，（即，整批都完成加工后，才有部件运输到下游），则所有的部件等待批次中的其他 k – 1 个部件，故

$$WIBT_{nonsplit} = (k - 1)t$$

总的周期时间为

$$\begin{aligned} CT_{nonsplit} &= CT_q + s + WIBT_{nonsplit} + t \\ &= CT_q + s + (k - 1)t + t \\ &= CT_q + s + kt \end{aligned} \tag{9.4}$$

如果批次分离（即，单个部件完成后立刻送往下游，故使用了一件的转运批量），则 WIBT 取决于部件在批次中的位置。第一个部件没有花费时间等待，因为它加工完成后立即离开了。第二个部件在第一个之后等待，因而在批中等等待了 *t* 时间。第三个在批中等待了 2*t* 时间，依此类推。故 *k* 个加工任务的平均批内等待时间为

$$WIBT_{split} = \frac{k - 1}{2}t$$

故有

$$\begin{aligned} CT_{split} &= CT_q + s + WIBT_{plit} + t \\ &= CT_q + s + \frac{k - 1}{2}t + t \\ &= CT_q + s + \frac{k + 1}{2}t \end{aligned} \tag{9.5}$$

（9.4）式与（9.5）式是图 9.5 的基础。我们可以用 $k = 10$ 时图 9.5 代表的例子的数据（r_a = 0.4，$c_a^2 = 1$，$t = 1$，$c_e^2 = 0.5$，$s = 5$），对其使用做具体的说明。故

$$t_e = s + kt = 5 + 10\times 1 = 15 \text{ 小时}$$

机器利用率

$$u = \frac{r_a t_e}{k} = \frac{0.4\times 15}{10} = 0.6$$

批次在队列中的期望时间

$$CT_q = \left(\frac{1+0.5}{2}\right)\left(\frac{0.6}{1-0.6}\right)15 = 16.875 \text{ 小时}$$

所以，如果不采用批次分离，平均周期时间为（309|310）

$$CT_{nonsplit} = CT_q + s + kt = 16.875 + 5 + 10\ (1) = 31.875 \text{ 小时}$$

如果将加工批次分离为一件的转运批次，平均周期时间为

$$CT_{split} = CT_q + s + \frac{k+1}{2}t = 16.875 + 5 + (10+1)\ /\ 2\ (1) = 27.375 \text{ 小时}$$

如所期望的，它比前者小。

上述分析的主要结论是，若换模时间可以做到足够短，使用一件的串联加工批量是个缩短周期时间的有效途径。然而，若短的换模时间不可能（至少是在近期），周期时间会对加工批量的选择敏感，并且“最佳”批量可能远大于 1。

并联批次（parallel process batching）。依赖于控制政策，串联批处理可以在批次尚未形成时就开始作业，还可以在整批尚未完成加工时就释放加工任务。（我们将在下一节李检视这种导致工站间周期时间“重叠”的方式。）但在并联处理，如热处理炉、烘焙箱或测试室，整批一次加工因而必须同时开始和结束。这使得对并联批次的分析与对串联批次的分析略有不同。

并联批次处理工站的总的周期时间包括等待成批时间（累积成一个整批的时间）、排队时间（整批在队列中等待的时间）以及加工时间。我们在下面的技术性注释中开发计算它们的公式。

技术性注释——并联批次的交互作用

假定部件一次一件地到达并联批处理工序。它们要等待形成批次，以批次的形式在队列中等待，然后以批次的形式被加工。我们使用下面的记号，与串联批次情形的相似。

k = 并联批量

r_a = 到达速率（件/小时）

c_a = 到达间隔时间的 CV

t = 批次的加工时间（小时）

c_e = 批次到达间隔时间的有效 CV

B = 最大批量（可送入加工的部件数目上限）

要计算平均等待成批时间（wait-to-batch-time，WTBT），请注意平均到达间隔时间为$1/r_a$。

批次中的第一个部件等待其他 k – 1 个的到达，所以等待了$(k-1)/r_a$小时。最后一个部件完全不需要等待，即刻成批。因此，部件等待成批的平均时间是这两个极值的均值，或是

$$WTBT = \frac{k-1}{2r_a}$$

一旦 k 个到达发生，就有了可以送入队列或工序的一个整批。因此，批次的到达间隔时间等于 k 个部件到达间隔时间之和。如在第八章所见，k 个 SCV 为c^2的独立同分布随机变量加和后得到的随机变量的 SCV 为c^2/k。所以，批次的到达 SCV 为（310|311）

$$c_0^2(batch) = \frac{c_a^2}{k}$$

批次 k 的加工产能为k/t，故最大产能为B/t。为了保持利用率低于 100%，有效产能必须大于或等于需求，所以我们令

$$u = \frac{r_a}{k/t} < 1$$

或者是

$$k > r_a t$$

若 B 小于或刚好等于$r_a t$，则产能不足以满足需求。

批次一旦形成，即进入批处理。如果利用率很高且存在变动性，很有可能出现队列。排队时间可由 *VUT* 方程计算

$$CT_q = \left(\frac{c_a^2/k + c_e^2}{2}\right)\left(\frac{u}{1-u}\right)t_e$$

结果，总的周期时间为

$$\begin{aligned} CT &= WTBT + CT_q + t \\ &= \frac{k-1}{2r_a} + \left(\frac{c_a^2/k + c_e^2}{2}\right)\left(\frac{u}{1-u}\right)t + t \\ &= \frac{k-1}{2ku}t + \left(\frac{c_a^2/k + c_e^2}{2}\right)\left(\frac{u}{1-u}\right)t + t \end{aligned} \qquad (9.6)$$

其中最后一个等式用到了$u = r_a/(k/t)$故$r_a = uk/t$的关系式。

注意到（9.6）式指出，当利用率接近零，正如它接近一时，周期时间会变得很大。原

因是利用率很低时，比起加工时间来到达显得很慢，因而成批的时间很长。

如在图 9.6 所见，并联批处理的周期时间显著地受批量的影响。依赖于作业的能力，最好以不满的批次运行。为了寻求最优批量，我们可以在电子表格中实现上面技术性注释中的计算并使用试错法。另外，还可以使用一种解析的方法，会在第十五章中讲到。

9.4.3 转运批次

在对一个组装厂的参观中，我们的向导自豪地展示他的一项最新举措——制造单元。铸件从铸造车间送入，在不到一个小时的时间里，完成钻孔（drill）、机加（machine）、磨（grind）、抛光（polish）作业。它们再从这个制造单元送入部装线作业。向导指出，通过将多种作业置于紧邻的位置并聚焦于单元内的流线型流（streamlining flow），这部分路线的周期时间已经从几天缩短为一个小时。我们感到震惊——直到我们发现由叉车以约 10,000 件的数量将铸件送到制造单元，再将完成的部件一起送入组装线！结果是第一件需要仅仅一小时通过该单元，却要在它被送入组装线之前等待其他的 9,999 件部件。该单元的产能是 100 件/小时，每车要等待 100 小时才填满。因此，尽管制造单元的设计初衷是削减 WIP 和周期时间，实际绩效确是我们所见的最接近第七章最差情形的。（311|312）

该工厂选择以 10,000 件的批量转运部件的原因是，错误地（却是普遍地）假定转运批量应等于加工批量。然而，在大多数生产环境中，并无必然的理由要这样。如我们在上面所见，批量的分离可以极大地缩短周期时间。当然了，小的批量也意味着较多的物料搬运。例如，若上述单元以 1,000 件（而非 10,000 件）的批量转运部件，则每车需要每 10 小时（而非每 100 小时）运一次。尽管这个组装厂很大且制程之间的转运冗长，这项额外的物料搬运显然是可管理的，它本可以在产线的这个部分将 WIP 和周期时间压缩到原来的 1/10。

这个例子的隐含意义总结为如下的工厂物理学定律。

定律（转运批量）：如果不需要等待运输设备，路线上某一段的周期时间大体上和其转运批量成比例。

这条定律指出在某些制造系统中压缩周期时间最简单的方法之一，就是降低转运批量。而事实上，它是如此简单以至于可能会被管理层忽略。降低转运批量可能是简单和廉价的，所以在诉诸更复杂的周期时间压缩策略之前一定要考虑它。当然了，较小的转运批量会要求较多的物料搬运，所以有“如果不需要等待运输设备”的补充说明。如果较频繁地在工站之间运输部件，它们等待物料搬运设备的时间就变长了，这项额外的排队时间可能会抵消等待成批时间的削减。因而，转运批量定律描述可能通过减低转运批量实现的周期时间的缩短，是在已经假定物料搬运设备能力足以无延迟地执行转运的情况下。

要正确评价周期时间与转运批量的关系，请注意其动力学原理和并联批次处理相同，物料搬运设备在并联批次处理中相当于并联批次的一道工序。如果批量太小，利用率将上升，并导致对物料搬运设备的排队等待时间过量。我们通过下面技术性注释中的数学模型来更精确地说明其机制。

技术性注释——转运批量

考虑图 9.7 所示的简单的两工站串联产线中批次的效应。第一个工站接受单个部件并一次一件地加工。部件收集为 k 件的转运批次，再送入第二个工站。第二个工站成批加工，并

以单件送入下游。为了简化问题，我们假定工站之间的转运时间为零。

图 9.7 一个成批和解批的例子

令 r_a 表示产线的到达速率，$t(1)$、$c_e(1)$ 分别表示第一个工站加工时间的均值和 CV，于是可以计算利用率 $u(1)=r_a t(1)$。应用 VUT 方程得出期望的排队等待时间

$$CT_q(1)=\left(\frac{c_a^2(1)+c_e^2(1)}{2}\right)\left(\frac{u(1)}{1-u(1)}\right)t \qquad (9.7)$$

在第一个工站处的总时间包括这项排队时间、本身的加工时间以及等待成批时间。(312|313) 观察到第一个部件不需等待其他 $k-1$ 个部件，而最后一个部件完全不需要等待，这样就可以计算平均成批时间。由于到达批处理的速率与部件到达工站的速率 r_a 相等（物流的守恒），平均成批时间是 $(k-1)(1/r_a)$ 和 0 的均值，故而是 $(k-1)/(2r_a)$。又有 $u(1)=r_a t(1)$，于是

$$\overline{WTBT}=\frac{k-1}{2r_a}=\frac{k-1}{2u(1)}t(1)$$

如所期望的，当批量 k 等于 1 时，它等于零。下面，我们给出部件在工站 1 处的总时间

$$CT(1)=CT_q(1)+t(1)+\frac{k-1}{2u(1)}t(1) \qquad (9.8)$$

为了计算第二个工站处的平均周期时间，我们可以将其视为一个整批的队列、一个单部件的队列（即，半批，partial batch）、与一个服务台。可以使用（9.7）式与 $u(1)$、$c_a^2(2)$、$c_e^2(2)$ 的批次调整值来计算整批的排队等待时间 $CT_q(2)$。我们这样做是因为批次的离开间隔时间等于 k 个部件的离开间隔时间之和。因而，如在第八章所见，k 个 SCV 为 c^2 的独立同分布随机变量加和后得到的随机变量的 SCV 为 c^2/k。所以，第二个工站的成批到达 SCV 为 $c_d^2(1)=c_a^2(1)/k$。类似地，由于必须加工 k 个独立的部件来完成批次的加工，第二个工站处成批加工时间的 SCV 为 $c_e^2(2)/k$，其中 $c_e^2(2)$ 为第二个工站处单个部件的加工 SCV。加工一个批次的有效平均时间为 $kt(2)$。如，如所期望地，利用率为

$$u(2)=\frac{r_a}{k}kt(2)=r_a t(2)$$

从而，由 *VUT* 方程，第二个工站处的平均周期时间为

$$CT_q(2)=\left(\frac{c_a^2(2)/k+c_e^2(2)/k}{2}\right)\left(\frac{u(2)}{1-u(2)}\right)kt(2)$$

$$=\left(\frac{c_a^2(2)+c_e^2(2)}{2}\right)\left(\frac{u(2)}{1-u(2)}\right)t(2)$$

有意思的是，整批的排队等待时间与我们本已计算的单个部件的排队等待时间相同（消去了 k 项，剩下的是一般的 *VUT* 方程）。

除了满批的队列，还必须考虑半批（partial batch）的队列。可以通过考虑部件在半批中的时间来计算。等待空闲机器的第一个部件完全不需等待，而批次中的最后一个要等在其他 $k-1$ 个之后才能完成加工。因而，批次中部件的平均等待时间为 $(k-1)t(2)/2$ 。

部件在第二个工站处的总的周期时间为批次在队列中的等待时间、在半批中的等待时间以及部件的实际加工时间之和：

$$CT(2)=CT_q(2)+\frac{k-1}{2}t(2)+t(2) \qquad (9.9)$$

现在我们可以将批量为 k 的两工站系统的总的周期时间表述为（313|314）

$$CT_{batch}=CT(1)+CT(2)$$

$$=CT_q(1)+t(1)+\frac{k-1}{2u(1)}t(1)+CT_q(2)+\frac{k-1}{2}t(2)+t(2)$$

$$=CT_{\sin gle}+\frac{k-1}{2u(1)}t(1)+\frac{k-1}{2}t(2) \qquad (9.10)$$

其中 $CT_{\sin gle}$ 表示无批次（即，$k=1$）时系统的周期时间。

（9.10）式定量地阐述了转运批量定律——周期时间相对于批量成比例地增长。可是要注意，批量 k 增长时发生的周期时间增长与加工时间或到达间隔时间变动性无关（即，（9.10）式中包含 k 的项中不包含 CV）。确实存在变动性——有些部件等待很长时间来成批而其他的根本不用等待——但那是不良控制或不良设计（与第七章中的最差情形相似）引起的变动性，而不是加工或流动不确定性引起的。

最后，我们注意到当第一个工站的利用率很低时转运批次的影响最大，因为它使（9.10）式中的 $(k-1)t(1)/[2u(1)]$ 变大。其原因是当到达速率相对于加工速率来说很低时，充满一个转运批次的时间很长。因而，部件用大量的时间以半批的形式等待。这很像并联加工批次情形（见（9.6）式）。（9.6）式与（9.10）式的唯一区别是，在前者中我们未将转运过程处理为有能力限制的。如果这样做了，两种情形就是一样的了。

单元制造（Cellular Manufacturing）。转运批次定律最重要的启示是大的转运批次直接抬高了周期时间。因而，降低转运批次可以是一个缩短周期时间的有效策略。保持转运批次较小的一种方法**是单元制造**，我们在第四章 JIT 的背景下已做介绍。

理论上，一个单元将生产一族部件所需的所有工站在物理上紧邻布置。由于物料搬运被

最小化了，工站之间以小批量，理想的批量是一件，转运部件成为可行的了。如果该单元真的只生产一个族的零件，那将不再有换模。加工批量可以成为一件或无穷大，或二者之间的任意数量（一般受控于需求）。

如果该单元处理多族部件，那就有明显的换模，并且从前面的讨论可知并联加工批次对单元的产能和周期时间都很重要。事实上，如我们将在第十五章所见，应当为不同的族设定不同的加工批量，甚至这些设定还要依时而变。不管加工批次如何完成，它都是与转运批次独立的决策。甚至在由于换模而需要大的加工批次时，我们也能用批次分离来以小的转运批次运输物料并充分利用单元的物理紧凑性（physical compactness）。

9.5 周期时间

已经考虑过利用率、变动性和批次的问题了，现在我们转向较复杂的绩效量度，周期时间。首先考虑单个工站处的周期时间，稍后再描述这些工站如何联合形成产线的周期时间。（314|315）

9.5.1 单一工站处的周期时间

作为开始，我们对单一工站处的周期时间进行拆解。

定义（工站周期时间）：工站处的平均周期时间由下列几部分构成:

周期时间 = 转运时间 + 排队时间 + 换模（生产准备）时间 + 加工时间
+ 等待成批时间 + 批内等待时间 + 等待匹配时间　　（9.11）

转运时间（move time）是加工任务从上一个工站移动过来所花费的时间。**排队时间（queue time）**是加工任务等待在工站处加工或转运到下一个工站所需的时间。**换模时间（setup time）**是加工任务等待工站换模所花费的时间。注意，在加工任务的移动途中换模已部分完成的情况下，它实际上可能小于工站换模时间。**等待成批时间（wait-to-batch-time）**是加工任务等待形成（并联）加工或转运批次使用的时间，**批内等待时间（wait-in-batch-time）**是部件在（加工）批次内等待轮到它在机器上加工所用的时间。最好，**等待匹配时间（wait-to-match-time）**发生在组装工站处组件等待其配合件来实现装配作业的时候。

注意到这些条目中，只有加工时间真正地贡献于产品的制造。转运时间可被视为一种必要的邪恶。因为不管工站离得多近，总是有些移动的时间。然而，其他的条目则完全是无效率。事实上，这些时间常被称为不增值时间、浪费或者 *muda*。它们也常被混为一谈，称为延迟时间或排队时间。然而如我们将要看到的，这些时间是极为不同的原因引发的后果，并因而应当有不同的矫正措施。由于它们常常占到周期时间的绝大部分，将它们区别对待有益于找到具体的改善政策。

我们已经讨论过成批时间，所以现在来看等待匹配时间，之后再转向整条产线的周期时间。

9.5.2 组装作业

大多数的制造系统都包含某种形式的组装。电子器件被插上电路板。车身、发动机和其他部分被组装成汽车。化学品通过化合反应产出其他的化学品。任何使用两种或多种输入来产生输出的制程，都可称为组装作业（an assembly operation）。

组装复杂了生产系统的物流，因为其中包含**匹配（matching）**。在匹配作业中，直到所

有必需的组件都就绪，加工才能开始。如果组装作业由几条制造组件的产线供给，任何一种组件的短缺都会扰乱（disrupt）组装，并因而扰乱所有其他的产线。因为它们对系统绩效有如此大的影响，通常的做法是组件产线的计划与控制服从于组装作业。实现方法是详细指定**最终组装计划（final assembly schedule，FAS）**，再向后计算得到组件产线的计划。我们将在第十二章从质量的立场、在第十四章从车间作业控制的立场、在第十五章从排配的立场讨论组装作业。（315|316）

现在，我们在下面的工厂物理学定律中总结组装作业背后的基本动力学原理。

定律（组装作业）：下面任何一项的上升都将降低组装工站的绩效:

1. 要装配组件的数量，
2. 组件到达的变动性，
3. 组件到达的不协调。

注意，这些中的每一项都可被视为变动性的上升。因而，组装作业定律是更为一般的变动性定律的一个具体例子。这条定律的论据和含义（reasoning and implication）都是相当直观的。为了更具体地说明其意义，考虑将器件插上电路板的作业。所有的组件按 MRP 计划采购。如果其中任何一个耗尽存货（out of stock），组装就不可能发生，计划就被扰乱了。

为了正确评估组件数量对周期时间的影响，假设 BOM 发生了一项变更，要求在最终产品上多加一个组件。其他的保持原样，额外的组件可能由于不时耗尽存货而只抬高了周期时间。

为了理解组件到达变动性的影响，假设企业更换了某个组件的供应商，却发现新的供应商比旧的有更多的变动。与到达变动性引起常规的非组装工站之前出现队列那样的发生，新增的等待变动性将引起工站等待迟到的配送，从而会抬高组装工站处的周期时间。

最后，为了鉴别组件到达缺乏协调的影响，假定企业目前从同一个供应商那里采购两种组件，并总是同时配送。如果企业转化到从两家供应商处采购这两种组件，则它们不再会同时配送了。即使两家供应商有着与先前一家同样的变动性水平，配送不协调的事实还是会引起更多的延迟。当然了，这样的分析忽略了所有其他的复杂因素，如有两种组件要交付可能使得一家供应商较不可靠，或某些供应商可能更适于交付某些组件。但如果其他的都相同，多种组件同步（synchronize）到达将降低延迟。我们将在第十四章讨论将制造产线与组装作业同步飞方法。

9.5.3 产线周期时间

在第七章的 Penny Fab 例子中，加工任务以一件的批量执行作业并被立即送走，故周期时间简单地为加工时间与排队时间之和。可是考虑到成批和转运，就不能总是将各个工站周期时间之和当作产线的周期时间了。由于一个批次可能同时在不止一个工站处被加工（即，若采用了批次分离），我们必须考虑工站件的重叠（overlapping）时间。如，我们定义产线周期时间如下。（316|317）

定义（产线周期时间）：产线的平均周期时间等于单个工站的周期时间之和减去两个或多个工站重叠的时间。

为了说明重叠周期时间的影响，我们来考虑表 9.5 所示的两条产线。产线 1 和产线 2 都是无加工变动性的三工站产线，接收每 35 小时一次、$k = 6$ 件的确定性批次到达。各批次都要有一次换模，之后加工任务一次一件地执行，再被送到下个工站。唯一不同之处是两条产

线的加工时间和换模时间不同（产线 2 与产线 1 相反）。因而，产线 1 工站的利用率递增，工站 1 为 49%，工站 2 为 75%，工站 3 为 100%。在产线 2，正好反过来。出于建模的目的，我们用 $t(i)$ 和 $s(i)$ 分别表示工站 i 的单件加工时间和换模时间。

表 9.5 说明周期时间重叠的例子

	工站 1	**工站 2**	**工站 3**
	产线 1		
换模时间（小时）	5	8	11
单件加工时间（小时）	2	3	4
	产线 2		
换模时间（小时）	11	8	5
单件加工时间（小时）	4	3	2

考虑产线 1。由于顺序地在有换模工站处加工并在作业完成后允许离开，可以应用（9.5）式来计算各工站处的周期时间。在工站 1，有

$$CT(1) = CT_q + s(1) + \frac{k+1}{2}t(1) = 0.0 + 5 + \frac{6+1}{2}(2) = 12$$

其中排队时间为零，因为系统无变动性。

对于工站 2 和工站 3，算法相同，有

$$CT(2) = CT_q + s(2) + \frac{k+1}{2}t(2) = 0.0 + 8 + \frac{6+1}{2}(3) = 18.5$$

$$CT(3) = CT_q + s(3) + \frac{k+1}{2}t(3) = 0.0 + 11 + \frac{6+1}{2}(4) = 25$$

这样就得出总的周期时间

$$CT = CT(1) + CT(2) + CT(3) = 12 + 18.5 + 25 = 55.5$$

可是这不对。批次的第一件加工任务已在工站 2 或工站 3 处作业，而最后一件却还在上一个工站。因而，（9.5）式中的等待成批时间项高估了成批引起工站 2 与工站 3 处的延迟。

对于这个确定性的例子，我们可以由跟随批次中的加工任务一次一件地通过工站来计算周期时间。如图 9.8 所示，第一个到达工站 2 的加工任务周期时间为 $s(2)+t(2)$。第二个工站在 $s(2)+2t(2)$ 时完成，却比第一个迟来 $t(1)$ 小时，故其周期时间为 $s(2)+3t(2)-2t(1)$。一直这样，直到批次中第 k 个（最后一个）加工任务在 $(k-1)t(1)$ 时开始并在 $s(2)+kt(2)$ 时结束，其周期时间为 $s(2)+kt(2)-(k-1)t(1)$。故工站 2 的平均周期时间为（317|318）

$$\begin{aligned} CT(2) &= \frac{1}{k}[ks(2) + (1+2+\cdots+k)t(1) - (1+2+\cdots+k-1)t(1)] \\ &= s(2) + \frac{k+1}{2}t(2) - \frac{k-1}{2}t(1) \\ &= 8 + 3.5(3) - 2.5(2) = 13.5 \end{aligned}$$

$[(k-1)/2]t(1) = 5$ 小时这一项代表了批次重叠的时间。

工站 3 的情形与工站 2 类似，其周期时间为

$$CT(3) = s(3) + \frac{k+1}{2}t(3) - \frac{k-1}{2}t(2)$$

$$= 11 + 3.5(4) - 2.5(3) = 17.5$$

因而，正确的产线 1 的总周期时间为 $CT(1)$、$CT(2)$、$CT(3)$ 修正形式的加和，得

$$CT(line) = s(1) + s(2) + s(3) + t(1) + t(2) + \frac{k+1}{2}t(3) = 43 \text{ 小时}$$

这可以在图 9.8 看到。批次中第一个加工任务的周期时间为 33 小时，第六个为 53 小时，故均值为 (33 + 53) / 2 = 43.4 小时。注意到它比各工站周期时间之和的 55.5 小时要小得多。

如果我们应用（9.5）式到各工站，并把结果加和来计算产线 2 的周期时间，则将得到与产线 1 同样的数值，55.5 小时。（318|319）原因是无变动性时，（9.5）式不受产线顺序的影响。然而，如果参透了产线的机理（if we work through the mechanics of the line directly），将会发现真正的平均周期时间为 38 小时（见图 9.9，第一个加工任务的周期时间为 33 小时，第六个的为 43 小时，故均值为（33 + 43）/ 2 = 38 小时）。再次地，这个值远小于原初的估计。它也比第一种情形小许多（较慢的制程在前时，重叠较多）。要点在于，不仅是重叠的周期时间，还有其他如工站顺序等机制对决定产线周期时间都很重要。

尽管有批次产线的行为很复杂，我们可以跟随一个加工任务通过产线的方式得到关于产线周期时间的见识。如在上述例子中，我们假定了

1. 加工任务成批到达。[8]
2. 各个批次中第一个加工任务经历各工站的完整换模（即，不允许在批次中第一个加工人物到达之前换模，尽管我们确实允许一个工站处换模时间为零的情形）。
3. 加工任务在工站之间一次一件地移动。

在这些条件下，我们在下面的技术性注释中发展了产线周期时间的上限和下限。

技术性注释——周期时间

我们称非排队时间（即，批内等待时间、换模时间和加工时间）为全部在制时间（*total in-process time*）。通过考虑无变动性（因而无队列）产线并检视批次中第一个与最后一个加工任务通过产线的时间 T_1、T_k，来为全部在制时间设定范围。[9]（319|320）

对一条 s_i、t_i 分别表示工站 i 的换模和加工时间的 k 工站产线。第一个加工任务需要各工站的换模和一次加工时间

$$T_i = \sum_{i=1}^{K} s_i + t_i$$

[8] 因为第一个加工任务投入产线后就认为整批进入了产线，出于计算周期时间的目的，假定整批同时到达也是合理的。

[9] 作者要对 Network Dynamics, Inc.的 Dr. Greg Diehl 在开发这些公式的帮助表示感谢。

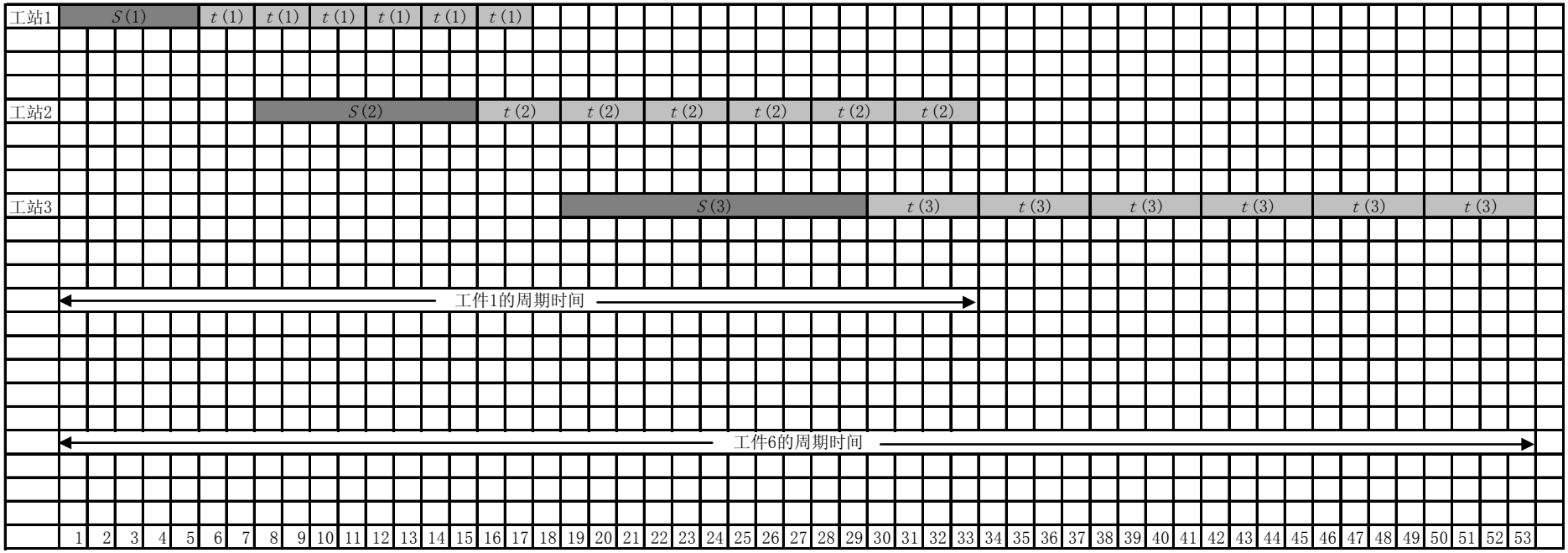

图 9.8 批次分离：从快到慢

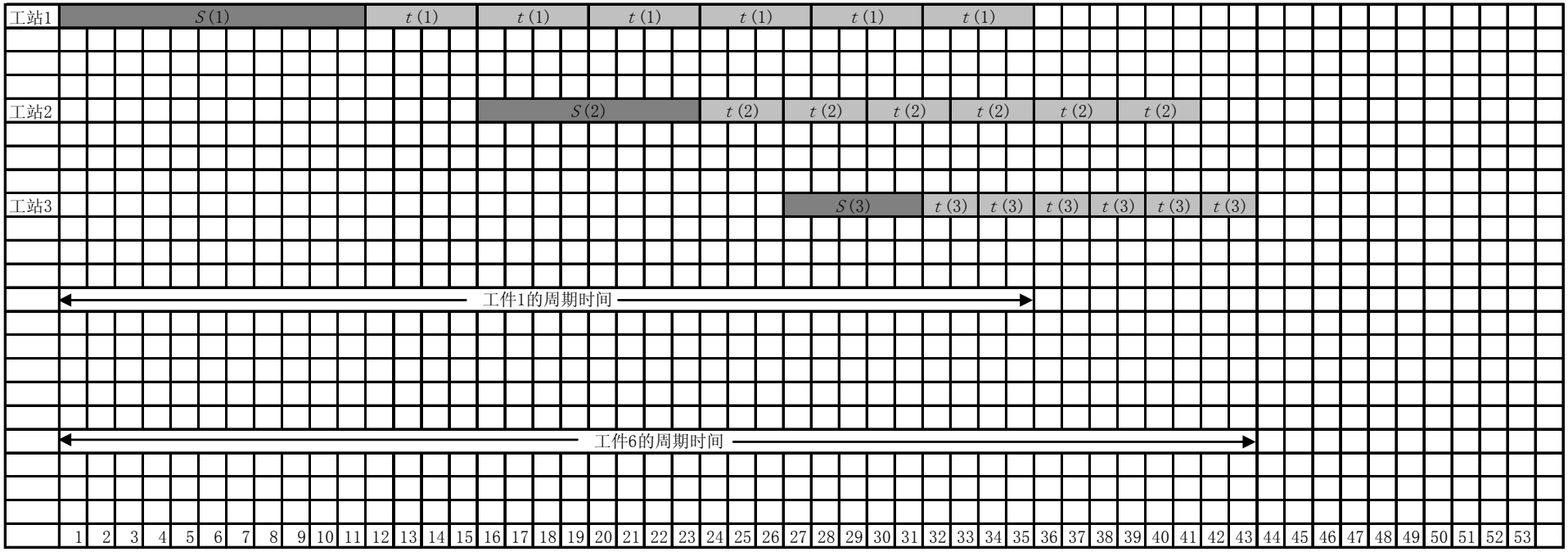

图 9.9 批次分离：从慢到快

对一条 s_i、t_i 分别表示工站 i 的换模和加工时间的 k 工站产线。第一个加工任务需要各工站的换模和一次加工时间

$$T_i = \sum_{i=1}^{K} s_i + t_i$$

最后一个加工任务需要这些时间，再加上在批中等待其他加工任务后等待的时间。最长的时间可能发生于最后一个加工任务遭遇所有其他加工任务要经历最长的加工时间（见图 9.8）。故

$$T_1 \le T_i + (k-1)t_b$$

其中 $t_b = \max_i\{t(i)\}$。平均在制时间的上限是 T_1 和 T_k 的均值，

$$\text{平均在制时间} \le \sum_{i=1}^{K}[s(i)+t(i)] + \frac{k-1}{2}t_b \qquad (9.12)$$

因为所有的加工任务同时到达第一个工站，最后一个加工任务将总是在最后一个工站处等待其他 $k-1$ 个完成后才能执行作业。可见的最小延迟发生于最后一个工站有着最短的加工时间并且没有空闲时间（见图 9.9）。所以平均在制时间的下限可由 $t_f = \min_i\{t(i)\}$ 代替 t_b 计算得到

$$T_k \ge T_1 + (k-1)t_f$$

因而有

$$\text{总的平均在制时间} \geqslant \sum_{i=1}^{K}[s(i)+t(i)] + \frac{k-1}{2}t_f \qquad (9.13)$$

为了给周期时间设定范围，必须在全部在制时间之间考虑排队时间。要做到这一点，回忆起我们对成批转运的讨论。在那里，总的排队时间独立于批量（记起 k 项如何被“消去”）。如果我们可以假定它对串联批次情形近似成立，则对排队时间的良好估计可通过使用 VUT 方程计算整批（*full batches*）在各工站处的平均等待时间得出。在第一个工站，由于成批到达，这个估值就如 *VUT* 方程本身那样精确。在其他一次一件地到达的工站，由于不确切知道 c_a^2，引入了较多的误差。当然了，这个问题也存在与无批次的系统。一系列例子的经验显示，其精度不低于为单个工站开发的公式（在第八章）。

令 $CT_q^b(i)$ 表示满批在工站 i 处的平均等待时间（按通常方法使用 *VUT* 方程计算得出），我们给出有关串联批次产线的总的周期时间的上下限

$$\sum_{i=1}^{n}[CT_q^b(i)+s(i)+t(i)] + \frac{k-1}{2}t_f \le CT \le \sum_{i=1}^{n}[CT_q^b(i)+s(i)+t(i)] + \frac{k-1}{2}t_b \qquad (9.14)$$

其中 $t_f = \min_i\{t(i)\}$，$t_b = \max_i\{t(i)\}$。（320|321）

例子：限定周期时间

再看表 9.5 中的两条产线。若无加工与等待变动性，则排队时间之和为零，换模与加工时间之和为 33.因而周期时间的限制为

$$33 + \frac{6-1}{2}(2) \le CT \le 33 + \frac{6-1}{2}(4)$$

$$38 \le CT \le 43$$

对于产线 1，上限是紧约束；对于产线 2，下限是紧约束。然而，如果我们将顺序调换使得最慢的工站在线首、最快的工站在线尾，结果是 CT = 40.5，在上下限之间。类似地，如果最慢的工站在线中、最快的在线尾，CT = 39.5，也在上下限之间。在这些例子中，批内无空闲时间（即，机器不在同批的加工任务之间发生空闲）。然而，它可能发生并确实发生与最慢的工站在第一位、最快的在第二位的系统（见问题 15）。

（9.14）式的周期时间限制在加工时间相似（即，故有$t_f \approx t_b$）的系统中将互相接近。但在最快的机器比最慢的（如，因为它也有一段很长的换模时间）快得多的产线中，上下限离得很远。更紧的约束需要更复杂的计算（见 Benjaafar 和 Sheikhzadeh 1997）

9.5.4 周期时间、提前期和服务水平

在无限产能和完全无变动性的制造系统中，周期时间和客户提前期的关系很简单——它们是相同的。这个系统的幸运的经理，他可以简单地按用来生产产品的周期时间来向客户预报提前期，并保证有 100%的服务水平。不幸的是，所有的实际系统都包含变动性，故这么完美的服务水平不可能达到，并且人们常常困惑于提前期、周期时间及其与服务水平的关系的区别。尽管在第三章和第七章略微接触过这些议题，我们现在更准确定义它们，并给出一条联系变动性与提前期、周期时间及服务水平的工厂物理学定律。

定义。 在这整本书中我们交替地使用术语周期时间和平均周期时间来表示加工任务通过产线的平均时间。然而，为了探讨提前期，需要在术语使用上更加准确。因而，这一节中我们定义**周期时间（cycle time）**为一个给出单一加工任务通过一条路线的时间的随机变量。特别地，我们定义 T 为表示周期时间的随机变量，其均值为 CT、标准差为σ_{CT}。

与周期时间不同，**提前期（lead time）**是一个用于说明加工任务预期或允许最长的周期时间的管理常量（*management constant*）。它有两种类型：客户提前期和制造提前期。**客户提前期（customer lead time）**是从开始到结束（即，整条路线）地完成客户订单所允许的时间，而**制造提前期（manufacturing lead time）**是一条特定路线上允许的时间。

在**备货生产（make-to-stock）**环境，客户提前期为零。当客户到达，产品要么能供给要么不能。如果不能，服务水平（在此种情形下长成为**补给率（fill rate）**）受损失。在**接单生产（make-to-order）**环境，客户提前期是客户允许企业生产、送达产品的期限。在这种情形下，有变动性时，提前期一般必须大于平均周期时间以获得可接受的服务水平（定义为准时送达的百分比）。（321|322）

一种压缩客户提前期的办法是制造和存储低层级组件（lower-level components）。由于客户只看到余下的作业，提前期将会显著地变短。我们在第十章推式与拉式生产的背景下讨论这种**接单组装（assembly-to-order）**类型的系统。

关系。 对于复杂的 BOM，计算合适的客户提前期将会很困难。一种解决办法是使用制造提前期，它指定了加工任务在一条路线上预期的最大量允许的周期时间。我们标记一条周期时间为 T 的具体路线的制造提前期为 l。制造提前期常用于计划投料（如，MRP 系统）和追踪服务水平。

服务水平（service） s 现在可以定义为接单生产模式下路线的周期时间小于或等于指定提前期的概率。故

$$s = \Pr\{T \le l\} \tag{9.15}$$

如果 T 有分布函数 F，则（9.15）式可用于设定 l

$$s = F(l) \tag{9.16}$$

如果周期时间服从正态分布，则对于服务水平 s

$$l = CT + z_s\sigma_{CT} \tag{9.17}$$

其中 z_s 是标准正态分布表中对应于 $\Phi(z_s) = s$ 的值。例如，一条给定产线上周期时间的均值为八天、标准差为三天，对应于 95%的 z_s 值为 1.645，故所需的提前期为

$$l = 8 + 1.645(3) = 12.94 \approx 13 \text{ 天}$$

图 9.10 显示了周期时间的分布函数 F 和概率密度函数 f。超出均值的额外的五天称为**安全提前期（safety lead time）**。

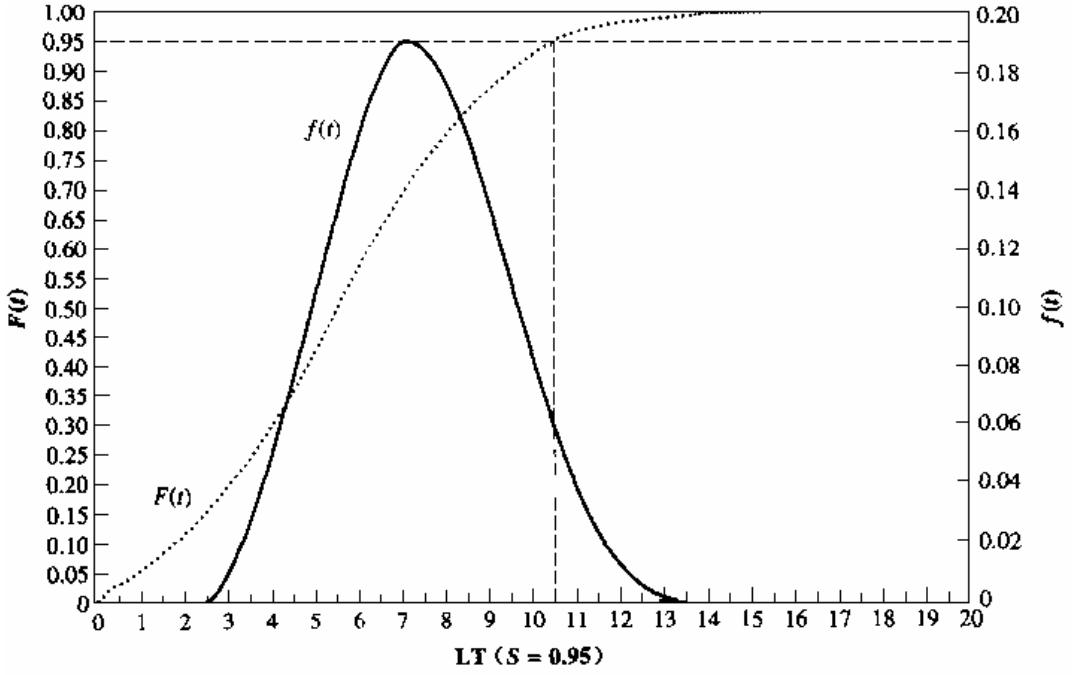

图 9.10 周期时间与所需提前期的分布函数

通过指定足够高的服务水平（以保证加工任务一般能按时完成），我们简单地加上 BOM 各个层级最长的制造提前期（当多条路线汇集到组装线时）来计算客户提前期。例如，图 9.11 显示了一个系统，我们有第零级提前期为六天。因而，总的客户提前期为 10 天。

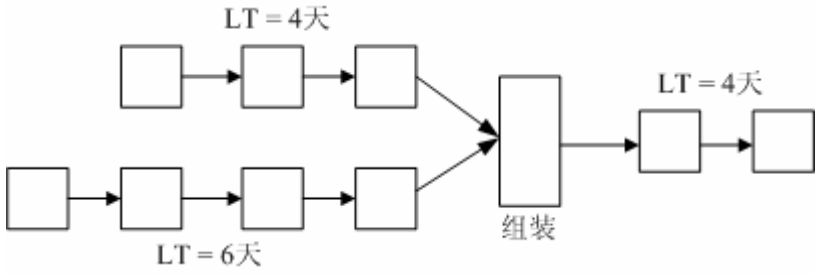

图 9.11 一个组装系统

不幸的是，使用 10 天的客户提前期的总体服务水平将比 95%低一些。这是因为我们未曾考虑组装作业前**等待匹配时间（wait-to-match-time）**的概率。如我们在组装作业定律所见，当变动性引起制造产线以不同步的方式像组装作业供给时，等待匹配时间就会发生。由于这一点，在任务有组装作业的地方都必须增添一些安全提前期。

我们可以在下面的工厂物理学定律中总结联系周期时间中变动性与所需提前期的基本原则。（322|323）

定律（提前期）：在给定服务水平的产线中，制造提前期是周期时间均值和标准差的增函数。

直观上看，这条定律意味着制造提前期等于周期时间加上一个依赖于周期时间标准差的“修正因子（fudge factor）”。周期时间标准差越大，为实现给定服务水平的修正因子也越大。在接单生产环境，我们希望周期时间较短从而保持客户提前期较短，所以就需要保持周期时间的均值和标准差较低。

如我们在第八章所见，抬高周期时间均值的因素一般来说与抬高加工时间标准差的因素相同，包括变动性、随机断供、换模、重工等等。然而，从周期时间角度看，重工尤其具有破坏性。一旦加工任务有可能被要求再次经过一部分产线，周期时间变动性剧烈地增长。我们将在第十二章讨论质量对物料的影响时回顾这个一起其他与周期时间变动性相关的议题。

（323|324）

9.6 诊断与改进

已探讨的工厂物理学定律揭示了制造系统行为的基本方面，并且突出了关键的权衡。可是它们自身却不能产生设计和管理的政策。其原因在于“最优的”运营结构依赖于环境约束和战略目标。在客户服务水平上竞争的企业需要聚焦于快速响应的配送，而在价格上竞争的企业则需要关注设备利用率和成本。幸运的是，无论系统属于何种类型，工厂物理学定律都有助于识别杠杠区域和改进的机会。

下面的例子说明了本章发展的原则在改进现有系统上的应用。主要考虑三个关键的绩效量度：产出、周期时间和客户服务水平。

9.6.1 提高产出

产线的产出由下式给出

$$\text{产出} = \text{瓶颈利用率} \times \text{瓶颈速率}$$

因此，提高产出的两种方式是提高瓶颈利用率或它的速率。谈及提高利用率可能显得不敬（blasphemous），因为我们已经知道它会延长周期时间。但是不同的目标有着不同的对策。在不限制 WIP 的系统，高的利用率引起排队从而延长周期时间。但是，如我们在现在支付还是稍后支付例子所见，在限制 WIP（有限缓冲或诸如看板等强加的逻辑限制）的系统，阻塞和饥饿将制约瓶颈利用率并进而降低产出。

提高产能的一份基本的政策清单如下。

1. 通过提高瓶颈的有效速率来**提高瓶颈速率（increase bottleneck rate）**。方法有添置设备、添置或培训人力、保持工站在休息与午餐时间内仍然运行、使用柔性人力、质量改进、产品设计变更以压缩在瓶颈处的时间等等。

2. 通过削减瓶颈处的阻塞与饥饿来**提高瓶颈利用率（increase bottleneck utilization）**。方法有两种：

- 以 *WIP 缓冲瓶颈*（*buffer bottleneck with WIP*）。可由扩大系统缓冲区的容量（或等效的看板卡片的数量）来做到。最有效的位置是瓶颈紧前的（允许队列增长有益于防止饥饿）和紧后的（建立队列有益于防止阻塞）缓冲区。扩大原理瓶颈的缓冲区也将有意，只是作用比近处的小。
- 以产能缓冲瓶颈（*buffer bottleneck with buffer*）。可由提高非瓶颈工站的有效速率来做到。较快的上游工站将减少瓶颈饥饿的次数，而较快的下游工站将减少瓶颈阻塞的次数。为高利用率的非瓶颈工站增添产能一般有着最大的效果，因为它们是最有可能引起阻塞/饥饿的工站。一般的产能提升政策，如在提升瓶颈工站产能时列举的那些，都是其具体措施。

（324|325）

例子：提高产出

HAL Computer 的一家印制电路板工厂里有一条两工站产线，第一个工站（覆膜 resist apply）将感光树脂涂到电路板，第二个工站（曝光 expose）将板子在紫外线下曝光来蚀刻图案。由于曝光作业必须在一间清洁室进行，两工站件的缓冲空间只能容纳 10 件加工任务。产能计算显示，瓶颈为曝光工站，其加工时间均值为 22 分钟、SCV 为一。覆膜加工时间为 19 分钟、SCV 为 0.25。另外的（不包含在上述加工时间之内），曝光平均失效间隔时间（MTTF）为 $3\frac{1}{3}$ 小时、平均恢复间隔时间（MTTR）为 10 分钟；而覆膜的 MTTF 为 8 小时，加工任务以中度变动性到达，故我们假定到达 SCV $c_a^2 = 1$。想要达到的产出为 2.4 件/小时。

根据以往的经验，HAL 知道产线能力不足以达到目标产出。为了补救，担当责任的工程师趋向于添置第二台曝光机器。然而，除了价格昂贵，第二台机器还需要扩大清洁室，这将显著地增加成本并导致建设期内实在的产出损失。故而，挑战在于用工厂物理学寻找一个较好的解决方案。

我们的两个处理工具是计算排队时间的 *VUT* 方程

$$CT_q = (\frac{c_a^2 + c_e^2}{2})(\frac{u}{1-u})t \qquad (9.18)$$

和连接方程

$$c_d^2 = u^2 c_e^2 + (1-u^2)c_a^2 \qquad (9.19)$$

将这些与第八章中的有效 SCV 公式一道使用，可以分析产线无法达到产出目标的原因。

由（9.18）、（9.19）式（还有计算平均加工时间$t_e(1)$、$t_e(2)$与加工 SCV$c_e^2(1)$、$c_e^2(2)$的公式，我们将在稍后回顾），到达速率设为 2.4 件/小时，我们估计出覆膜工站的排队等待时间为 645 分钟，曝光工站的排队等待时间为 887 分钟。工站 1 和工站 2 的平均 WIP 水平分别为 25.8 件、35.5 件。

这揭示了系统为何做不到 2.4 件/小时，尽管瓶颈（曝光）利用率仅为 92.4%。也就是，清洁室只能容纳 20 件，而模型预测的平均队长是 35.5 件。由于系统不会允许曝光之前的 WIP达到这个水平，覆膜将不时被阻塞（即，因下游缓冲区空间不足，上游完成的部件送不出去导致其闲置），覆膜损失的产能也逐渐引起曝光饥饿（即，因缺乏可加工的部件而闲置）。结果是没有一台机器能维持 2.4 件/小时的产能所需的利用率。[10]

因此，我们认为问题的根源在于曝光处长长的队列。根据里特定律，削减平均队长等效于缩短周期时间。故我们更加详细地考察曝光处的排队时间：（325|326）

$$CT_q(2) = (\frac{c_a^2(2) + c_e^2(2)}{2})(\frac{u(2)}{1-u(2)})t_e(2)$$

= (3.16)(12.15)(23.1 分钟)

= 887 分钟

第三项$t_e(2)$为曝光处的有效加工时间，它是原始加工时间与可用率的简单相除

$$t_e(2) = \frac{t(2)}{A(2)} = \frac{t(2)}{m_f(2)/(m_f(2)+m_r(2))}$$

$$= 22(\frac{31/3+1/6}{31/3})$$

= 23.1 分钟

由于它只是稍大于 22 分钟的原始加工时间，提高可用率的改进效果不大。

$CT_q(2)$表达式中的第二项为利用率项$u(2)/(1-u(2))$。乍看上去 12.15 的数值很大，可它对应于 92.4%这个很大却不过分的利用率水平。提升该工站的产能必定会缩短排队时间（与队长），可我们已经解释过它是一个昂贵的选择。

故而我们来看第一项变动性膨胀因子$(c_a^2(2)+c_e^2(2))/2$。回忆起中度变动性到达（即，$c_a^2(2)=1$）和中度变动性加工时间（即，$c_e^2(2)=1$）的结果是这一项的值为一。因此 3.16 的数值在任何系统中都是过大的。为了探究其原因，我们将其拆解到构成要素，得到

$$c_e^2(2) = 1.04$$

$$c_a^2(2) = 5.27$$

显而易见，到达进程是变动性的支配性源头。这就指出问题存在于上游的覆膜制程。所以现在要来调查$c_a^2(2)$值很大的原因。回忆起$c_a^2(2) = c_d^2(1)$，有（9.19）式可得

$$c_d^2(1) = u^2(1)c_e^2(1) + [1-u^2(1)]c_a^2(1)$$

$$= (0.887^2)(6.437) + [1-0.887^2](1.0)$$

= 5.05 + 0.22

= 5.27

使$c_d^2(1)$很大的部分是覆膜的有效 SCV$c_e^2(1)$。这个系数也是由两部分构成的：自然

[10] 注意到我们已经在 8.7.2 小节中使用阻塞模型分析过这种状况。我们推荐读者尝试问题 13，来看看这个更精巧的工具如何用于获得同样的定性结果，虽然它可以得到更精确的定量结果。

SCV $c_0^2(1)$ 和机器失效引起的膨胀因子。使用第八章的公式，可将 $c_e^2(1)$ 拆解为（326|327）

$$A(1) = \frac{m_f(1)}{m_f(1) + m_r(1)} = \frac{48}{48+8} = 0.8571$$

$$t_e(1) = \frac{t(1)}{A(1)} = \frac{19}{0.8571} = 22.17 \text{ 分钟}$$

$$c_e^2(1) = c_0^2(1) + \frac{2m_r(1)A(1)[1-A(1)]}{t(1)}$$

$$= 0.25 + \frac{2(480)(0.8571)(0.1429)}{19} = 6.44$$

$c_e^2(1)$ 的最大值（lion's share）是随机断供的结果。这意味着提升曝光处产能的替代方案是改善覆膜处的停机状况。注意到曝光是瓶颈但问题出在覆膜，这很重要.由于变动性在产线中传递，一个工站处的拥塞可能实际上是上游工站处过量变动性的结果。

有多种可行的选择可以缓解覆膜处的断供问题。例如，HAL 可以通过保持“随时待命（field-ready）”的备件来缩短平均修复时间。如果这个措施能将 MTTR 减半，其引起的覆膜处有效产能提升与离开 SCV 降低将使覆膜处排队时间降到 146 分钟（比原值小四分之一），是曝光处的排队时间降到 385 分钟（比原值小一半）。

还有一种方式是，HAL 可以争取更频繁的预防性维护。假设可以由每 30 分钟关停机器做一个五分钟的调整，从而避免长时间（八小时）的失效。产能将与原值相同（即，可用率未发生变化），但由于断供更有规律，覆膜处的排队时间下降到 114 分钟，曝光处为 211 分钟。估计里特定律，它对应于曝光处平均有 8.4 件加工任务，正在空间限制之内。

上述任何一种改善实施后，以所要的 2.4 件/小时的速率运行就成为可行了（实际上还要高些）。任何有助于削减覆膜处输出间隔时间变动性的政策都有类似效果。因为改善覆膜处的维修时间分布可能比在曝光处添置新设备便宜而且破坏性小。应当认真考虑这个替代措施。

9.6.2 缩短周期时间

综合考虑工站与产线周期时间的定义，我们可以将生产系统中的周期时间拆解如下：

1. 转运时间。
2. 排队时间。
3. 换模（生产准备）时间。
4. 加工时间。
5. 加工批次时间（等待成批与批内等待时间）。
6. 转运批次时间。
7. 等待匹配时间。
8. 减去工站重叠时间。

大多数的生产系统中，实际加工和转运时间都只占总的周期时间的很小比例（5&~10%）（Bradt 1993）。事实上，这两项占主要的产线已经非常有效率，改善的机会不多。对于无效率的产线，主要的杠杠在于其他项中。下面是削减这些项的一般政策的简单清单。（327|328）

排队时间有由利用率和变动性引起。因此，两类的改进政策如下：

1. 通过提高瓶颈处的有效速率来降低利用率。类似方法可以是提高瓶颈速率（通过添置设备、缩短换模时间、缩短修复时间、工艺改进、作业员在休息和午餐式轮值，等待）或降低进入瓶颈的物流（通过计划变更使物流进入非瓶颈工站、提高产出或减少重工）。

2. 削减任何工站，尤其是高利用率工站处的加工时间或到达变动性。加工时间变动性可以通过缩短修复时间、缩短换模时间、改进质量以减少重工或产出损失、由更好的培训削减作业员变动性等等方法。修建到达变动性可以通过降低上游工站加工变动性、使用更好的

计划与车间作业控制工具平滑物流、取消成批释放（即，一次释放超过一件加工任务）以及安装推式系统（见第十章）来实现。

加工批次时间由加工批量引起。减低（串联或并联）加工批次的两种基本方法是：

1. 批次优化以更好地平滑批次时间与高利用率引发的排队时间。在本章早些时候，我们已经给出一些检视。更详细的优化计算在第十五章中。
2. 换模时间压缩以在不提高利用率的情况下允许更小的批量。有明确的分析和压缩换模时间的技术（Shingo 1985）。

等待匹配时间由组件到达组装工站时缺乏协同引起。改进协同的主要措施有：

1. 削减制造变动性以削减到达组装作业的波动。可通过与缩短排队时间相同的变动性削减技术来实现。
2. 通过车间作业控制和/或计划体系来协同释放到产线来完成组装作业的进程。我们将在第十四章讨论车间作业控制机制，在第十五章讨论计划的程序。

工站重叠时间。与其他的时间不同，我们要增加工站重叠时间，因为它从总的周期时间中减去。可通过在可行处使用批次分离来增加。流线型物流搬运（如，使用制造单元）使较小的转运批次成为可能，从而增大了周期时间在批次分离上的收益。（328|329）

例子：缩短周期时间

SteadyEye 是一家商用照相机支架制造商，以接单生产方式向电影行业供应产品。近来公司感到忧虑，因为它的客户提前期不再具有竞争优势。SteadyEye 从两周的订单时段之后，提供 10 周的提前期。（例如，如果订单在 1999-09-05 至 1999-09-18 的两周间隔内任何时候接收，它的预计交期为 1999-09-18 后的第十周。）然而，主要竞争对手在下单后提供五周的提前期。更糟糕的是，SteadyEye 的库存水平和平均周期时间（现在为九周）都达到了历史最高水平，客户服务水平（准时交付的订单的比例）很差（低于 70%）并且还在下降。

SteadyEye 的业务始于客户订单的进入，一名职员每天处理一次。很让这名职员沮丧的是，大部分订单似乎都在两周时段的末尾到来，使得她即便超时很长时间也落在后面，并且每两周这种情况就出现一次。使用最新的客户订单，ERP 系统生成一份每日的采购订单和分派清单。这些单据被送到各个加工中心，尤其是重要的组装区因为部件在那里结合来完成订单。不幸的是，由于需要的部件未送达，清单实现不了的情况常常发生。

SteadyEye 制造腿、臂及相机支架的其他结构组件，它还生产进入控制组装线的齿轮与变速箱。所有的电动机和电子器件都从外部供应商采购。原材料和子装配件在接收码头处接收。棒料被锯成适合各种齿轮的长度，在堆装栈板由叉车送到磨床作业。由于磨床处的换产时间很长，加工批次非常大。其他的作业包括钻孔、研磨和抛光。抛光机很快，所以只有一个。不幸的是，它也难于调整，停机时间也很长，所以堆积了许多要报废的部件。热处理作业要三小时，那里有一座可容纳约 1,000 个部件的巨大的烤箱。由于大多数加工批次都比单一订单的大，每个作业之后部件返回到一个中间库存点。

SteadyEye 问题的根源在于过量个周期时间。从工厂物理学的角度看是变动性（到达与加工）和利用率的结果。因此，改善政策必须关注这两个问题。

首先，到达变动性被订单处理体系不必要地放大了。通过建立一个所有订单预报同一个交期的两周的时间窗口，系统怂恿了客户和销售工程师的延迟。（如果不能较早地送出，为什么要在时间窗口后期结束之前接收订单？）其结果是救火式的短期行为导致订单爆发式地到达到系统，因而极大地提高了有效到达 CV。幸运的是，这个问题可有简单地取消订单窗口而补救。更好的政策是预报第 t 天接收的订单将在第 $l + t$ 天（其中 l 为提前期，我们希望它降到五天或更短）送到。通过稍晚些时候吸收订单的主生产计划，订单仍可以在系统内成批，但这样就对客户有着透明度。

其次，对有效加工时间变动性的分析显示，抛光机有着约 7 的巨大 c_e^2。这又进一步倍抛光机的利用率加重，因为这个利用率在考虑各种扰动后超过了 90%。故，一项有吸引力

的改善措施是分析影响抛光机的参数来寻找缩短调整时间的办法。这也减少了报废和加速小批加工任务来替换报废品的需要。净效果是降低了瓶颈作业的和利用率；它将显著降低排队时间，进而还有平均周期时间。由于这些量度还将削减周期时间变动性，它们使客户周期时间的缩短甚至大于平均周期时间的缩短。（329|330）

系统中变动性与周期时间的另一个大的来源是成批，我们接下来转而讨论它。成批受出于物料搬运和加工的考虑驱动。转运批量很大（常常是满栈板），是因为制程相互远离故而叉车的能力不允许频繁运输。因此，一项吸引人的政策是将制程纳入接近组装线的制造单元。它加上在物料搬运设备（如，传送带）上的投资，使得转运批量降至一成为可能。加工批量很大，是因为长的换模时间。因此，合乎逻辑的改进步骤是实施强有力的换模削减项目（如，快速换模（single minute exchange of die，SMED）技术，见 Shingo 1985）。既然按原来四分之一（by a factor of four）地压缩换模时间不是不可能的，这些步骤能使 SteadyEye 以 75% 或更多地降低加工批量。

除了这些在制程本身的改进，系统的一些变更也能进一步压缩周期时间。人们可能会限制 ERP 用于提供外购件的采购订单以及生成“计划的订单”却不将这些转换成实际的加工任务。需要一个单独的模块将订单纳入加工任务，使得在仍满足交期的同时同类的订单能一起处理（共享磨床处仍非常显著的换模时间）。这个模块的机制将在第十五章给出。

另外，有必要将一些通用组件由接单生产转换成备货生产。现在以大批次存储剩余许多部件的 Crib 将转化为这些部件的存货。（The crib that is now storing remnants of large batches of many parts would be converted to storage of stocks of these parts.）由于批量大为减小，所有其他的部件将永远不会进入 crib，却将以已制品被使用。故而，尽管择出部件（那些消除其周期时间将显著压缩客户提前期的通用件）的存货水平会上升，crib 的总体存货水平将显著下降。

净效果是这些变更将充分地压缩周期时间。周期时间从平均 10 周降到少于两周也不是一个不可能的预期。如果该企业能努力实现这点，它可以使其制造运营从有竞争力的里程碑变成战略优势。

更详尽的压缩周期时间的例子，参见第十九章。

9.6.3 提高客户服务水平

以运营的术语来讲，满足客户需求主要与提前期（快速反应）和服务水平（准时交付）相关。如我们早先提到的，大幅削减提前期的一种方法是将备货生产体系改造成接单生产体系，或由制造、存储通用组件并且接单组装来不完全地实现。我们在第十章更详细地讨论这种方法。

对于接单生产体系的产线，提前期定律指出

$$\begin{aligned}\text{提前期} &= \text{平均周期时间} + \text{安全提前期} \\ &= \text{平均周期时间} + z_s \times \text{周期时间的标准差}\end{aligned}$$

其中 z_s 是随着目标服务水平上升而上升的安全因数。因此，缩短固定服务水平的提前期（或是提高固定提前期的服务水平）需要缩短平均周期时间和/或削减周期时间的标准差。缩短平均周期时间的政策已在前文给出。幸运的是，它们同样对削减周期时间标准差有效。我们同时也看到，有一些政策，如缩短长程的重工环路（reducing long rework loops），就只对削减周期时间变动性有效。（330|331）

例子：提高客户服务水平

SteadyEye 例子的焦点在于缩短平均周期时间。其中深层次的原因，当然是企业对其响应能力的担忧。但是，只说明提前期却不在这同时考虑服务水平就由问题了。承诺短的提前期，后来又不能实现，显然不是提高客户服务水平的方法。幸好，我们给出的改善建议能使系统既压缩提前期又提高服务水平。

例如，回忆起一项提议是降低抛光机处的报废，从而减少最终组装工站处加速小批部件来追赶批内其他部件的必要。这样做将会显著地降低周期时间的标准差以及平均周期时间。

因此，即使我们提高服务水平（即，提高安全因数 z_s），总的客户提前期仍将被缩短。其他的变动性削减方法有着类似的效果。

为了说明这一点，假设原来的平均周期时间为九周，其标准差为三周。10 周的提前期允许仅为标准差三分之一的安全提前期。$z_{0.33} = 0.63$，所以这导致仅仅约为 63%的服务水平，与观测值相符。

假设在所有的提前期压缩步骤实施之后，平均周期时间降到七个工作日（1.4 周），标准差降到一周半。在这种情况下，两周的提前期代表着安全提前期为 0.6 周，即 1.2 倍的标准差，服务水平为 88%。（可能更为合理的）三周的提前期代表着 3.2 倍的标准差，服务水平为 99.9%。显著短于竞争对手的提前期和可靠的配送的组合将成为 SteadyEye 的竞争利器。

最后，我们指出变动性与周期时间削减的好处并非局限于接单生产体系。回忆起周期时间削减中的一项改进提议即是将一些部件改用备货生产。例如，假设 SteadyEye 存储一种通用的齿轮，其每周平均需求为 500 件，标准差为 100 件。该部件的生产周期时间为九周，标准差为三周。因此，补给间期的平均需求为 4,500 件，标准差为 1,500 件。如果我们一次生产 Q = 500 件，则可以使用第二章的（Q，r）模型计算出，为了确保 99%的补给率将需要 r = 7,800 的再定购点。这项政策将导致 3,555 件的平均持有库存。然而，如果上述提议的变动性削减方法将周期时间降低到 1.4 周，标准差降低到 0.4 周，再定购点将减少到 r = 1,080 并且平均持有库存减少到 631 件，降幅为 92%。这使得通用部件转向响应性更好的备货生产成为经济上可行的选择。

9.7 结论

本章主要关注于变动性对产线绩效的影响。要点总结如下：

1. *变动性降低绩效*。任何种类的变动性——加工、流动、成批——的升高，都有其代价。库存将累积，产出将下降，提前期将增长，或某些其他的绩效量度将受损。故而，几乎所有有效的改进运动中至少都有一些变动性削减措施。（331|332）

2. *变动性缓冲是制造生活的实情*。所有的系统都用库存、才能或时间来缓冲变动性。所以，如果无法削减变动性，你将不得不面对下列中的至少一项：

   a. 长的周期时间和高的库存水平
   b. 浪费的产能
   c. 损失的产出
   d. 长的提前期和/或差的客户服务水平

3. *柔性缓冲比固定缓冲更有效*。使产能、库存或时间不止一种用途可以降低给定的系统所需缓冲的总量。这条原则是现代制造实践中强调柔性或敏捷的根本原因。

4. *物料守恒*。流入工站的物料都将流出，不管流出的是良品还是废品。

5. *长期中，投料速率总是低于产能*。人们意图以 100%的产能作业，但考虑到包含超时、外包等因素的真实产能，它永远不会发生。最好是在系统“爆炸”之前就计划（*plan*）降低投料速率，以及其他无论如何都要降低的速率。

6. *产线前端的变动性比后端的破坏性大*。推式产线前端的高度加工变动性传递到下游并导致后续工站处的排队，而后端的高度加工变动性只影响本身的工站。因而，应用于产线前端的变动性削减一般比后端的效果明显。

7. *周期时间随利用率上升而非线性增长*。当利用率接近 1 时，长期 WIP 和周期时间趋向于无穷大。这意味着在高的利用水平处，系统绩效对投料速率非常敏感。

8. *加工批量影响产能*。加工批量与换模时间的交互作用很微妙。增大批量将提升产能，进而减少排队。然而，增大批量也增加了等待成批和批内等待时间。所以，串联批次情形的首要关注点是压缩换模时间，它将促进小的、有效率的批量的使用。如果无法压缩换模时间，周期时间将在某个大于 1 的批量处达到最小。类似的，取决于产能和需求，最有效率的并联加工批次将在一与制程所能容纳的最大数量之间。

9. *周期时间与转运批量成比例地增长*。等待成批时间和解批占周期时间的很大一部分。因此，减低转运批量是许多生产环境下可供使用的缩短周期时间的最简单的办法。

10. 匹配（*match*）可能是组装系统中延迟的重要来源。变动性引起的缺乏协同、不良排配或不良车间作业控制可能导致 WIP 的显著累积，并因而在组装处有延迟。

11. 诊断是工厂物理学的重要部分。工厂物理学的定律和概念有助于寻找制作系统中绩效问题的源头。解析的公式一定在这点上有用，然而诊断过程中最重要的还是直觉。

因为变动性在制造业中并未被很好理解，本章的观点是这本书中最有用的。在第三篇解决具体的制造管理问题时，将深深依赖于它们。