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

第八章 变动性基础

上帝不掷骰子。

——阿尔伯特·爱因斯坦

不要吩咐上帝该做什么。

——尼尔斯·波尔

8.1 引言

由里特定律（TH = WIP/CT）得出，周期时间长在制品水平高，或者周期时间短在制品水平低，可能达到相同的产出。当然，周期时间短在制品水平低的制造系统更优，是什么原因导致这种不同呢？大量实例表明，答案是变动性。

当运行在最佳情形（best case）下，第七章的硬币制造系统会以W_0= 4（临界在制品数量）的在制品水平实现最大产出（0.5 件/小时）；但是当处于实际最差情形（practical worst case）时，它要求 27 件在制品以实现 90%的产能（57 件在制品实现 95%的产能）；当处于最差情形（worst case）状态，甚至连 90%的产出都不可能实现。为什么存在这么大的差别呢？那就是变动性！

作为工厂的一部分，Briar Patch 制造厂有两个非常相似的工站。它们都是由一台机器组成，该机器以 4 件/小时的产出运行；它们都是为了满足相同的需求，平均工作量为 69 件/天（2.875 件/小时）；而且它们都存在定期的随机停机。然而，对于 Hare X19 机器的工站，停机不会频繁发生，一旦发生便会持续很长时间；对于包括 Tortoise 2000 机器的另一个工站，停机会更加频繁而时间会相应更短。两种机器都能达到 75%的利用率（即机器正常工作的时间）。因此，这两个工站的产量都是 4×0.75= 3 件/小时。既然产量相同且满足相同的需求，那它们是不是应该有相同的绩效——周期时间、在制品数量、提前期、客户服务水平？不是。事实证明，在所有这些绩效指标上，Hare X19 都要比 Tortoise 2000 差。为什么？答案依旧是变动性！

变动性存在于所有的制造系统中，并且对绩效会产生很大的影响。（248|249）由于这个原因，测度、理解和控制变动性的能力对于有效的制造管理是至关重要的。这一章，我们将介绍描述制造系统变动性的基本工具以及直觉的知识。下一章，我们将深入探索变动性降低系统绩效的方式以及对其进行控制的方法。

8.2 变动性与随机性

确切地讲，什么是变动性？书面定义是一组实体的不均匀性（*the quality of nonuniformity of a class of entities*）。例如，一组体重完全相同的人群在体重上就不存在变动性，然而体重差异很大的一组人在这方面则是高度变动性的。在制造系统中，很多属性存在变动性，如物理尺寸、运行时间、机器失效/修复时间、产品质量测度、温度、材料硬度、生产准备时间等等都是易于产生不均匀性的例子。

变动性与**随机性（randomness）**密切相关（但不相同）。因此，为了理解变动性的产生原因和影响，必须掌握随机性的定义和相关概念——**概率（probability）**。在这一章，我们将尽可能用非量化（loose）和直觉的方式来研究必要的概念。然而，为了保证精确性，有几点我们必须运用正式语言来描述可能性，通过**均值（mean）**和**标准差（standard deviation）**体现的**随机变量（random variable）**的概念和特性在这点上尤为必要。对这一术语感到陌生的读者在开始本章之前可以参考附表 2A 中对概率基础知识的回顾。

正如前面所提及的，最差情形和实际最差情形均代表绩效被变动性削弱的系统，然而最差情形下的变动性是可预测的——不良控制的结果——而变动性在实际最差情形里是不可预测的、随机的。为了理解其中的不同，我们必须区分可控的变动与随机的变动。

可控的变动（controllable variation）是由决策直接导致的。比如，当一个工厂生产几种产品时，产品描述（如，它们的物理尺寸、制造时间等）将存在变动性。同样，当材料按批量搬运到下一工序时，首先完成的部件将不得不比最后完成的等待更长的时间才被搬运，因此批量搬运情况下等待时间将比每次搬运一件更具有变动性。

与此相反，**随机的变动（random variation）**是由超出我们即时控制的事件产生的结果。例如，客户需求的时间间隔通常是无法控制的，因此我们应该在各个特定工站的波动范围内预测其工作量。同样，我们不知道一台机器何时会出故障。由于在作业完成之前必须等待机器恢复，这样便增加了有效加工时间（effective process time）。既然此类突发事件无法预测和控制（至少在短期内如此），机器中断就由于随机性增加了有效加工时间的变动性。

尽管两种变动都会影响工厂的正常生产，随机变动的影响更隐蔽且需要更精密的工具来刻画。由于这个原因，本章我们将主要集中于随机变动。

8.2.1 随机性的根源

很遗憾的是，随机性这个概念困扰着大多数人（也包括哲学家）。独立于与初始环境的事情怎么会发生呢？难道这没有违反因果对等的原则吗？（249|250）完全透彻地讨论这一哲学难题超出我们的能力范围，因而针对随机性的特性进行一些基本观察会更有趣。

随机性的一种解释是因为我们拥有的信息不完整（或不完全），系统因而表现出随机地行为。这一观点的潜在前提是：当我们理解了所有的物理定律并对宇宙有了完整刻画的时候，理论上来讲从那时开始我们便能确定地预测宇宙发展的各个细节。

随机性的另一种解释是宇宙的运行实际上就是随机的。也就是说，对宇宙和物理定律的完整刻画是不足以预测未来的，这些至多能对将要发生的事情提供统计估计，而且相同的初始环境也可能产生不同的结果。由于这一解释明显违反了因果对等原则，哲学界对此进行了严厉的批判，但是它的支持者已经指出，因果原则可以通过定义其他不受随机性影响的更为基本的量度来修正。[1]

在 20 世纪初的物理学界，这两个学派之间的争论变得异常激烈。爱因斯坦支持第一种观点（知识不足）并强调说“上帝从不掷骰子”；而玻尔和其他人则相信第二种观点（随机的宇宙 random universe），并建议爱因斯坦“不要吩咐上帝该做什么”。（见 普朗克 1936 年对这一争议的讨论）近些年，实验证据已趋向于支持随机性宇宙的看法，这对一些哲学家来说无疑是巨大的打击。

不管随机性是自然存在的还是知识缺乏所导致的，其影响都一样——生活中的许多方面包括制造管理，都是难以预测的。这意味着管理活动的结果永远得不到保证。事实上，在相同的环境下，运用相同的控制政策也会因时间不同而产生不同的结果。

这并非意味着我们应该放弃对工厂的管理，只是表明我们应该集中于寻找稳健政策，它

[1] 一些数量，如量子数（*quantum number*），定义明确并且确定了随机现象，如位置和速度的概率分布，而不是实际结果。

在大部分时间是起作用的。稳健政策不同于最适用特定环境的最佳政策，它几乎从来不是最佳的而通常是“比较好”的。相反，最佳政策可能会在针对其设计的特性环境里非常有用，但是在其他很多环境中表现极其不佳。面对随机性，管理者用来识别有效的稳健政策的最有力工具是良好的概率直觉。遗憾的是，这种直觉很不常见。本章的主要目的就是培养这样一种重要技能。

8.2.2 概率直觉

直觉在日常生活的许多方面起着重要作用，我们做的大多数决定都是基于某种形式的直觉。例如，汽车转向时我们会减速，这是经过一段时间的驾驶产生的直觉而不是对汽车物理结构的详细了解；我们决定是否融资购置住房，取决于对经济的直觉而非正式的经济分析；我们决定申请加薪的时机，主要根据对老板心情的直觉而不是基于对其心理倾向分析的深奥理论。

许多情况下，对于“首因（first-order）”效应我们的直觉是相当有用的。例如，当我们加速产线瓶颈（最忙的工站）而不改变其他部分时，我们期望获得更多的产品。（250|251）这种直觉通常来自于行动时自认为的不存在随机性的**确定性（deterministic）**世界。用概率统计的话说，这一推理主要基于**第一属性（first moment）**或其中的随机变量的**均值（mean）**。只要均值的变化（如，提高机器的平均速度）与其中的随机性高度相关，首因直觉通常表现良好。

而对于第二属性（即，包含随机变量变动的数量），我们的直觉就贫乏得多。例如，哪一个的加工时间有更多变动，一个部件还是一批部件？哪一种机器故障更具有破坏性，短期、高频还是长期、低频？在哪里削减加工时间的变动性更能改进产线的绩效，接近线首还是线尾？与发现加快瓶颈速率可以增加产出相比，以上问题和其他涉及变动性并与工厂运营相关的问题需要更为微妙的直觉。

因为对于第二属性人们通常缺乏已建立的良好直觉，他们常常会曲解随机现象。发生在学校的一个典型例子是，在第一次考试中成绩低的学生会在第二次考试中取得相对进步，而第一次得到高分的学生则会在第二次表现得相比较差。这是**向均值回归（regression to the mean）**现象的一个实例。第一次考试的极端分数（极高或极低）有可能，至少是部分由于随机性（如，侥幸或倒霉的猜题、测试当日头痛等等）。既然对于某个学生，随机性的影响不大可能连续两次导致极端现象，那么第一次成绩极端的学生有可能在第二次得到更为正常的分数。遗憾的是，许多老师就此认为终于获得差学生的进步却正在失去好学生。现实中，简单的随机性也能很好地说明这种影响。

对向均值回归的总体趋势的曲解也会发生在制造经理之中。在特别低的产出时期后，经理可能会做出苛刻的评价和处分。当然，产量提升了。类似地，超额绩效和表扬之后，产量下降——很显著的原因是员工开始自满了。当然，只要随机性存在，即使没有任何变动，同样的行为——由好变坏和由坏变好——随时有可能发生。

除了前两个属性（均值和方差），随机现象还受到第三（偏态）、第四（峰态）甚至更高属性的影响。这些高级属性的影响通常不如前两个的显著，因此我们将仅集中在均值和方差。此外，正如上面所提到的，由于均值的影响相当直观而方差的影响隐蔽得多，我们将侧重于对方差的刻画。

8.3 加工时间变动性

工厂物理学中首先提到的随机变量是工站中工件的**有效加工时间（effective process**

time）。在此我们使用实际这个标签是因为指的是我们“看到”加工任务在工站处的时间。这样做是因为：从逻辑角度看，当工件在机器A处加工，机器B因等待而处于空闲时，工件事实上是正在加工还是因机器 A 的原因（正在修理、启动、因质量问题返工或等待作业员从休息中回来）而被延迟都无关紧要。（251|252）对于B而言，所有的影响都是一样的。因此，我们将这些和其他影响合并到一个总的变动性量度中。

8.3.1 变动性的量度与分类

为了有效分析变动性，必须要能够对其量化。主要通过统计学中的标准量度来定义工厂物理学中的一系列变动性的种类。

方差（Variance），通常用σ^2表示，和标准差都是绝对变化的量度，而标准差σ用方差的平方根来定义。但是，绝对变化通常不如相对变化重要。例如，10μm 的标准差对于两英寸长的螺栓表现出非常小的变动性，但是对于宽度均值只有 5μm 的芯片来说却是高度变动性。一个表示随机变量变动性的合理量度是标准差除以均值，称作**变异系数（coefficient of variation，CV）**。如果用t表示均值（用t是因为这里要考虑的主要随机变量是时间），用σ表示标准差，那么变异系数c就可以写为

$$c=\frac{\sigma}{t}$$

事实证明，许多情况下用**变异系数的平方（squared coefficient of variation，SCV）**表示更方便

$$c^2=\frac{\sigma^2}{t^2}$$

我们将广泛运用 CV 和 SCV 来表示和分析生产系统的变动性。当一个随机变量的变异系数小于 0.75 时称为**低度变动性（low variability，LV）**，当变异系数介于 0.75 与 1.33 之间时称为**中度变动性（moderate variability，MV）**，而当变异系数大于 1.33 则称为**高度变动性（high variability，HV）**。表 8.1 列举了这三种情况并举出实例。

表 8.1 变动性的种类

变动性类型	变异系数	典型情况
低（LV）	$c<0.75$	加工时间中无断供
中（MV）	$0.75\leqslant c<1.33$	加工时间中有短时间调整（如，换模）
高（HV）	$c\geqslant 1.33$	加工时间中有长时间断供（如，停机）

8.3.2 低度与中度变动性

提到加工时间，我们倾向于考虑一台机器或一个作业员在工件上所花费的真实时间（即，不包括故障失效时间和生产准备时间）。这些时间趋向于像钟形曲线那样的概率分布。图 8.1 显示了均值为 20 分钟，标准差为 6.3 分钟的加工时间的概率分布，。注意曲线下面的大部分面积是如何在 20 分钟左右对称分布的。这个例子的变异系数大约是 0.32，因此它在低度变动性（LV）的范围内，这也是大多数 LV 加工时间都有的钟形概率密度曲线的特征。（252|253）

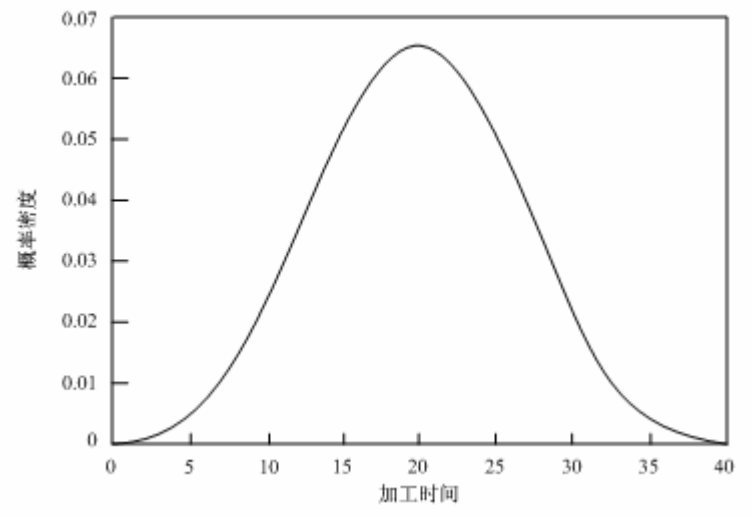

图 8.1 一个低度变动性分布

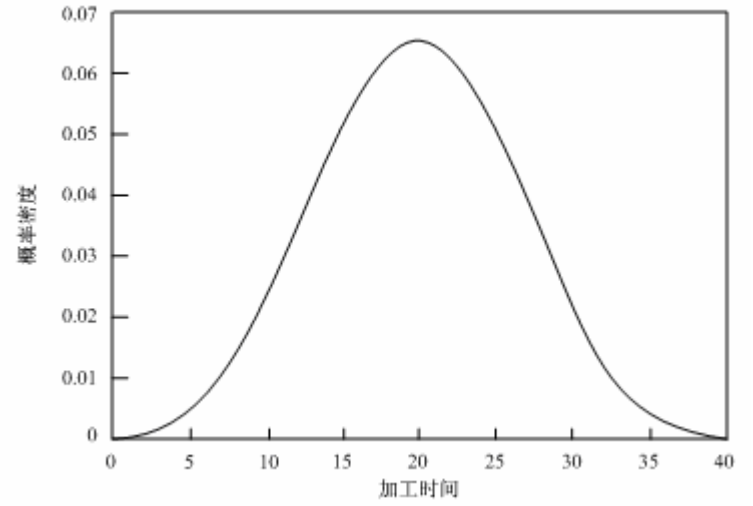

图 8.2 低度与中度变动性分布

现在考虑一种状况，加工时间均值为 20 分钟但 CV 大约是 0.75，即刚刚进入中度变动性。一个例子可能是手工作业的加工时间，大部分时间操作很简单但偶尔会有困难出现。图 8.2 对两种分布进行了比较。注意到 LV 情形下大部分概率都集中在均值 20 的左右，而在 MV 情情形下最可能出现的加工时间，大约是 9 分钟，实际上低于均值。然而，LV 图在 40 附近结尾，而 MV 图则一直延续到 80 附近。因此尽管均值相同，方差却极为不同。正如我们将看到的，这种差异对工站的运营绩效是至关重要的。

为了得到变动性的一系列运行效果，假设LV制程供给MV制程。刚开始的一会儿，MV制程能不费力地跟上。然而，一旦出现较长的加工时间，MV制程前将建立起待加工的工件队列。我们可能会不假思索地认为第二个工站中中一连串短的加工时间可能会消耗完队列，导致第一个工站出现空闲。当这种情况发生时，产能损失掉了而且不能“储存”起来在下一个较长的加工时间中使用。[2]

另一种对此的看法是，当一个制程供给另一个时，所有的输入必须输出，即**物料守恒**

[2] 在图 8.2 所示的中度变动性制程，20%的加工时间小于或等于 20 分钟，另有 20%大于或等于 31 分钟。为了保持均值在 20 分钟，大小两种数值都将出现。

（**conservation of material**）。除了第二个制程满溢（称作阻塞，并将在后续讨论）时关掉从第一个制程来的工件流，第二个制程前的工件数量将无限地增长。由于存在第二个工站比第一个工站运行快得多的时刻，并且平均输出速率一定等于平均输入速率，因此趋向于出现待加工工件队列。

我们将在 8.6 节对此进行完整地讨论。现在我们注意到有效加工时间的变动性越大，平均队列会越长。应用里特定律也意味着变动性越大，周期时间越长。（253|254）

8.3.3 高度变动的加工时间

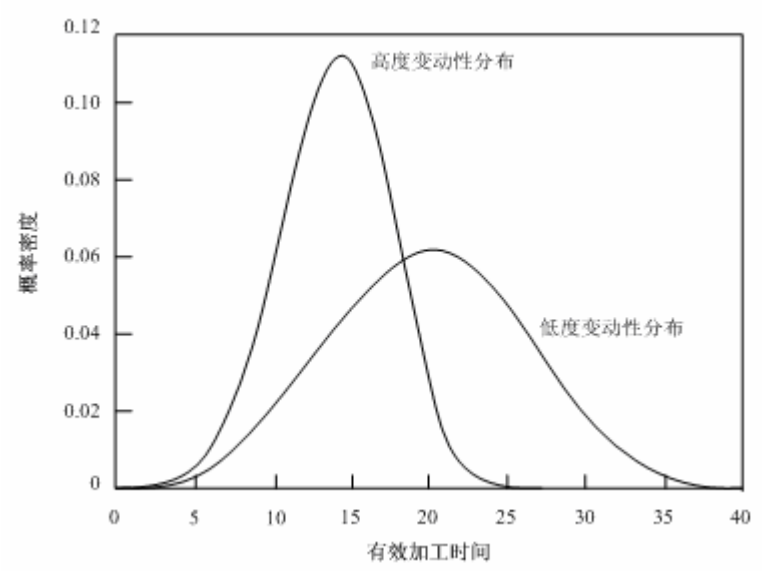

图 8.3 高度与低度变动性的比较

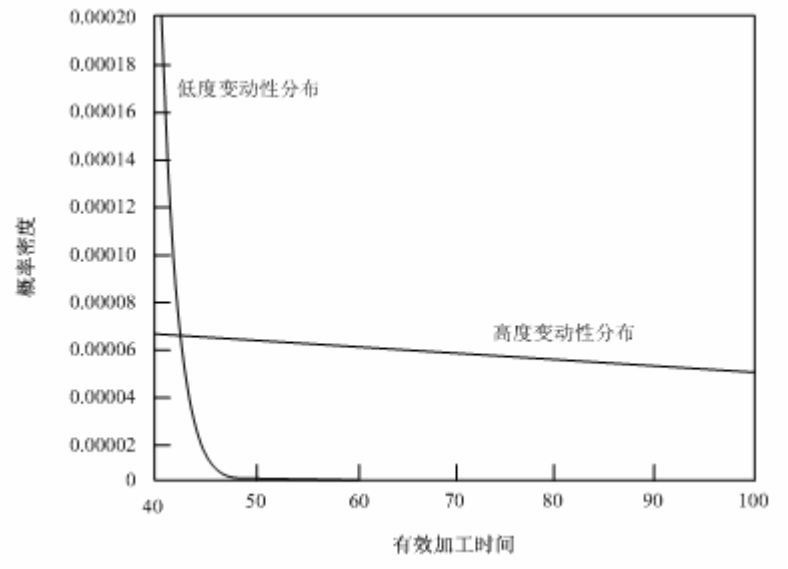

图 8.4 在 40 分钟以上的时间尺度上比较高度与低度变动性

可能很难想象 CV 大于 1.33 的加工时间，但却能很容易用如此大的变动性构建有效加工时间。假设没有断供时，某台机器的平均加工时间为 15 分钟，CV 为 0.225。这比前面的低度变动性情形具有更少的变动。但是现在假设机器一般在平均 744 分钟的生产后出现平均

时间为 248 分钟的断供，我们能证明（细节稍后给出）这时会导致 20 分钟（如前）的有效加工时间和值为 2.5 的特大 CV！图 8.3 对这个 HV 分布和前面的 LV 分布进行了比较。由于 HV 分布更高更陡，初看起来它似乎显示了比 LV 分布更少的变动。这是因为我们没有看到远期将出现的情况。一旦超过 40 分钟，图像改变了。图 8.4 在大于 40 分钟的时间尺度下对两种分布进行了比较。从中可以发现，LV 分布迅速下降接近为零，而 HV 分布则几乎没什么变化。事实上，它下降得非常缓慢。这说明极长的加工时间会以一个小的概率出现，这也是为什么 HV 的加工时间会在图 8.3 显示了较低的均值。大部分时间里大约会是 15 分钟，但是每 50 件中大约有一件为 17 倍。这使得均值抬高至 20 分钟左右，并驱使 CV 到 2.5。

该水平的变动性对产线的影响是非常严重的。例如，假设产出为每 22 分钟一件。由于包含断供的平均加工时间为 20 分钟，因此从产能方面考虑是不存在问题的。然而，250 分钟的断供将建立起差不多 12 件的队列。当机器重新开始工作后，该队列的消耗速度将是 $\frac{1}{15}-\frac{1}{22}\approx\frac{1}{47}$。因此，*假设不再有任何停工发生*，清空该队列的时间差不多是 536 分钟！若在此期间在此发生停工，队列又会增长。在用复杂方法（即，指数分布的失效时间）发现的普遍情况下，这种停工的概率是 $1-e^{-536/744}=0.51$。这说明队列清空之前，再次发生断供的可能性超过 50%。所以，平均队列中的工件数会多于 12，事实上会在 20 左右（如我们将在 8.6 节中所见）。（254|255）

8.4 变动性的起因

为了确定面对变动性时的生产系统管理战略，首先理解变动性的起因是非常重要的。制造环境，变动性最普遍的来源是：

- “自然”变动性，包括由作业员、机器和原材料的差异引起的加工时间的较小波动
- 随机断供（random outages）
- 生产准备（setup）
- 作业员可用性（operator availability）
- 重工（recycle）

以下将分别对它们进行讨论。

8.4.1 自然变动性

自然变动性是**自然加工时间（natural process time）**的内在变动性，包括随机停机、生产准备或其它外部作用。从某种意义上来说，由于该类含有源头未详细指出（如，作业员眼里的一片灰尘）的变动性，故而它是包罗万象的。因为许多未识别的变动性的来源都与作业员相关，所以手工作业一般比机器作业有更多自然变动性。但即使在大多数最严密控制的制程中，也一直存在某些自然变动性。比如，在完全自动作业中，材料的结构可能存在不同，导致加工速度有轻微变化。

我们分别以 t_0、σ_0 标记自然加工时间的均值和标准差，其变异系数可以表示为

$$c_0=\frac{\sigma_0}{t_0}$$

在大多数系统中，自然加工时间为 LV，因此$c_0 < 0.75$。

自然加工时间只是计算有效加工时间的起点。在任何真正的生产系统中，工站受各种**扰动（detractor）**的制约，如，停机、生产准备、无可用的作业员等等。正如前面讨论的，这些将会抬高有效加工时间的均值和标准差。我们现在提供一种量化这种影响的方法。

8.4.2 源于占先断供（故障）的变动性

在前面讨论过的高度变动性例子中，我们发现不定期故障会极大地抬高有效加工时间的均值和标准差。确实，在许多系统中，这是产生变动的最大原因。幸运的是，有许多实用的方法可以削弱它的影响。由于这是一个普遍问题，我们将进行详细讨论。

通常将故障停机看作是**占先断供（preemptive outages）**，是因为不论我们是否需要它都会发生（如，正当加工工件时，它就发生了）。停电、作业员因紧急情况离开、消耗品（如，切削油）断料）是占先断供的其他可能源头。（255|256）由于这些因素对产线的运行会产生类似的影响，因此用前面讨论的方法将它们合并起来当作机器停机来处理是讲得通的（如，包括当计算 MTTF 与 MTTR 时由这些原因产生的停工和其他真正停机）。我们将在下一节讨论**非占先断供（nonpreemptive outages）**（即，发生于加工任务之间，而不是作业过程中的中断）。

为了发现机器断供如何引起变动性，让我们回头看看 Briar Patch Mfg 的例子并提供一些用数字表示的细节。Hare X19 和 Tortoise 2000 有相同的加工时间均值和自然标准差，分别是$t_0 = 15$ 分钟、$\sigma_0 = 3.35$ 分钟。因此，两个工站的自然 CV 值$c_0 = \sigma_0 / t_0 = 3.35 / 15.0 = 0.05$。两台机器都会发生失效且长期可用率（即，正常运行时间的比例）均为 75%。但是，Hare X19 存在长时间、低频率的断供，而 Tortoise 2000 则为短时间、高频率的断供。特别是，Hare X19 的平均失效间隔（MTTF），用m_f表示，为 12.4 小时或是 744 分钟；平均恢复间隔（MTTR）用m_r表示，为 4.133 小时或是 248 分钟。Tortoise 2000 的 MTTF 为$m_f = 1.90$ 小时或是 114 分钟，MTTR 为$m_r = 0.663$ 小时或是 38.0 分钟。注意到，Hare X19 的失效时间、恢复时间都比 Tortoise 2000 的大三倍。最后，假设恢复时间是变动的，并且两台机器的 CV = 1.0（中度变动性）。

当计算平均产能时，工业中的大多数产能计划工具都会涉及随机断供。这一点通过计算**可用率（availability）**来实现，由m_f、m_r给出

$$A = \frac{m_f}{m_f + m_r} \tag{8.1}$$

因此，对以上两台机器，可用率 A

$$A = \frac{m_f}{m_f + m_r} = \frac{744}{744 + 248} = 0.75$$

假设用来说明机器不可用的时间比例的自然加工时间t_0产生了**平均有效加工时间（effective mean process time）**t_e，表示为

$$t_e = \frac{t_0}{A} \tag{8.2}$$

因此，在两种情形下，t_e = 20 分钟。回忆在第七章中，我们推导出工站的产能是机器数量 m 除以平均有效加工时间。所以当r_0表示自然产能（产出）时，有效产能（*effective capacity*）（产出）r_e为

$$t_e = \frac{m}{t_e} = A\frac{m}{t_0} = Ar_0 = 0.75\times 4 = 3 \text{ 件/小时} \tag{8.3}$$

因此 Hare X19 和 Tortoise 2000 有着相同的有效产能。既然在工业中所有用来分析故障的维护体系仅考虑对可用率和产能的影响，那么这两个工站通常被认为是相当的。

但是，当考虑变动性的影响，两个工站极为不同。为了找出原因，考虑它们作为产线的一部分时有何种行为。如果Hare X19 工站失效 12.4 小时（它的平均失效持续时间），它将需要保持 12.4 小时的在制品以防止下游饥饿。而另一方面，Tortoise 2000 需要少于以上 1/6 的在制品来应付平均长度的失效。由于失效有着随机的特性，所以下游缓冲中的在制品量必须随时维持，以防止产出的损失。很显然，使用Tortoise 2000 的产线与使用Hare X19 的产线相比，将能够保持较少的WIP来实现相同水平的防止效果，从而产生相同水平的产出。[3]（256|257）净效果是，Hare X19 的产线比Tortoise 2000 的产线效率低（如，在给定的在制品水平下产出较低，或者需要更高水平的在制品量和周期时间来达到相同的产出）。

前面已说明Hare X19 的CV为 2.5，该值是通过使用现在要描述的数学工具得到的。我们假设失效时间间隔服从指数分布（即，它们为MV）。[4]除了服从某种概率分布，我们不对修复时间做其他假设。定义修复时间的标准差为σ_r、CV为$c_r = \sigma_r / t_r$。在这个例子中，c_r为 1.0（即，我们假定修复时间具有中度变动性）。

在这些假设下，我们可以计算有效加工时间的均值、方差和变异系数平方（分别是t_e、σ_r^2和c_e^2）

$$t_e = \frac{t_0}{A} \tag{8.4}$$

$$\sigma_e^2 = \left(\frac{\sigma_0}{A}\right)^2 + \frac{(m_r^2 + \sigma_r^2)(1-A)t_0}{Am_r} \tag{8.5}$$

$$c_e^2 = \frac{\sigma_e^2}{t_e^2} = \sigma_0^2 + (1+c_r^2)A(1-A)\frac{m_r}{t_0} \tag{8.6}$$

[3] 实际上，由于这些只是平均的故障期，使用Hare X19的产线需要超过12.4小时的WIP，而使用Tortoise 2000的产线需要超过 4.133 小时的 WIP。然而不变的是，使用 Hare X19 的产线需要多得多的 WIP 来达到同样的产出。

[4] 在实践中这通常是一个好的假设，尤其适用于那些诸如由新旧设备结合而成的复杂设备。这样，指数分布的无机已经趋于在有旧的或者新的组件的失效引发的任何断供的时间内成立。

有效加工时间的 CV 值c_e可以通过给c_e^2开方来计算。

注意到（8.4）式给出的平均有效加工时间只依赖于平均自然加时间和可用率，因此两个工站的值是一样的，为：

$$t_e = \frac{t_0}{A} = \frac{15}{0.75} = 20.0 \text{ 分钟}$$

但是（8.6）式中有效加工时间的 SCV 不仅仅依赖于平均加工时间与可用率。为了理解其中的影响，我们可以将（8.6）式改写为

$$c_e^2 = c_0^2 + A(1-A)\frac{m_r}{t_0} + c_r^2 A(1-A)\frac{m_r}{t_0}$$

第一项源于加工中的自然变动性（未说明的）。第二项则来源于随机断供的事实。即使断供本身（即，修复时间）是恒定的（即，$c_r = 0$），第二项依然存在。例如，一直需要同样加工时间完成的定期检修，其$c_r^2 = 0$。因此消除修复时间的变动性对削减这一项没有用处。然而最后一项很明显是由于修复时间的变动性，而且将随变动性被消除而消失。可用率给定时，后两项都会随着m_r的增加而增大。因此，其他量度均相等时，较长的修复时间会引发较大的变动性。

将例子中的数值代入这些式子，得

$$c_e^2 = 0.05 + (1+1)0.75(1-0.75)\frac{248}{15} = 6.25$$

或是$c_e = 2.5$，表明 Hare X19 正是落在 HV 范围内。而 Tortoise 2000 有

$$c_e^2 = 0.05 + (1+1)0.75(1-0.75)\frac{38}{15} = 1.0$$

因此c_e=1，这说明它落在 MV 范围内。

因此使用 Hare X19 的产线比使用 Tortoise 2000 的具有更高的变动性。我们将会在第 8.6 节详细探讨它如何影响 WIP 和周期时间。

通过分析得出结论，即，当可用率相同时，高频率、短时间的断供比低频率、长时间的断供更可取。这个结论可能与我们的非概率直觉有点相反，直觉可能暗示（suggest）我们的是，每月一次大的头痛可能比每天的小阵痛好。但是从逻辑上来讲，每天的阵痛更容易医治。

这是一直有潜在价值的洞察力，因为在实际中我们会将长时间、低频率的失效转化为短时间、高频率的失效（如，通过预防性维护程序）。但是，为防读者变得自满，完全没有失效会比短时间、高频率的失效更好。没有什么理由值得降低为提高整体可靠性而付出的努力。

8.4.3 源于非占先断供的变动性

非占先断供（Nonpreemptive outages）代表不可避免要发生的停机；但其何时发生，我们能施加一些控制。相反地，占先断供可能由机器的灾难性故障或剧烈的失调引起，迫使当前工件完成作业与否都要暂停。非占先断供的例子有工具变钝需要替换，电路板的载具磨损等。在类似的状况下，我们可以一直等到当前工件完成作业后再停止生产。

生产工艺变化（如，更换外罩）而非产品变化引起的工艺转换（换模）可以被视为非占先断供。产品变化（如，准备生产一个新部件）引起的转换较容易被控制，且是第九章、第十章的主题。其他非占先断供，包括预防性维护、中途休息、作业员会议和（我们希望是）换班等。它们一般发生于不同的作业之间，而非某一项作业之内。非占先断供需要一些与占先断供有点不同的处理方法。因为它的最常见来源是机器换模，我们将据此建立讨论框架。然而，该方法适用于所有形式的非占先断供，正如我们对停机的分析适用于所有形式的占先断供那样。

与占先断供一样，通常的产能计算没能完全分析非占先换模的影响。平均产能分析只告诉我们短时间换模优于长时间换模，它不能评估当具有相同的有效产能时，短换模时间的低速机器与长换模时间的高速机器之间的区别。（258|259）

例如，决定是否用不需换模的低速、柔性机器替代需要定期换模的相对高速机器。机器 1，高速，平均可以每小时生产一件，但是平均生产四件需要一次两小时的换模；机器 2，柔性，不需要换模但速率较低，平均生产一件需要 1.5 小时。机器 1 的有效产能 r_e 是

$$r_e = \frac{4}{6} = \frac{2}{3} \text{ 件/小时}$$

由于它是单机工站，有效加工时间就是有效产能的倒数，因此 t_e = 1.5 小时。由此可知，机器 1 和机器 2 具有相同的有效产能。

只考虑平均产能的传统产能分析认为这两台机器是相当的，因此不会对用机器 2 替换机器 1 提供任何支持。但是，前面关于机器停机的工厂物理处理方法显示，在评估存在停机的机器时考虑变动性是非常重要的。其他量度均相等时，机器 2 将比机器 1 有着更小的变动有效加工时间（即，机器 1 上每生产四个工件就会有一个长时间的换模加入其有效加工时间）。因此，用机器 2 代替机器 1 将有助于降低加工时间 CV 从而使产线更有效率。这种变动性削减效应给 JIT 的短换模时间倾向提供深入支持，它也是柔性制造技术的显著动机。

然而，对柔性收益的评估可能是微妙的（subtle）。上面的条件“其他量度均相等时”要求机器 1、2 的自然变动性相同（即，确保机器 1 的换模不会显著地增加其有效加工时间 CV）。但是如果柔性机器也有较多的自然变动性呢？在这种情况下，我们必须计算并比较两台机器的有效加工时间 CV。

计算有换模的机器的有效加工时间CV时，我们首先需要关于自然加工时间的数据，即均值 t_0 和方差 σ_0^2（等效地，由于 $\sigma_0^2 = c_0^2 t_0^2$，也可以使用均值 t_0 和CV c_0）。接下来我们必须对换模进行刻画。主要通过假设两次换模之间机器平均加工 N_s 个部件（或加工任务），换模时间均值为 t_s、CV为 c_s。我们还假设每个工件完成作业后发生换模的概率相同。[5]也就是说，如果两次换模之间平均生产 10 件，那么无论上一次换模之后已经生产了多少件，当前工件完成加工后发生换模的可能性为 1/10。

在这些假设下，均值、方差与有效加工时间的 SCV 的计算公式分别为

$$t_e = t_0 + \frac{t_s}{N_s} \qquad (8.7)$$

[5] 这个假设意味着两次换模之间加工的工件数量是中度变动性的（即，均值与标准差相等）。也可以对其他考虑换模间隔时间变动性的的假设进行类似的分析。

$$\sigma_e^2 = \sigma_0 + \frac{\sigma_s^2}{N_s} + \frac{N_s - 1}{N_s^2} t_s^2 \qquad (8.8)$$

$$c_e^2 = \frac{\sigma_e^2}{t_e^2} \qquad (8.9)$$

为了说明这些公式的作用，考虑另一个对两台机器进行比较的案例。机器 1 是柔性的，无换模，但加工时间有变动性。特别地，它的平均自然加工时间$t_0 = 1.2$ 小时、CV$c_0 = 0.5$。（259|260）机器 2 在两次换模之间平均产出 $N_s = 10$ 件，平均自然加工时间$t_0 = 1.0$ 小时、CV$c_0 = 0.25$，平均换模时间$t_s = 2$ 小时、CV$c_s = 0.25$。哪台机器较好？

首先，考虑有效产能。机器 1

$$r_e = \frac{1}{t_0} = \frac{1}{1.2} = 0.833$$

而机器 2

$$r_e = \frac{1}{t_e} = \frac{1}{1 + \frac{2}{10}} = 0.833$$

在这方面两台机器是等效的。因此，哪台机器较好的问题就变成，哪台机器的变动性较低？

应用（8.9）式，对机器 1 我们可以计算出$c_e^2 = c_0^2 = 0.25$，机器 2 为 $c_e^2 = 0.31$。因此，无换模、变动性较多的机器 1 比有换模、变动性较少的机器 2，有着较低的总体变动性。

当然，该结论是例子中具体数据的结果。柔性机器不总是有着较低的变动性。例如，考虑如果机器 2 在平均生产 $N_s = 5$ 件之后有较短的换模时间（$t_s = 1$ 小时）会发生什么。有效产能保持不变。然而机器 2 的有效变动性显著降低，$c_e^2 = 0.16$。这种情况下，有换模的机器 2 是较好的选择。

8.4.4 源于重工（recycle）的变动性

制造系统中变动性的另一个主要来源是质量问题。用来分析的最简单的质量案例是单一工站处的返工。当工站执行一项任务并检查该任务是否被正确完成时，它就会发生。如果没有正确地完成，那么该任务就要重复。倘若把花费在“校正工件”上的时间看作是停机，那就会很容易发现这种情况与非占先断供是等效的。因此，重工（rework）与换模有着类似的效应，也就是，它即损失产能，又极大地提高了有效加工时间的变动性。

与停机、换模一样，传统的减少重工的理由是防止有效产能的损失（即，减少浪费）。当然，与停机、换模的传统的分析一样，这种方法会把两台有效产能相同、返工比例不同的机器看作是等效的。但是，类似分析换模的方法表明当重工比例上升时，有效加工时间的 CV 会随之增大。因此，较多的重工意味着较大的变动性。较大的变动性会导致较多的阻塞、WIP 和周期时间。所以，这些变动性与产能的损失一起发生影响，使得重工成为实在的破坏性问题。我们将在第十二章更详细地讨论这个质量与运营之间的重要界面（interface）。

8.4.5 变动性相关公式的汇总

占先断供与非占先断供情形下t_e、σ_e^2与c_e^2的计算公式总结在表 8.2 中。注意，如果遇到同时包括占先、非占先断供（如，同时发生停机和换模）的例子，这些公式必须衔接地使用。（260|261）例如，我们始于自然加工时间参数t_0、c_0^2，然后运用占先断供公式计算有效加工时间的t_e、σ_e与c_e^2将机器失效的影响考虑进来。最后在非占先断供的公式中，通过t_e、σ_e与c_e^2取代t_0、σ_e来考虑环换模的影响。最终的均值t_e、标准差σ_e和 SCVc_e^2将因此被“膨胀”而反映两种类型的断供。

表 8.2 计算有效加工时间参数的公式的汇总

情况	自然	占先	非占先
例子	可靠的机器	随机失效	换模；重工
参数	t_0, c_0^2 （基本参数）	基本参数 以及 m_f, m_r, c_r^2	基本参数 以及 N_s, t_s, c_s^2
t_e	t_0	$\frac{t_0}{A}, A=\frac{m_f}{m_f+m_r}$	$t_0+\frac{t_s}{N_s}$
σ_e^2	$t_0^2 c_0^2$	$\frac{\sigma_0^2}{A^2}+\frac{(m_r^2+\sigma_r^2)(1-A)t_0}{Am_r}$	$\sigma_0^2+\frac{\sigma_s^2}{N_s}+\frac{N_s-1}{N_s^2}t_s^2$
c_e^2	c_0^2	$c_0^2+(1+c_r^2)A(1-A)\frac{m_r}{t_0}$	$\frac{\sigma_e^2}{t_e^2}$

8.5 流动变动性（Flow Variability）

以上所有讨论只集中于单一工站的加工时间变动性。但是在产线上，一个工站的变动性会通过另一种类型的变动性来影响其他工站的行为，我们称之为**流动变动性（flow variability）**。流动是指工件或部件在工站之间的转移。很明显如果上游工站有高度变动的加工时间，那么它供给下游工站的流也将是高度变动的。因此，为了分析变动性对产线的影响，我们必须刻画流动的变动性。

8.5.1 刻画流动变动性

研究流动的起点是工件到达单个工站。工件离开这个工站也就是到达下一个工站。因此，一旦我们刻画工件到达工站的变动性并确定它离开此工站（并因此到达下一个工站）的变动性，我们将能够刻画整条产线的流动变动性。

用来描述工件到达工站的第一个参数就是**到达速率（arrival rate）**，以单位时间的工件数量来量度。为了一致，到达速率的单位必须与产能相同。例如，如果我们规定产能的单位为件/小时，那么到达速率也必须表达为件/小时。正如既可以用平均加工时间t_e，又可以用

工站平均速率 r_e 来刻画产能，我们可以用**平均到达间隔时间（mean time between arrivals）** t_a 或平均到达速率 r_a 来描述工站的到达速率。（261|262）这两个量度恰好是简单的倒数关系

$$r_a = \frac{1}{t_a}$$

并且能够有着完全等效的信息。

为了使工站能够跟得上工件的不断到达，工站产能基本上要大于到达速率，即

$$r_e > r_a$$

事实上所有现实的例子中（即，存在变动性的），产能必须严格大于到达速率以保证工站不至于超负荷。我们将在下文较精确地检视其原因。

正如加工时间存在变动性，到达间隔时间同样是有变动性的。同加工时间一样，我们可以定义到达间隔时间为一个合理的变动性量度。如果 σ_a 为到达时间间隔的标准差，则到达间隔时间的变异系数 c_a 为

$$c_a = \frac{\sigma_a}{t_a}$$

我们将其称为**到达 CV（arrival CV）**，区别于**加工时间 CV（process time CV）** c_e。直观地看，低的到达 CV 表示有序的、节拍平稳的到达进程，而高的 CV 表示不平均的、或是“爆炸式”的到达。它们的不同之处如图 8.5 所示。到达 CV c_a，同平均间到达隔时间 t_a 一起概述了工件到达工站过程的基本情况。

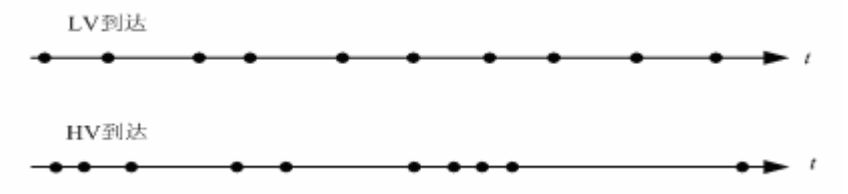

图 8.5 低度 *CV* 与高度 *CV* 的到达进程

下一步是刻画工件离开工站的情况。我们可以用与描述到达事件相似的量度，即平均离开间隔时间 t_d，离开速率 $r_d = 1/t_d$，离开 CV c_d。在工站 *i* 的全部为工站 *i*+1 的输入的串联产线中，工站 *i* 的离开速率必须等于工站 *i*+1 的到达速率，即

$$t_a(i+1) = t_d(i)$$

当然，在没有产出损失或重工的串联产线，每个工站的到达速率等于产出 TH。同样，工站 *i* 的离开变成工站 *i*+1 的到达的串联产线，工站 *i* 的离开 CV 等于工站 i+1 的到达 CV。

$$c_a(i+1) = c_d(i)$$

图 8.6 描述了这些关系。（262|263）

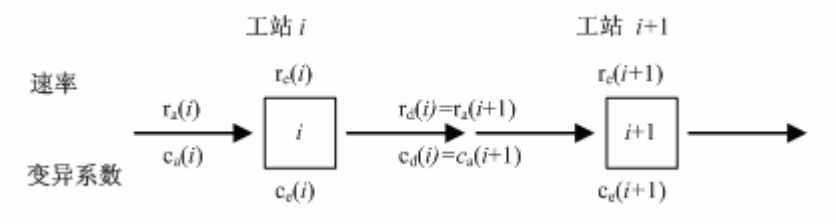

图 8.6 紧邻工站之间的变动性传递

最后一个决定流动变动性的问题是如何根据到达和加工时间的变动性来刻画离开的变动性。离开的变动性是到达和加工时间变动的共同作用结果。这两个影响因素的相对贡献由工站的**利用率（utilization）**决定。如前所述工站利用率用 u 表示，是从长期来看它处于运转的时间的比例。对于一个有 m 台相同机器的工站，利用率的正式定义为

$$u = \frac{r_a t_e}{m}$$

注意到这里 u 同到达速率和平均有效加工时间一同增长。利用率的上限为 1（即，100%），意味着有效加工时间必须满足

$$t_e < \frac{m}{r_a}$$

如果 u 接近 1，工站几乎总是繁忙的。因此，在这些条件下，离开间隔时间实质上同加工时间一致。因此，我们可以认为离开 CV 同加工时间 CV 一致（即，$c_d = c_e$）。

另一种极端情况是，当 u 接近 0 时，工站近乎空负荷。事实上每次工件被加工完，工站都要为下一个等待很长一段时间。因为加工时间只是离开间隔时间的一小部分，离开间隔时间基本等于到达间隔时间。因此，在这些条件下我们认为到达和离开 CVs 一致（即，$c_d = c_a$）。

一个好的、简单的在这两种极端情况下插值的方法是利用如下的利用率的平方：[6]

$$c_d^2 = c_e^2 u^2 + c_a^2(1-u^2) \qquad (8.10)$$

如果工站总是处于繁忙状态，则 $u = 1$, 进而有 $c_d^2 = c_e^2$。类似地，如果机器几乎总是闲置，则 u = 0, 进而有 $c_d^2 = c_a^2$。利用率的中间水平，$0 < u < 1$，离开 SCV c_d^2 是到达 SCV c_a^2 和加工 SCV c_e^2 的组合。

当工站处不止一台机器（即，$m > 1$）时，下式是估计 c_d^2 的合理方法（也有其他的方法；参见 Buzacott 和 Shanthikumar 1993）

[6] 注意到包含 CVs 的等式再次以它们的 SCVs 表示。

$$c_d^2 = 1 + (1-u^2)(c_a^2 - 1) + \frac{u^2}{\sqrt{m}}(c_e^2 - 1) \qquad (8.11)$$

注意当 $m = 1$ 此式简化成（8.10）式。

净效果就是，流动变动性与加工时间变动性一样，在实际状况中可能有很大的不同。利用与加工时间变动性相同的分类表，我们可以根据到达 CV c_a 将到达事件分类如下

低度变动性（LV）　　$c_a \leqslant 0.75$

中度变动性（MV）　　$0.75 < c_a \leqslant 1.33$

高度变动性（HV）　　$c_a > 1.33$

根据离开 CV c_d，离开事件可以用同样的方式分类。

例如，高负荷 LV 工站的离开事件趋向于 LV，而高负荷 HV 工站的离开事件趋向于 HV。（263|264）由 MV 到达事件供给的 MV 产线将产生 MV 的离开。所有这些离开事件依次转化成其他工站的到达事件，所以实际中各种类型的到达事件都会发生。

实际中 MV 到达事件发生是另一种情况是工站由多个源头供给。例如，热处理作业可能从多条不同的产线接受工件。当这种情况发生时，上一个到达之后的间隔时间并不能为下一个达到会在何时发生提供多少信息（因为工件可能从很多地方送来）。因此，到达间隔时间趋向于无记忆性（即，服从指数分布），因此 c_a 将接近 1。甚至从每个给定的源头过来的到达事件都可能是规则的（即，LV），所有到达叠合（*superposition*）将趋向于显示为 MV。

8.5.2 成批到达与离开

流动变动性的一个重要原因是**成批到达（batch arrivals）**。当工件形成批量以便运输到工站时，它就会发生。例如，假设一个叉车每次（8 小时）能运送 16 个工件到一个工站。因为到达事件总是无随机性地以这种方式发生，因此，人们可能无可非议地认为变动性和 CV 是零。

然而，从单个工件的角度看，批次中工件的到达间隔时间展示了一幅极为不同的图景。批次中第一个工件的到达间隔时间（即，从前一个到达算起）是 8 小时。其他 15 个工件间隔到达时间是 0。因此，平均到达间隔时间 t_a 为 1.5 小时（8 小时除以 16 个工件），这些时间的方差

$$\sigma_a^2 = [\tfrac{1}{16}(8^2) + \tfrac{15}{16}(0^2)] - t_a^2 = \tfrac{1}{16}(8^2) - 0.5^2 = 3.75$$

因此到达 SCV 为

$$c_a^2 = \frac{3.75}{(0.5)^2} = 15$$

通常，如果批量是 k，这种分析将得出结果 $c_a^2 = \mathrm{k} - 1$。

那么，究竟哪个是正确的，是 $c_a^2 = 15$ 还是 $c_a^2 = 0$？事实上系统表现于“中间的某处”。

因为成批混淆两种作用效果。第一种作用是成本身。这并不是随机性，而是如我们在第七章讨论过的最差情形，一种不良控制（*bad control*）。另一种作用就是成批到达本身的变动性（即，由成批到达 CV 来刻画）。我们将在第九章中更详细地讨论成批与变动性的关系。

8.6 变动性交互作用----排队

加工时间变动性与流动变动性的上述结果是刻画整个产线中变动性的影响的构建模块。现在我们将关注转移到评估这些类型的变动性对产线绩效量度，即 WIP、周期时间、产出的影响上。

为了达到这个目的，我们首先要观测到实际加工时间（包括有生产准备、停工等），一般只占工厂总周期时间的一小部分（5%~10%）。它被记载于众多已发表的调查中（如，Bradt 1983）。其他时间中的大部分都用来等待各种资源（如，工站、运输设备、机器作业员等）。因此，工厂物理学的一个基本议题就是其理解造成这些等待的根本性原因。（264|265）

关于排队的科学被成为**排队论（queueing theory）**。在英国，人们并stand in line，而是stand in a **queue**。因此，排队论是关于队列的理论。[7]因为工件诸如等待加工、等待转运、等待部件时都在“排队”，所以排队论是分析制造系统的一个有力工具。

排队系统（queueing system）合并了到目前为止的已经考虑的部分：到达进程、服务（即，生产）进程以及队列。到达进程可以包含单个的工件或成批工件。工件可以是相同的或有着不同的特征。到达间隔时间可以是恒定或随机的。工站可以是只有一台机器的或有多台并联机器，并且其加工时间可以是恒定或随机的。排队规则可以是先到先服务（FCFS）、后到先服务（LCFS）、最早交期（EDD）、最短加工时间（SPT）或任何形式的优先权。排队的空间可以是无限或有限的。排队系统的类型计划无限。

不管当下考察的排队系统属性如何，排队论的任务是用一系列的描述性参数来刻画系统的绩效量度。我们用这个方法对几个适用于大多数制造系统的排队系统进行描述。

8.6.1 排队系统的记号与量度

为了用排队论来描述单个工站的性能，我们设定以下的参数：

r_a = 工件到达工站的速度。在没有产出损失或重工的串联产线，每个工站处 $r_a = TH$

t_a = $1/r_a$ = 平均到达间隔时间

c_a = 到达 CV

m = 工站中并联机器的数量

b = 缓冲区的容量（即，系统中容许的工件的最大数量）

t_e = 平均有效加工时间。工站的速度（产能）为 $r_e = m/t_e$

c_e = 有效加工时间 CV

[7] **排队**（**queueing**）也是我们能想起来的唯一的包含五个元音的单词，这点信息可能对于填字游戏有帮助（which could be useful if one is a contestant on a game show）。

我们将关注的绩效量度有

p_n = 工站中存在 n 个加工任务的概率

CT_q = 在队列中的期望等待时间

CT = 在工站处的期望驻留时间（，即，等待时间加上加工时间）

WIP = 工站处的平均 WIP 水平

WIP_q = 队列中的期望 WIP 水平

除了以上参数，排队系统还以一系列特定的假设来刻画，包括有到达的种类和加工时间的概率分布、调度规则、阻行约定（balking protocols）、成批到达或加工等，而不论它是否由等待工站的网络构成，是否有单一或各种工件类型，及其他状况。*Kendall* 记号（*Kendall's notation*）表示一类单机、单工件类型的排队系统，它用四个参数刻画排队系统中的工站：（265|266）

$$A/B/m/b$$

这里，*A* 表示到达间隔时间的概率分布，*B* 表示加工时间的概率分布，*m* 表示工站中机器的台数，*b* 表示系统中能够容纳的工件的最大数量。也有一些其他类型，但 *A* 与 *B* 的典型取值为

D：常值（确定性）分布
M：指数（马尔科夫过程）分布
G：完全一般类型的分布（如，正态、均匀）

在许多情况下，队列长度并未被严格被限制（如，缓冲区可以很大）。我们用 *A*/*B*/*m*/∞或更简单的 *A*/*B*/*m* 表示这种情形。

例如，*M*/*G*/3 排队系统表示，有三台机器的工站，到达间隔时间服从指数分布，加工时间服从一般类型的分布，缓冲区容量无限大。

最开始我们将聚焦于 *M*/*M*/1 与 *M*/*M*/*m* 的排队系统，因为它们能提供需要的直觉并可能作为更加一般的系统的构建模块使用。然后我们将会考虑 *G*/*G*/1 与 *G*/*G*/*m* 的排队系统，原因是它们可以直接有效地用于工站建模。最后，我们将在 *M*/*M*/1/*b* 与 *G*/*G*/1/*b* 情形下讨论限制缓冲区容量时的状况。

简单地说，我们将限制所考虑的系统在单任务类型（即，一种产品）。当然，大部分的制造系统有多种产品。但通过单任务类型的模型，我们可以发展出对于制造系统中变动性的作用的关键洞察力。而且，这些模型有时可以用来近似地的估计多任务类型系统的行为。Buzacott 与 Shanthikumar（1993）的工作详述了实现的细节以及更精巧的多任务类型模型的开发过程。

8.6.2 基本关系

在考虑特殊的排队系统之前，我们注意到一些基本关系对于所有的单工站系统都成立（即，不论到达与加工时间分布的假设，不论机器数量，等等）。首先是**利用率（utilization）**，即工站处于运转状态的概率，表述如下

$$u = \frac{r_a}{r_e} = \frac{r_a t_e}{m} \tag{8.12}$$

第二个是驻留于工站的平均时间 CT 与驻留于队列的平均时间 CT_q 之间的关系。均值是可加的，因而有

$$CT_q = CT + t_e \tag{8.13}$$

第三个，应用里特定律于工站，可以得到 WIP、CT 与到达速率之间的关系：

$$WIP = TH \times CT \tag{8.14}$$

第四个，应用里特定律于队列，可以得到 WIP_q、CT_q 与到达速率之间的关系：

$$WIP_q = r_a \times CT_q \tag{8.15}$$

有了上面的关系式并知道四个绩效量度（CT、CT_q、WIP 或 WIP_q）中的任何一个，我们能计算出其他三个。（266|267）

8.6.3 M/M/1 队列

用于分析的最简单的排队系统之一是 *M*/*M*/1。这个模型假设指数分布的到达间隔时间、有着指数分布的加工时间的单台机器、先到先服务的规则，并且队列中的等待工件有无限大的可用空间。尽管不能精确地代表大多数的制造工站，*M*/*M*/1 队列易于处理并且为更复杂与现实的系统提供了有价值的洞察力。

分析 *M*/*M*/1 队列的关键是指数分布的无记忆性。理解其原因，我们考察刻画系统的未来需要什么信息。即，为了回答诸如一段特定时间之后系统出清的可能性如何之类的问题，我们应当知道系统当前状态的哪些方面？就是，我们需要知道的系统的准确的状态比如回答这样的问题在一段特定的时间后的系统是否是空闲的？或者是在一个加工任务被执行之前它的等待时间低于一个特定值的可能性如何？现在的议题不是如何计算出答案，而是简单地想，为了这么做，需要该系统的什么信息。

作为开始，我们需要到达间隔和加工时间的信息。因为我们假设它们服从指数分布的，所以要知道的是它们的均值（即，因为指数分布标准差与均值相等）。到达间隔时间的均值是 t_a，所以到达速率是 $r_a = 1/t_a$。加工时间的均值是 t_e。所以加工速率是 $r_e = 1/t_e$。

除了这些，惟一一项需要的信息是系统当前有多少加工任务。因为到达间隔时间与加工时间的概率分布都是无记忆性的，上一个到达事件发生了多久、当前工件已被加工多久与系统的未来行为无关。正因为如此，系统**状态（state）**可以用一个数字 n 来表示，意为当前系统中的工件数目。通过计算出现每种状况的长期概率，我们可以刻画所有的长期（稳定状态）绩效量度，包括 CT、CT_q、WIP 或 WIP_q。我们在以下的技术性注释中用这种方法分析 *M*/*M*/1 队列。（267|268）

技术性注释

定义 p_n 为发现系统处于状态*n*的长期概率（即，正在加工与排队等待的加工任务总数为

n)。[8]由于加工任务一次只来一件，机器一次只加工一件，系统状态一次也只能改变一个单位。例如，若工站处现在有n个加工任务，则状态只可能改变到n+1（来了一个）或n-1（走了一个）。系统从当前状态为n转移到n+1 的速率就是到达速率 r_a 。类似地，系统从当前状态为n转移到n-1 的速率就是加工速率 r_e 。系统的这种动态特性由图 8.7 显示出来。

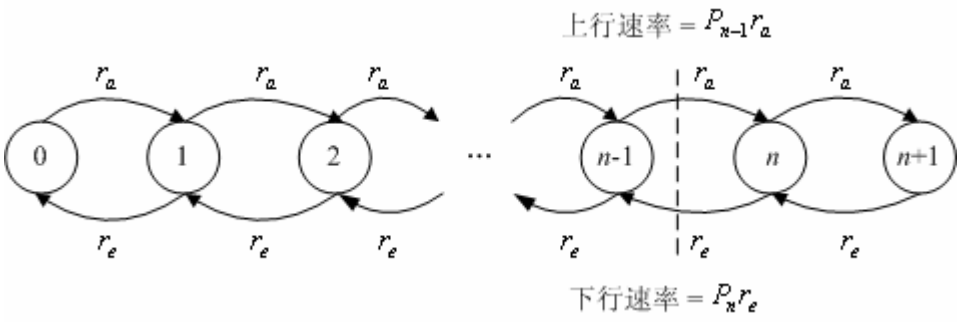

图 8.7 *M*/*M*/1 队列的系统状态转移图

系统从 n-1 转移到 n 的绝对（即，稳态）**速率**是 $p_{n-1}r_a$，也就是处于状态 n-1 的概率乘以从 n-1 转移到 n 的速率。类似的，系统从 n 转移到 n-1 的速率是 $p_n r_e$ 。为了使系统稳定，这两个速率必须相等（即，否则处于某个给定状态的概率将随时间“漂移”）。因此有，

$$p_{n-1}r_a = p_n r_e$$

或是（267|268）

$$p_n = \frac{r_a}{r_e} p_{n-1} = u p_{n-1} \qquad (8.16)$$

其中 $u = r_a t_e = r_a / r_e$ 是利用率，也就是在无阻塞时，长期内机器处于繁忙状态的时间比例。

按照利用率的定义，可以知道工站未作业的概率（长期时间比例）为 1-u。机器只在系统中无加工任务时才空闲，这就意味着 p_0=1-u。它给出一个 p_n 值。为了算出其他的，我们按 n = 1, 2, 3, ..., 写出（8.16）式

$$p_1 = u p_0 = u(1-u)$$

$$p_2 = u p_1 = u \cdot u(1-u) = u^2(1-u)$$

$$p_3 = u p_2 = u \cdot u^2(1-u) = u^3(1-u)$$

……

[8] 这些概率仅仅对于**稳态**（**steady state**）有意义。这也意味着我们只能从 p_n 的值中计算出长期量度。幸运的是，我们的关键量度 CT、WIP、CT_q 与 WIP_q 都是长期量度。分析排队系统的**暂态**（**transient**）（即，短期）行为很困难，我们不在这里讨论。

依此类推就有任意状态的表达式

$$p_n = u^n(1-u) \qquad n=0, 1, 2, ... \qquad (8.17)$$

这些 p_n 值都是概率并因而必须加和为一，故有

$$p_0 + p_1 + p_2 + \cdots = (1 + u + u^2 + \cdots)p_0 = 1$$

或是

$$p_0 = 1-u \qquad (8.18)$$

然而，若$u \geq 1$，括号中的加和就无穷大，并违背概率的性质。因此，为了使工站有稳定的长期行为（即，不会出现“爆炸”的队列），必须有$u < 1$（即，利用率严格小于 100%）。[9]

WIP（即，系统中的期望工件数目）是最易于计算的绩效量度了，对于 *M*/*M*/1 情形

$$WIP = \sum_{n=0}^{\infty} np_n$$

$$= (1-u)\sum_{n=0}^{\infty} nu^n$$

$$= u(1-u)\sum_{n=0}^{\infty} nu^{n-1} \qquad (8.19)$$

很容易得到$\sum_{n=1}^{\infty} nu^{n-1} = (1-u)^{-2}$，所以（8.19）式将产生一个WIP简明表达式。[10]（268|269）

8.6.4 绩效量度

依据技术性注释的结果，可以计算出多个稳态下的绩效量度。由（8.19）式，期望 WIP 为

$$WIP(M/M/1) = \frac{u}{1-u} \qquad (8.20)$$

它和里特定律可以推导出平均周期时间

$$CT(M/M/1) = \frac{WIP(M/M/1)}{r_a} = \frac{t_e}{1-u} \qquad (8.21)$$

然后由（8.13）式我们可以计算出排队平均时间

$$CT_q(M/M/1) = CT(M/M/1) - t_e = \frac{u}{1-u}t_e \qquad (8.22)$$

最后，对于队列中的 WIP，再次应用里特定律可以推导出

---

[9] 如果 $u < 1$，通过$1 + u + u^2 + \cdots = 1 + u(1 + u + u^2 + \cdots)$并令$x = 1 + u + u^2 + \cdots$，我们发现 $x = 1 + ux$。解 x 得$x = (1-u)^{-1}$。由于$xp_0 = 1$，这也得出 $p_0 = 1-u$，与前面考虑利用率时的结论一致。

[10] 这是因为$\sum_{n=1}^{\infty} nu^n$ 是$\sum_{n=0}^{\infty} u^n$ 的倒数，而后者等于$1/(1-u)$。和的倒数等于倒数的和，$\sum_{n=1}^{\infty} nu^n$ 就等于$1/(1-u)$的倒数，即$1/(1-u)^2$。注意这只在 $u < 1$ 时成立，且 $u < 1$ 已为队列稳定所需。

$$WIP_q(M/M/1) = r_a \times CT_q(M/M/1) = \frac{u^2}{1-u} \qquad (8.23)$$

观察到 WIP、CT、CT_q 与 WIP_q 都是 u 的增函数繁忙的系统比轻量负载系统表现出更多的阻塞，这并不奇怪。另外，对于一个固定的 u，CT 与 CT_q 都是 t_e 的增函数。因此，在一个给定的利用率的水平下，较慢的机器引起较长的等待时间。最后，注意到这些表达式的分母上是 1-u，当 u 接近 1 时所有这些阻塞的量度将发生“爆炸”。这就意味着当利用率接近100%时，WIP 水平与周期时间剧烈（即，非线性地）上升。我们将在第九章中更加详细地讨论它的影响。

例子：

回到前面提到的Briar Patch Mfg例子中，Tortoise 2000 的到达速率 2.875 件/小时（r_a = 2.875）。现在假设到达间隔时间服从指数分布（当工件来自于不同的地点时，它并非一个坏的假设）。另外，回想起生产速率是 3 件/小时（或者是 t_e = 1/3）并且 c_e = 1.0。因为有效加工时间的CV为 1，正如指数分布的性质，利用M/M/1 模型表示Tortoise 2000 是合适的。[11] 计算出的利用率是u = 2.875 / 3 = 0.9583，并且绩效量度在下面给出：（269|270）

$$WIP = \frac{u}{1-u} = \frac{0.9583}{1-0.9583} = 25 \text{ 件}$$

$$CT = \frac{WIP}{TH} = \frac{23}{2.875} = 8 \text{ 小时}$$

$$CT_q = CT - t_e = 8 - 0.3333 = 7.6667 \text{ 小时}$$

$$WIP_q = TH \times CT_q = 2.875 \times 7.6667 = 22.0417 \text{ 件}$$

我们看到 WIP 和 CT 远小于 Hare X19 在同等要求情况下的值。然而，为了给非指数分布的 Hare X19 建模相比，我们需要一个比 *M*/*M*/1 更一般的模型。

8.6.5 有着一般类型加工时间和到达间隔时间的系统

现实世界中大多数的制造系统不能满足 M/M/1 排队模型的假设条件。加工时间很少满足指数分布。当工站由加工时间并非服从指数分布的上游工站供给时，到达间隔时间也不大可能服从指数分布。为了说明到达间隔时间与加工时间都不是指数分布的系统，我们必须转向 *G*/*G*/1 队列。

不幸的是，在没有指数分布的无记忆特性来方便分析的情况下，我们不能计算出 *G*/*G*/1 队列的精确绩效量度值。但是，我们可以通过“两属性”法来近似地估计，也就是只使用到达间隔时间与加工时间的均值与标准差（或者 CV）。尽管能构造出这种估计结果糟糕的情形，但是它对于典型的制造系统（即，除了 c_a、c_e 远大于 1，或者 u 大于 0.95 或小于 0.1 的大多数情形）有着合适的精度。因为它表现地很好，这种近似成为几种可用的商业化的制

[11] 可是加工时间实际上并非服从指数分布的，因为 c_e = 1 只是失效与低度变动性的自然加工时间合成的结果。所以 *M*/*M*/1 队列并不精确，却是一个合适的估计。

造系统排队分析软件包的基础。

正如 M/M/1 情形，我们首先建立排队等待时间CT_q的表达式，再计算其他的绩效量度。

对CT_q的近似估计，最先由 Kingman 提出（见 Medhi 1991 的溯源），表示如下

$$CT_q(G/G/1)=\left(\frac{c_a^2+c_e^2}{2}\right)\left(\frac{u}{1-u}\right)t_e \qquad (8.24)$$

这个近似有着几个良好的性质。首先，它精确地适用于*M*/*M*/1 队列。[12]它恰好精确地适用于*G*/*G*/1 队列，尽管我们没有在这里详细讨论。最后，它可以完美地分成三项：无量纲的**变动性V**，**利用率U**，**以及时间T**，正如

$$CT_q(G/G/1)=\underbrace{\left(\frac{c_a^2+c_e^2}{2}\right)}_{V}\underbrace{\left(\frac{u}{1-u}\right)}_{U}\underbrace{t_e}_{T}$$

或者

$$CT_q=VUT \qquad (8.25)$$

我们将其称为 **Kingman 方程**或是 **VUT 方程**。根据它，我们可以看到如果 V 的值小于 1，那么 *G*/*G*/1 队列的排队等待时间与其他的阻塞量度都小于 *M*/*M*/1 队列。相反地，如果 *V* 值大于 1，*G*/*G*/1 队列阻塞的程度大于 *M*/*M*/1 队列。因此，与产线的实际最差情形相似，*VUT* 方程说明 *M*/*M*/1 队列是单个工站绩效的中等情形。（270|271）

例子：

让我们回到 Briar Patch Mfg 并考虑 Hare X19。记起这台机器有着高度变动性（$c_e^2=6.25$），并且再次假定到达间隔时间服从指数分布（即，$c_a^2=1$）。Hare X19 的利用率 u = 0.9583，因此，我们可以用 *VUT* 方程来计算期望的排队的时间

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

$$=\left(\frac{1+6.25}{2}\right)\left(\frac{0.9583}{1-0.9583}\right)20$$

= 1 667.5 分钟 = 27.79 小时

结果正如我们在本节开始的引言中讲到的。

现在，假设用 Hare X19 供给 Tortoise 2000。因为没有产出损失，Hare X19 和 Tortoise 2000 的到达速率相等；因为两个机器有着相同的有效加工时间，它们也有着相等的 $u=0.9583$。但是，要使用 *VUT* 方程，必须找到 Tortoise 2000 的到达速率 CV c_a。我们首先用关联方程（8.10）从 Hare 的 c_d 中找到离开 CV

[12] 当c_a与c_e都等于 1 时，第一项变成 1，其他的项就是 *M*/*M*/1 队列中的等待时间$CT_q(M/M/1)$。

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

$$= 6.25（0.9583\times 0.9583）+1.0（1- 0.9583\times 0.9583）$$

$$= 5.8216$$

因为 Hare X19 供给 Tortoise 2000，Tortoise 2000 的c_a^2就等于 Hare X19 的c_d^2。因此，Tortoise 2000 的期望排队时间将是

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

$$= \left(\frac{5.82+1.0}{2}\right)\left(\frac{0.9583}{1-0.9583}\right)20$$

$$= 1\ 568.97 \text{ 分钟 } = 26.15 \text{ 小时}$$

结果再一次如我们在引言中讲到的。

注意到尽管 Hare X19 的加工时间变动性远高于 Tortoise 2000，但 Tortoise 2000 的排队时间几乎和 Hare X19 的一样大。这种现象的原因是 Tortoise 2000 到达速率的高度变动性（$c_a = 2.41$）。如果由一个中度变动性的到达（$c_a = 1.0$）来供给 Tortoise 2000，那么它的绩效可由 *M*/*M*/1 队列表示，预计的平均排队时间为 7.6 小时。超额的时间（与阻塞）是上游的 Hare X19 的变动性传递的后果。

8.6.6 并联机器

VUT 方程提供了分析单机工站的工具。然而在现实的系统里，工站通常由多台并联机器组成。理由当然是需要不止一台机器来实现预定的产能。要分析和理解并联机器工站的行为，我们需要一个更一般的模型。

并联机器工站的最简单类型是，到达间隔时间和加工时间都服从指数分布（$c_a = 1$、$c_e = 1$）。（271|272）它相对于 *M*/*M*/1 排队系统。在这个模型中，所有的工件都排在一个单独的队列中等待下一个可用的机器（不像很多食品杂货店里，每一个收银员那里都排着一个独立的队列；而是像大多数银行里，所有的人只排成一个队列）。虽然 *M*/*M*/1 排队系统的稳态的概率可以很精确地计算，但却是繁杂而意义甚微的。更有用的是下面由 Sakasegawa（1977）提出的估算排队等待时间的近似形式，它不仅提供直觉，同时也相当精确（见 Whitt（1993）对它的优点和用处的讨论）：

$$CT_q(M/M/m) = \frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}t_e \qquad (8.26)$$

注意到当 $m = 1$ 时，这个表达式可以简化为（8.22）式，正是 *M*/*M*/1 队列排队时间的精确计算式。利用这个表达式，再与通用关系（8.13）~（8.15）一道，就可以得到$CT(M/M/m)$、$WIP(M/M/m)$与$WIP_q(M/M/m)$的表达式。

例子：

再考虑 Briar Patch Mfg 的例子。回忆起 Tortoise 2000 的有 $c_e = 1$ 的加工时间，因此可以用指数分布模型很好地估计。现在假设，Tortoise 2000 的到达速率是 207 件/天，并且有着服从指数分布的到达间隔时间（$c_a = 1$）。因为它已经超出了一台 Tortoise 2000 的产能，现在我们假设 Briar Patch Mfg 有三台 Tortoise 2000。

首先，考虑如果三个机器都有各自的到达流会发生怎样的情况。那就是每台机器都占总需求的三分之一，或者是 69 件/天（2.875 件/小时）。因为加工时间是三分之一小时，每台机器的利用率就是 0.958。这样，每台机器的状况与我们在 8.6 节中建立的模型完全相同，而那时候我们计算的平均排队时间是 7.67 小时。

现在假设将三台 Tortoise 2000 合并为一个工站，207 件/每天，或者 8.625 件/小时的总需求，到达一个由三台并联机器服务的队列。利用率是相等的，因为

$$u = \frac{r_a t_e}{m} = \frac{(8.625)(1/3)}{3} = 0.958$$

然而，现在队列的平均排队时间

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

$$= \frac{(0.958)^{\sqrt{2(3+1)}-1}}{3(1-0.958)} (\frac{1}{3})$$

$$= 2.467 \text{ 小时}$$

显著低于三台机器各有队列的情形。我们可以得出结论，当变动性和利用率相等的时候，使用并联机器的工站优于使用专用机器的工站。具体原因在食品杂货店里选错队的人都明白，专用机器处一段较长的加工时间将耽误在在队列中得到的每个人。当把队列合并，像在银行，经历一段较长加工时间的机器会被绕开，因此不会对平均排队时间造成那么大的破坏性影响。这是**变动性汇聚（variability pooling）**更加一般属性的例子。我们将在 8.8 节介绍变动性汇聚。（272|273）

8.6.7 并联机器和一般类型分布的时间

一个有着一般类型（非指数）分布加工时间与到达间隔时间的并联机器工站，可用 *G*/*G*/*m* 队列表示。为了建立此种状况的近似表达式，注意到（8.24）式可以写成

$$CT_q(G/G/1) = \left(\frac{c_a^2 + c_e^2}{2}\right) CT_q(M/M/1)$$

这里的 $CT_q(M/M/1) = [u/(1-u)]t_e$ 是 *M*/*M*/1 队列的排队等待时间。这就暗示了我们可以近似地将它用于 *G*/*G*/*m* 队列（见 Whitt 1983 的讨论）

$$CT_q(G/G/m) = \left(\frac{c_a^2 + c_e^2}{2}\right) CT_q(M/M/m) \qquad (8.27)$$

使用（8.26）式近似替代（8.27）式中的 $CT_q(M/M/m)$，得到如下的 *G*/*G*/*m* 队列排队等

待时间的近似形式：

$$CT_q(G/G/m)=\left(\frac{c_a^2+c_e^2}{2}\right)\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right)t_e \qquad (8.28)$$

（8.28）式是并联机器形式的 *VUT* 方程。*V* 项和 *T* 项与单机形式的（8.25）式相同，而 *U* 项不同。尽管看起来复杂，但它的求解不需要任何迭代算法，因而很容易由电子表格实现。这样就可能同时使用（8.28）式单工站近似估计与（8.11）式多工站“关联方程”来开发电子表格工具分析产线的绩效。

8.7 阻塞效应

到目前为止，我们仅仅考虑了没有限制队列长度的系统。事实上，在我们已经检视的任意系统，利用率达到 100%时平均队长（与周期时间）增长至无限大。然而在现实世界中，队列永远不会变成无限长。它们受空间、时间和运营政策的限制。因此，工厂物理学的一个重要主题就是有限队列空间的系统的行为性。

8.7.1 M/M/1/b 队列

考虑加工时间与到达间隔时间服从指数分布，如 *M*/*M*/1 队列那样，但系统（队列中与加工中）只够容纳 *b* 件加工任务的情形。在肯道尔标记中，它相当于 *M*/*M*/1/*b* 队列。这种系统的行为与 *M*/*M*/1 队列大多相同；唯一不同的是一旦系统饱和，到达进程将被终止。这时，我们称机器被**阻塞（blocked）**。*M*/*M*/1/*b* 模型代表了生产实际中一种非常普遍的状况。

例如，考虑一个由两个工站与其间有限容量缓冲区构成的制造单元。第一个工站处理原材料并送入第二个工站的缓冲区。（273|274）假设原料总是可用的（如，原材料是可足量供应的横木或金属片），则 *M*/*M*/1/*b* 队列是对第二个工站行为的良好近似估计。事实上，如果两个工站的加工时间都服从指数分布，这个模型将是精确的。这种类型的结构并不罕见。事实上，所有的看板系统都会表现出其固有的阻塞特性。

像 *M*/*M*/1/*b*，在存在阻塞的排队模型中，到达速率 r_a 有着与其在无限容量队列中不同的意义。这里，它代表假定系统未饱和时的潜在到达速率。因此 $u=r_at_e$ 不再是工站繁忙的长期概率，而是到达未被拒绝时的工站利用率。所以，*u* 可能等于或者大于 1。在接下来的技术性注释中我们来计算 *M*/*M*/1/*b* 队列的概率分布与量度。

技术性注释

正如在 *M*/*M*/1 队列中一样，我们定义 *M*/*M*/1/*b* 队列的状态为系统中加工任务的数量。然而，与 *M*/*M*/1 情形不同的是，*M*/*M*/1/*b* 队列有着有限数量的状态，$n=0, 1, 2, \cdots, b$。与在 *M*/*M*/1 队列中做的一样，我们可以这样表示 *M*/*M*/1/*b* 队列处于状态 *n* 的长期概率

$$p_n=u^n p_0$$

代数学知识显示，为了使 $p_0+\cdots+p_b=1$ 成立，必须有

$$p_0 = \frac{1-u}{1-u^{b+1}} \tag{8.29}$$

从而，

$$p_n = \frac{u^n(1-u)}{1-u^{b+1}} \tag{8.30}$$

注意到 b 趋于无穷大时（8.29）式和（8.30）式简化为 $M/M/1$ 队列的表达式（因为 b→∞时，$u^{b+1}\to 0$）。

$u\neq 1$ 时，（8.30）式都是成立的。对于 $u=1$ 这种特殊情形，所有的系统状态等可能地发生，并且有着相同的概率，故

$$p_n = \frac{1}{b+1}\ , n = 0, 1, \cdots, b \tag{8.31}$$

我们可以由下面的公司计算平均 WIP 水平

$$WIP = \sum_{n=0}^{b} np_n \tag{8.32}$$

因为系统只要未饱和就会接受到达，并且输入速率等于输出速率，我们可以由下式计算产出

$$TH = (1-p_b)r_a \tag{8.33}$$

对于 $u\neq 1$ 的情形，平均 WIP 与产出为（274|275）

$$WIP(M/M/1/b) = \frac{u}{1-u} - \frac{(b+1)u^{b+1}}{1-u^{b+1}} \tag{8.34}$$

$$TH(M/M/1/b) = \frac{1-u^b}{1-u^{b+1}} r_a \tag{8.35}$$

对于 $u=1$ 的情形，平均 WIP 与产出为

$$WIP(M/M/1/b) = \frac{b}{2} \tag{8.36}$$

$$TH(M/M/1/b) = \frac{b}{b+1} r_a = \frac{b}{b+1} r_e \tag{8.37}$$

对于以上两种情形，我们都可以利用里特定律来计算周期时间、排队时间和队列长度

$$CT(M/M/1/b) = \frac{WIP(M/M/1/b)}{TH(M/M/1/b)} \tag{8.38}$$

$$CT_q(M/M/1/b) = CT(M/M/1/b) - t_e \tag{8.39}$$

$$WIP_q(M/M/1/b) = TH(M/M/1/b) \times CT_q(M/M/1/b) \tag{8.40}$$

当把 $M/M/1/b$ 模型理解为两个串联工站组成的系统时，我们能从这些公式获得有益的洞察力。第一个工站被假设有着充足的原材料，所以永远不会饥饿。类似地，第二个工站总是能把它的产品移出（即，它永远不会被阻塞）。然而，两个工站之间的缓冲区是有限度的，

且容量等于 B。如果两个工站的加工时间均服从指数分布，第二个工站和缓冲区的行为可由 *M*/*M*/1/*b* 队列刻画，这里 $b = B + 2$。两个额外的缓冲区空间就是这两个工站本身。

注意到 M/M/1/b 队列的 WIP 总是比 M/M/1 系统的少。这是因为第二个工站存在阻塞，防止了 WIP 增长到超过 *b*。如果 *b* 很小，这种效应可能会非常明显。事实上，像有限容量缓冲区那样发挥作用的看板，就专门地意在防止 WIP 的积累。

然而，限制 WIP 有代价——损失产出。回想起在 M/M/1 情形中到达速率等于输出率。这是因为在稳态下，输入的也必须输出。当存在阻塞的情形下就不是这样了，因为输入速率与输出速率（产出）加上阻行速率（balking rate，到达被拒绝的速率）。由（8.35）、（8.37）式，可见

如果 $u \neq 1$，
$$TH = \frac{1-u^b}{1-u^{b+1}} u r_e < u r_e$$

或者，如果 $u = 1$，
$$TH = \frac{b}{b+1} r_e < r_e$$

最后这两个公式显示存在阻塞的系统的产出总是低于无阻塞的系统。进一步地，缓冲区容量越小，产出降低得越多

例子：

考虑一条由两台串联机器组成的产线。第一台机器完成一件加工任务的平均时间为 $t_e(1) =$ 21 分钟。第二台机器 $t_e(2) =$ 20 分钟。两台机器的加工时间均服从指数分布（$c_e(1) = c_e(2) = 1$）。两台机器之间有放置两件加工任务的空间，因此 b = 4（两个在缓冲区，两个是机器本身）。

首先考虑无限容量缓冲区的情形。由于第一台机器连续运转，所以第二台机器的到达速率等于第一台机器的加工速率。因此，第二台机器的利用率为 $u = r_a / r_e = \frac{1}{21} \Big/ \frac{1}{20} = 0.9524$。

（275|276）第二台机器的其他绩效量度可由 *M*/*M*/1 公式算出

$$WIP = \frac{u}{1-u} = \frac{0.9524}{1-0.9524} = 20 \text{ 件}$$

$$TH = r_a = \tfrac{1}{21} \text{ 分钟} = 0.0476 \text{ 件/分钟}$$

$$CT = \frac{WIP}{TH} = 420.18 \text{ 分钟}$$

现在，考虑有限容量缓冲区的情形。首先用 *M*/*M*/1/*b* 排队模型计算 TH。

$$TH = \frac{1-u^b}{1-u^{b+1}} r_a$$

$$= \frac{1-0.9524^4}{1-0.9524^5} \left(\frac{1}{21}\right)$$

$$= 0.039 \text{ 件/分钟}$$

我们现在可以计算 M/M/1/b 模型所表示的系统的不完全 WIP（*partial* WIP，标记为 WIPP），也就是第二台机器、两加工任务缓冲区以及第一台机器处的缓冲区。我们注意到第一台机器

处的 WIP 只有排队等待时才被计入 WIPP（即，第一台机器阻塞时）。正在第一台机器处加工中的 WIP 不包含在内，因为我们认为这些在制品“正在进入”*M*/*M*/1/*b* 模型所表示的系统。由（8.34）式，不完全 WIP

$$WIPP = \frac{u}{1-u} - \frac{(b+1)u^{b+1}}{1-u^{b+1}}$$

$$= 20 - \frac{5(0.9524^5)}{1-0.9524^5} = 20 - 18.106 = 1.894 \text{ 件}$$

产线的周期时间等于第二台机器处作为不完全 WIP 的时间加上第一台机器处的加工时间。注意我们没有考虑第一台机器处的排队时间，因为由原材料无限供应的假设它有可能无限长。

$$CT = \frac{WIPP}{TH} + t_e(1) = \frac{1.894}{0.039} + 21 = 69.57 \text{ 分钟}$$

再次应用里特定律，得到产线的 WIP

$$WIP = TH \times CT = 0.039 \text{ 件/分钟} \times 69.57 \text{ 分钟} = 2.71 \text{ 件}$$

有缓冲情形与无缓冲情形之间的对比非常明显。工站之间队列的限制显著地削减了 WIP 与 CT（超过 83%），但也降低了产出（但仅为 18%）。然而，产出下降 18%的损失可能超过库存成本节约的收益。这个例子突出说明了为什么看板不能简单地由削减缓冲区容量来实施。产出的损失往往是巨大的。既能降低 WIP 和 CT 又不会损失太多产出的唯一方法是，同时削减变动性（即，我们必须搬走礁石，而不仅仅是降低水位）。不幸的是，我们无法由 *M*/*M*/1/*b* 模型来检视变动性削减，因为它假设加工时间服从指数分布。我们将在下一节中讨论非指数分布模型。

由 *M*/*M*/1/*b* 模型我们还可以观察到，不论 r_a 和 r_e 为何值，有限容量的缓冲区将建立其稳定性。原因是 WIP，进而有 CT，在一个缓冲区容量有限的系统中不会“爆炸”。例如，假设前面那两台机器顺序颠倒一下，较快的供给较慢的。（276|277）如果缓冲区无限大，WIP 和 CT 都将增长到无限大（长期情况下）。但在缓冲区容量有限的情形，$u = 21 / 20 = 1.05$，因此

$$TH = \frac{1-u^b}{1-u^{b+1}} r_a = \frac{1-1.05^4}{1-1.05^5}\left(\frac{1}{20}\right) = 0.0390 \text{ 件/分钟}$$

不完全 WIP 为

$$WIPP = \frac{u}{1-u} - \frac{(b+1)u^{b+1}}{1-u^{b+1}}$$

$$= \frac{1.05}{1-1.05} - \frac{5(1.05^5)}{1-1.05^5} = 2.097 \text{ 件}$$

周期时间为

$$CT = \frac{WIP}{TH} + t_e(1) = \frac{2.097}{0.0390} + 20 = 73.78 \text{ 分钟}$$

最后，整个产线的 WIP 为

$$WIP = TH \times CT = 0.0390 \times 73.78 = 2.88 \text{ 件}$$

它比较快的机器处于下游时的 WIP 大一些，因为系统的到达速率大一些。不过，产出未受机器顺序的影响。后一个结果就是可逆性（*reversibility*），成立于有着多于两台机器、一般类型加工时间的产线（见 Muth 1979 的证明）。这是一个非常诱惑人的理论结果，然而，由于企业几乎没有机会逆着原来的产线机器顺序运行，因此它在实际中并不常见。

8.7.2 一般阻塞模型

为了分析变动性的影响，我们需要将 *M*/*M*/1/*b* 模型扩展到更加一般的达到间隔时间与加工时间分布类型。一般而言，这非常难。我们推荐有兴趣的读者查阅 Buzacott 和 Shanthikumar（1993，第四章）来获得较全面的描述。然而，通过像修正 *M*/*M*/1 队列来为 *G*/*G*/1 队列建模那样的方式，我们修正 *M*/*M*/1/*b* 队列，就可以获得一些有益的近似估计。

我们考虑以下三种情形：（1）到达速率小于生产速率（u <1），（2）到达速率大于生产速率（u >1），以及（3）到达速率等于生产速率（u =1）。

到达速率小于生产速率。首先应用 Kingman 方程和里特定律来计算无阻塞系统的预期 WIP，记为WIP_{nb}。

$$\begin{aligned} WIP_{nb} &\approx \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u}{1-u}\right)t_e + t_e \\ &= \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u^2}{1-u}\right) + u \end{aligned} \qquad (8.41)$$

回忆起对于 *M*/*M*/1 队列，$WIP = u/(1-u)$，因此（277|278）

$$u = \frac{WIP - u}{WIP}$$

我们可以用类似的方式通过WIP_{nb}计算出“修正的”利用率ρ

$$\rho = \frac{WIP_{nb} - u}{WIP_{nb}} \qquad (8.42)$$

然后在 *M*/*M*/1/*b* 的 TH 表达式中用将ρ替代（几乎）所有的 u 项，得

$$TH \approx \frac{1 - u\rho^{b-1}}{1 - u^2\rho^{b-1}} r_a \qquad (8.43)$$

联立通过 Kingman 方程（来计算ρ）与 *M*/*M*/1/*b* 模型，我们综合了变动性与阻塞。尽管这个表达式显著复杂于 *M*/*M*/1/*b* 模型的，但用电子表格计算也非常简单。此外，由于当$c_a = c_e$= 1 时可以很容易得出$\rho = u$，所以当到达间隔时间和加工时间服从指数分布时（8.43）式就恰好简化为（8.35）式。

不幸的是，期望 WIP 和 CT 的表达式麻烦得多。不过，对于小的缓冲区，WIP 将接近（但总是小于）缓冲区容量（即，*b*-1）；对于大的缓冲区，WIP 将接近（但总是小于）*G*/*G*/1 队列的。这样，

$$WIP < \min\{WIP_{nb}, b-1\} \tag{8.44}$$

根据里特定律，我们结合前面计算出的 TH 得到 CT 的近似范围

$$CT > \frac{\min\{WIP_{nb}, b-1\}}{TH} \tag{8.45}$$

不过这仅仅是一个近似范围因为 TH 的表达式本身就是近似值。

到达速率大于生产数率。在 *M*/*M*/1/*b* 队列早先的例子中，我们发现两台机器的顺序调换之后 WIP 的均值水平是不同的，但差别不大。这启发我们去考虑如何由机器顺序调换之后的 WIP，来估计在到达速率大于生产速率情形下的 WIP。当调换了机器顺序，我们发现产出进程变成了到达进程，反之亦然。因此利用率为 1/*u*（因为 *u*>1，所以 1/*u* 小于 1）。调换之后的产线的平均 WIP 可以这样估算

$$WIP_{nb} \approx \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{1/u^2}{1-1/u}\right) + \frac{1}{u} \tag{8.46}$$

在这里可以采用与 *u*< 1 时相同的方式，来计算调换之后产线的“修正的”利用率 ρ_R，得

$$\rho_R = \frac{WIP_{nb} - 1/u}{WIP_{nb}}$$

然后定义 $\rho = 1/\rho_R$，并像同前面那样计算 TH。一旦估算出了 TH，我们就可以通过不等式（8.44）、（8.45）分别得出 WIP 和 CT 的范围。

到达速率等于生产速率。最后，下面这个式子很好地估计了 $u = 1$ 情形下的 TH 值（Buzacott 和 Shanthikumar 1993）：（278|279）

$$TH \approx \frac{c_a^2 + c_e^2 + 2(b-1)}{2(c_a^2 + c_e^2 + b - 1)} \tag{8.47}$$

再一次地，利用这个 TH 近似值和不等式（8.44）、（8.45）得到 WIP 和 CT 的范围。

例子：

让我们回到 8.7.1 小节中的例子。第一台机器（加工时间为 21 分钟）供给第二台机器（加工时间为 20 分钟），并且两台机器之间有两件加工任务的缓冲区（因此 $b = 4$）。前面的例子中，我们假设加工时间服从指数分布后，发现限制缓冲区容量会导致产出下降 18%。一个抵消由于限制 WIP 而导致产出下降的方法是削减变动性。因此我们再考虑削减了加工变动性之后，两台机器都有 $c_e = 0.25$ 时的情形。

利用率仍然是 $u = r_a / r_e = \frac{1}{21}/\frac{1}{20} = 0.9524$，因此可以计算无阻塞时的 WIP 为

$$WIP_{nb} \approx \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u^2}{1-u}\right) + u$$

$$= \left(\frac{0.25^2 + 0.25^2}{2} \right) \left(\frac{0.9524^2}{1 - 0.9524} \right) + 0.9524$$

$$= 2.143$$

修正的利用率为

$$\rho = \frac{WIP_{nb} - u}{WIP_{nb}} = \frac{2.143 - 0.9524}{2.143} \quad = 0.556$$

最后，计算产出

$$TH \approx \frac{1 - u\rho^{b-1}}{1 - u^2\rho^{b-1}} r_a$$

$$= \frac{1 - 0.9524(0.556^3)}{1 - 0.9524^2(0.556^{3)}} \frac{1}{21}$$

$$= 0.0473$$

因此，这里的产出比无缓冲速率（$\frac{1}{21} = 0.0476$）下降了不到 1%。削减两台机器之间的加工变动性使得通过限制工站之间的缓冲区容量，又不显著降低产出地减少 WIP 成为可能。这突出说明了为什么削减变动性是 JIT 实施（implementation）的一个极为重要的部分。

8.8 变动性汇聚（variability pooling）

在这一章中，我们已经识别了若干项变动性的起因（失效、生产准备等等），并且观察了它们如何引起制造系统的拥塞。很清楚地，如我们将在第九章进行的更全面讨论，减轻这种拥塞的一种办法是通过寻找变动性的原因来削减变动性。不过还有另外一种更巧妙的办法，就是把多个变动性源头合并起来。这就是**变动性汇聚（variability pooling）**。它在制造业中已有许多应用。

变动性汇聚的一个日常应用的例子是财务规划。事实上所有的财务顾问都会建议投资于多样化的金融工具组合。（279|280）其原因，当然是规避风险。多种投资同时表现极差的状况不可能出现。同样，它们同时表现极好的状况也不可能出现。因此，我们期待多样化投资的收益变动低于单一资产。

在很多制造情景中，变动性汇聚都起着很重要的作用。在这里我们将讨论它如何影响成批处理、集结安全库存以及共享队列。

8.8.1 成批处理

为了阐明变动性汇聚背后的基本思想，我们来考虑一个问题：一个部件的加工时间和一批部件的加工时间，哪一个的波动较大？为了回答这个问题，我们必须定义变动性。本章中我们已经证明，变异系数是一个客户变动性的合理量度。因此我们就以变异系数 CV 来构建分析。

首先考虑一个部件。它的加工时间是一个均值为 t_0、标准差为 σ_0 的随机变量，则其加工时间的 CV 就等于

$$c_0 = \frac{\sigma_0}{t_0}$$

现在再来考虑 n 个部件组成的批次，每个部件加工时间的均值为 t_0、标准差为 σ_0。这样整个批次的加工时间均值为各部件加工时间均值的简单加和

$$t_0(batch) = nt_0$$

并且成批处理时的方差为各部件的方差之和

$$\sigma_0^2(batch) = n\sigma_0^2$$

因此，这个批次的加工时间 CV

$$c_0(batch) = \frac{c_0(batch)}{t_0(batch)} = \frac{\sqrt{n}\sigma_0}{nt_0} = \frac{\sigma_0}{\sqrt{n}t_0} = \frac{c_0}{\sqrt{n}}$$

因而，加工时间 CV 依批量的平方根递减（the CV of the time to process decreases by one over the square root of the batch size）。我们可以推断成批的加工时间比单件加工时间的波动要小（前提是所有的加工时间独立同分布）。其原因类似于投资组合。所有 n 个部件同时出现极长或极短的加工时间是极不可能的。因此成批趋向于“平均化”各个部件的变动性。

这是否意味着我们应该成批处理零件来削减变动性呢？不一定。在第九章我们将会看到，成批有另一个负面效应，它甚至会抵消较低变动性的所有好处。但是也有成批的变动性削减效应非常重要的时候，比如质量控制的抽样。在一批部件中抽取一定数量的样本降低了估计的不准确性，因而成为构建统计过程控制图的标准实践。

8.8.2 集结安全库存

变动性汇聚在库存管理中也非常重要。为了阐明原因，先来考虑这样一个例子。一个计算机设备制造商出售的系统的处理器、硬盘驱动器、CD ROM、可移动的媒体存储设备、RAM 以及键盘都各有三种配置可供选择。这一共就可以组合成 3^6 = 729 种不同的配置。为了使这个例子简单些，假设所有的组件的成本均为\$150，因此，制成品的成本为 6×\$150 = \$900。（280|281）进一步地，我们还假设每种类型计算机的需求均服从均值为每年 100 台的泊松分布，补给提前期均为三个月。

首先假设制造商存储有所有类型计算机的制成品，并且根据基础库存点模型设定存货水平。运用第二章中的技术，我们得出保持 99%的客户服务水平（供给率）需要 38 个单位的基础库存点水平，并且导致每种配置的平均库存水平为\$11 712.425。因此，总的库存投资为 729×\$11 712.425 = \$8 538 358。

现在假定制造商只存储组件并接单组装，而不再存储制成品计算机。我们假设从客户提前期的角度来看这是可行的，因为三个月的补给提前期可假定为取决于组件的可获得性。进一步地，相对于 729 种不同配置的计算机，因为只有 18 种不同的组件，存储物品的种类较少。然而，因为我们要组装这些组件，它们每种都必须有 0.99 ^ (1/6) = 0.9983 的供给率来确保 99%的客户服务水平。[13]假设每种组件要三个月的补给提前期，实现 0.9983 的供给率需要 6 306 件的基础库存点水平，并且使每种组件的平均库存水平为\$34 655.447.因此，现在总

[13] 注意如果组件价格不同，我们将希望能设定不同的供给率。为了降低总的库存成本，合适的做法是为便宜的组件设定较高的供给率，为贵重的组件设定较低的供给率。我们聚焦于汇聚带来的效率提升，所以忽视上述情况。第十七章提供了优化多部件库存系统存储规则的工具。

的库存投资是 18×$34 655.447 = $623 798，下降了 93%！

这种效应并不局限于基础库存点模型。它同样适用于使用（Q, r）或其他库存规则的系统。重点是保持通用的（*generic*）库存，从而可用于满足多个源头的需求。这使变动性汇聚的特性可以大幅削减所需的安全库存。我们将在第十章推式与拉式生产的背景下检视其他的接单组装型生产系统。

8.8.3 共享队列

我们在前面提到杂货店一般为各个收银台建立独自的队列，而银行一般为所有出纳建立一个队列。银行这样做是为了通过服务时间的变动性汇聚降低拥塞。如果一个出纳由于服务于一个坚持声称某个账户没有透支的顾客而陷入困境，队列继续移动到其他出纳。相反地，如果一个收银员由于等待顾客核定价格细目而不能接受新的顾客，这个队列的所有人都被堵住了（或者开始排到其他收银台队列后，这样是系统变现更像合并队列情形，但效率低、排队等待时间也不公平）。

工厂里，共享队列可用于削减 WIP 堆积于加工时间较长的机器之前的可能性。例如，在 8.6.6 节我们给出一个例子，及如果三台机器都有各自的队列时，周期时间为 7.67 小时；而如果三台机器共享一个队列时，周期时间只有 2.467 小时（降低了 67%）。

考虑另一个例子。假定一个五台机器构成的工站的到达速率为 13.5 件/小时（且$c_a = 1$）。每台机器名义上用 0.3 小时加工意见，自然 CV 为 0.5（即，$c_0^2 = 0.25$）。每台机器的平均失效间隔为 36 小时，修复时间假设服从指数分布且均值为 4 小时。由（8.6）式，我们能计算有效 SCV 为 2.65，因此$c_e = \sqrt{2.65} = 1.63$。（281|282）

应用 8.6.6 节的模型，我们可以为专有队列情形与合并队列情形分布建模。在专有队列情形，平均周期时间为 5.8 小时；而在合并队列情形，平均周期时间为 1.27 小时，下降了 78%（见问题 6）。在这里，巨大差异的原因很明显。合并的队列使加工任务避开长时间的失效。所有的机器同时失效不太可能，所以如果这些机器由一个共享队列供给，加工任务可以前往其他机器从而避开失效的机器。这是一个利用共享的机器削减加工变动性的有力途径。

然而，如果分散的队列实际上是不同类型的加工任务，将它们合并会给机器带来一项耗时的挑拣合适的加工任务的生产准备，这种状况就较复杂了。通过专有队列避免生产准备带来的产能节约，可能抵消通过合并队列带来的变动性削减。我们将在第九章中检视有变动性系统中生产准备与成批之间的权衡。

8.9 总结

从随机性的本质到产线中变动性的传递及其影响，这一章自始至终在讲变动性的复杂及微妙之处。从工厂物理学的角度来看，重要的观点有：

1. 变动性是生活的实情。事实上，物理学领域日益强调随机性或许是存在（existence）本身不可避免的一面。从管理的视角来看，很显然，有效地处理变动性与不确定性将是在可预见的未来的一项重要技能。

2. 制造系统中的变动性有多种来源。加工变动性可能产生于作业程序变更之类的简单

事件，也可能产生于生产准备、随机断供、质量问题等较复杂的效应。流动变动性产生于加工任务投入系统的方式或工站之间搬运的方式。因此，系统呈现的变动性，是工艺选择、系统设计、质量控制、管理决策等一系列因素的作用结果。

3. *变异系数是品目变动性的关键量度*。使用标准差除以均值得到的这个无量纲比率，我们能对加工变动性与流动变动性做出一致的比较。在工站的水平，有效加工时间的 CV 因为机器失效、生产准备、重工及多种其他因素而变大。对于恒定的可用率水平，与短期、高频的中断相比，长期、低频的中断趋向于更多地抬高 CV。

4. *变动性传递*。一个工站高度变动的输出会成为另一个工站高度变动的输入。利用率较低时，工站输出进程的流动变动性很大程度上取决于工站的到达进程变动性。然而，随着利用率的提高，流动变动性逐渐取决于工站自身加工时间的变动性。

5. *排队等待时间通常是周期时间的最大成分*。两个因素会导致较长的排队等待时间：高的利用率水平和高度变动性。本章中讨论的排队模型显示，提升有效产能（即，降低利用率水平）和降低变动性（即，缓解拥塞）都有利于压缩周期时间。

6. *限制缓冲区容量压缩了周期时间，却以降低产能为代价*。因为限制工站之间的缓存区容量逻辑上等效于设置（installing）看板，这条属性就是变动性削减（通过生产平滑、改进设施布置、物流控制、全面预防性维护与增强质量保证）在 JIT 系统中至关重要的主要原因。（283|284）它也指出在实现设定的产出与周期时间绩效时，产能、WIP 缓冲与变动性削减之间的相互替代关系。理解其中的权衡，是设计支持战略业务目标的运营系统的基础。

7. *变动性汇聚降低了变动性的影响*。通过集结，使其中之一不大可能支配绩效，变动性汇聚削弱的整体变动性的影响。这种效应在工厂物理学中有着多项应用。例如，保持低层次（a generic level）的存货并接单组装，以削减安全库存。还有，多机加工中心共享一个队列，以压缩周期时间。

下一章里，我们将应用这些见识，与已开发的概念和公式一道，来检视变动性如何降低了制造单位的绩效，并提出应对的方法。