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

第七章 工厂动力学基础

我不知道在别人看来,我是什么样的人;但在自己看来,我不过就像是一个在海滨玩耍的小孩,为不时找到比寻常更为光滑的一块卵石或比寻常更为美丽的一片贝壳而沾沾自喜,而对于展现在我面前的浩瀚的真理的海洋,却全然没有发现。

——伊萨克·牛顿

7.1 引言

上一章里，我们认为制造管理需要一门制造科学。在这一章，我们通过检视产线的一些基本行为，来充实这门科学。

为了引出即将重点讨论的种种量度（measure）与机制（mechanic），我们从一个实例开始。HAL 是一家生产印制电路板（PCB）的计算机公司，它的产品卖给其他企业，在上面安装组件，然后送入个人电脑的组装线。生产 PCB 的基本工艺如下：

1. 层压。铜和玻璃纤维层被压制在一起形成核心（空板）。
2. 机加工。核心被修整到一定尺寸。
3. 印制电路。通过感光和蚀刻过程，电路被印制在空板的铜层上，赋予“个性”（独特的产品特性），使其成为面板。
4. 光学检验与修补。电路被光学检验缺陷，如果缺陷不严重则修补。
5. 钻孔。在板上钻孔连接多层板的不同部分的电路。要注意多层板在印制完电路形成层后返回层压。单层板通过层压一次不需要钻孔和镀铜。
6. 镀铜。多层板在铜镀液过滤，在钻孔中沉淀铜，从而连接不同板上的线路。(213|214)
7. 加保护层。在面板上加层保护性的塑料膜。
8. 定尺寸。加工面板到最终尺寸。在大多数情况下，复合 PCBs 在一个单独的面板上加工制造，在定尺寸阶段形成单独的板。依据板的尺寸，很少有两个板从一个面板中制造出来的，或最多 20 个。
9. 线尾检测。对板的功能进行电器检测。

HAL 的工程师监控着 PCB 产线的产能与性能。他们对产能的最高估计总结在表 7.1，其中给出了每个工站的平均加工速率（每小时生产的板数）和平均加工时间（小时）。(要注意的是因为 PCB 通常成批加工，许多工艺由并行设备作业，所以加工速率不是时间的倒数。)这些都是平均值，用来说明 HAL 生产的不同型号与不同工艺路线的 PCB（如，一些板子也许要经过两次层压）。它们也可以说明“扰动”，如机器故障、换模时间、作业员效率等现象。这样，加工速率就估计出在无限投料的假定下每道工艺每小时可以产出多少板子。加工时间表示典型的板子在每道工序加工的平均时间，包括扰动引起的等待时间但不包括排队等待加工的时间。

表 7.1 *HAL PCB* 产线的产能数据

工艺	速率（件/小时）	时间（小时）
叠片	191.5	1.2
机加	186.2	5.9
布线	150.5	6.9
光学检测/维修	157.8	5.6
钻孔	185.9	10.0
铜板	136.4	1.5
覆膜	146.2	2.2
裁切	126.5	2.4
线尾检测	169.5	1.8

HAL 强调的主要绩效量度是产出（单位时间生产多少片 PCB）、周期时间（生产一块典型 PCB 所需的时间）、在制品（产线上的存货）以及客户服务水平（能准时送达客户的订单的比例）。在过去的几个月里，产出平均是每天 1,100 片，或是每小时 45.8 片（HAL 是 24 小时工作制）；产线上的在制品平均是 37,000 板；周期时间大概是 34 天，或者 816 小时；客户服务水平平均是 75.5%。

问题是，HAL 做得怎么样？

我们可以立刻回答这个问题的一部分。HAL 管理层并不满意 75%的客户服务水平，因为企业的目标是 90%，所以这方面的绩效不好。然而，产生这个问题的原因可能是热情过头的销售人员向客户承诺了不现实的交付期。它可能并不是产线出了问题的迹象。

其他绩效量度——产出、在制品、周期时间——则更难处理。我们需要建立某种基准用来比较。（214|215）一种方法就是以竞争对手的经营情况为标杆进行衡量。然而即使 HAL 能获取这些数据，它们实际上有多大的可比性仍然是个问题。毕竟，每个工厂都是独特的。比一个不同类型的工厂好或差并不能必然性地意味着什么。较好的基线会是将实际表现与该工厂的理论可能表现进行比较。

这一章里，我们审视简单、理想化产线的极端性能，并用产生的模型开发一个评价实际工厂的尺度。我们将返回 HAL 的例子，并用这个尺度去评价 PCB 产线的绩效。但首先我们必须定义术语。

7.2 定义与参数

科学方法绝对需要精确的术语。不幸的是，工业及 OM 文献中的制造术语完全没有标准化。这使得来自不同公司（甚至同一公司）的经理和工程师们很难交流及相互学习。这就意味着我们最好先认真地定义术语，并提醒读者注意在别的资料里可能会对同一术语有不同的定义，或者用不同的术语来代替我们的。

7.2.1 定义

在第二篇中我们聚焦于产线的特性，因为他们联系着单个流程与整个工厂。因此，以下的术语以使我们能精确描述产线的方式来定义。一些术语应用于工厂时可能有更广泛的意义，正如我们在定义中提到并在第三篇中一贯采用的那样。然而，为了培养对产线的敏锐直觉，我们将为第二篇的剩余章节保持这些狭义的定义。

工站（Workstation）：工站是执行（本质上）相同功能的一台或台设备或手工岗位的集

合。例如车削工站由数台立式车床组成，检验工站由数个配置了品质检验员的工作台组成，烧结工站包含为测试的目的而加热组件的一个单独的房间。在**工艺导向的布置（process-oriented layouts）**中，工站依据它们执行的操作（如，所有的磨床布置在磨床部门）而被物理性地组织起来。或者，在**产品导向的布置（production-oriented layouts）**中，工站被组织在生产不同产品的产线中（如，一台磨床被分配到一条产线）。站（station）、工作中心（workcenter）、加工中心（process center）都与工站同义。

工件（Part）：工件是原材料、组件、半成品（subassembly），或是在工厂工站上加工的成品。**原材料（Raw material）**是指从工厂外部购入的部件（如，棒料）。**组件（Components）**是装配成复杂一些制品的多个单件（如，齿轮传动装置）。**半成品（Subassembly）**是组成更复杂制品的装配单元（如，传动轴）。**组装品（Assemblies）**（或最终组装品）是完全装配的产品或终端品目（如，汽车）。注意一个厂的最终产品可能是另一个厂的原材料。例如，传动轴是传动轴工厂的最终产品，却是汽车组装厂的原材料或外购组件。

终端品目（End item）：不管是不是成品，直接卖给客户的部件称为终端品目。（215|216）终端品目与其组成部分（原材料、组件、半成品）的关系由物料清单（BOM）维护，这在第三章中有详细说明。

消耗品（Consumable）：在大多数情况下，消耗品是诸如钻头、化学试剂、气体、润滑油之类的在工站使用却不构成出售的产品的物料。更正式地，我们这样区分部件和消耗品，部件被列在物料清单上而消耗品没有。这就意味着，一些构成产品的品目，如焊料、胶水、电线，如果记录在物料清单上就是部件，如果不在则是消耗品。因为有典型适用于部件和消耗品的不同采购计划（如，部件可能根据 MRP 体系下定单，而消耗品由再定货点方法采购），这项选择可能影响这些品目是如何被管理的。

工艺路线（Routing）：工艺路线描述了部件经过工站的序列。它起始于原材料、组件、半成品存储点，终止于中间产品存储点或制成品库存。例如，齿轮的工艺路线开始于棒料的存储点，经过切削、滚齿、修毛刺，结束于成品齿轮的存储点。这个齿轮存储点可能给组成齿轮半成品的工艺路线提供原料。物料清单及相关的工艺路线包含了制造终端品目所需的基本信息。

订单（Order）：客户订单是客户对特定的料号、质量、交付时间的要求。客户发送的纸质或电子采购单可能包含若干客户订单。因此，我们把客户订单简称为订单。在工厂内部，订单也可能暗示着某些库存（如,安全库存）需要补充。对于客户发起的订单，定时（timing）可能更关键；但两种类型的订单都代表需求。

加工任务（Job）：加工任务是指有着相关逻辑信息（如，图纸、BOM）的、经过一条工艺路线的一组实体物料的集合。虽然每一个加工任务都是由实际的客户订单或对客户订单的预期（如，预测的需求）触发的，但加工任务与订单通常没有一一对应的关系。这是因为（1）加工任务以确切的部件（由料号唯一决定）衡量，而不是用来组成成品满足订单的部件集合；（2）加工任务中部件的数量依赖制造效率（如，对批量大小的考虑），因此可能与客户订购的数量不匹配。

产出（Throughput）：一个生产过程（机器、工站、产线、工厂）单位时间的平均产量（如，件/时）定义为该系统的产出。在企业水平，产出被定义为单位时间销售的产量。然而，产线主管一般控制生产而不是销售。因此，对于一个工厂、产线或者工站，我们定义产出为单位时间优良（无缺陷）部件（经理确实控制品质）的平均产量。如果产线由生产单一族类产品的串行工站构成，所有产品经过每个工站一次，则每个工站的产出将会是一样的（假定没有产出损失）。在复杂一些的工厂，工站服务于多条工艺路线（如，车间），一个工站的产出为经过它的所有流程的产出之和。

产能（Capacity）：一道生产工艺的产出的上限就是它的产能。在大多数情况下，在产

能或其之上水平投料生产会引起系统的不稳定（即，建立无限度在制品库存）。只有非常特殊的系统能在产能水平平稳生产。（216|217）因为这个概念难以把握却很重要，所以在引进合适的记号和概念之后，我们将在本章晚些时候对其进行更透彻的研究。

原材料库存（Raw material inventory，RMI）：如前所述，生产流程开始时的输入实体一般被称为原材料库存。它可以是要切削磨削形成齿轮的棒料，要层压形成电路板的铜片和玻璃纤维，要形成纸浆进而造纸的木片，或是要压成汽车保险杠的钢板等等。一般地，流程开始处的存储点称为原材料库存，即使物料已经过某些处理。

中间产品和制成品库存(“Crib” and finished goods inventory，FGI)：流程结束时的存储点是中间产品（即，中间的库存位置）或制成品库存。中间产品用于进一步加工或组装之前和厂内其他不同的部件进行连接。例如，生产齿轮组装件的工艺路线可能由若干包含如齿轮、外壳、曲轴等等的中间产品库存供给。制成品库存就是要运送客户之前的终端品目的存放位置。

在制品（Work in process，WIP）：工艺路线开始与结束之间的库存称为在制品库存。因为工艺路线开始并结束于存储点，在制品是不包括终点库存的所有产品。虽然口语中认为在制品包括中间产品，但我们区别了二者以使讨论更加明晰。

库存周转率（Inventory turns）：库存周转率是通常衡量库存利用效率的量度，定义为产出与平均库存的比值。一般地，产出以年为单位，因此这个比率反映了库存补给或周转的平均次数。具体应包含哪些库存取决于要测度的内容。例如，一个仓库中所有的库存是 FGI，则周转率为 TH/FGI。工厂中我们通常一并考虑在制品（仍在产线的库存）和制成品（等待销售的库存），所以周转率为 TH/（WIP+FGI）。无论如何，产出与库存的计量单位要保持一致。库存通常以成本价（即，不是销售价）计量，所以产出也应以成本价计量。

周期时间（Cycle time）：一条给定工艺路线的周期时间（也被多样地称为平均周期时间、加工时间、产出时间和暂留时间）是从加工任务投放于起点到它抵达终点的库存点的平均时间（即，部件作为在制品的时间）[1]。虽然这是周期时间的一种精确定义，但它仍是狭义的，只允许我们用在单一的工艺路线。通常人们所指的是由多个复杂的半成品组成（如，汽车）的产品的周期时间。这时用这个定义就不是很清楚了。汽车生产从什么时候开始？底盘何时下线？引擎何时生产？或者，以亨利·福特的方式，矿石何时从地下采出？我们将在稍后介绍如何计算此类组合件的周期时间，但现在将定义限制在单一工艺路线。

提前期（Lead time）、客户服务水平（Service level）、供给率（Fill rate）：一条给定工艺路线或产线的提前期是分配给该部件上的生产时间。正如此，它是一个管理常量[2]。相比而言，周期时间通常是随机的。（217|218）因此，在*接单生产*（即，以一定的交期生产部件满足订单）环境下的产线，一个很重要的产线绩效量度是客户服务水平，定义为：

$$\text{客户服务水平} = P\{\ \text{周期时间} \leqslant \text{提前期}\ \}$$

注意这个定义意味着对于给定分布的周期时间，客户服务水平可以由改变提前期来影响（如，提前期越长客户服务水平越高）。

在*备货生产*（即，补给一个使客户或其他产线预期可以无延迟地获得部件的缓冲）环境下的产线，另一个绩效量度可能比客户服务水平更合适。一个理性的选择就是我们在第二章中讨论过供给率，即由库存满足的订单所占的比例。因为供给率和许多其他绩效量度常常被称作“客户服务水平”，所以读者无论何时遇到这个术语，都需要注意寻找其精确的定义。

---

[1] 周期时间在装配线上可能还有另一个意思。因为时间被分配给每个工站来完成它的任务，它也可以指每个机器的加工时间（如，冲床的周期）。我们将避免该术语的其他用法来防止混淆。

[2] 回忆起 MRP 的时间计划功能就是严重依赖对这类提前期的选择。

我们在第二篇一直使用客户服务水平的前述定义，但将在第十七章回到供给率的量度方法。

利用率（Utilization）：工站的利用率是非停工待料时间所占的比例，包括工站加工部件时间，以及由于机器故障、换模与其他扰动带来的含料等待时间的比例。利用率的计算式为：

$$\text{利用率} = \text{到达速率} / \text{有效加工速率}$$

有效加工速率是考虑到整个计划期内都与之相关的的机器故障、换模及其他扰动的影响，工站加工部件所能达到的最大平均生产速率。

7.2.2 参数

参数是生产流程的数字化描述，因此在不同工厂取值不同。描述单个产线（工艺路线）的两个个关键参数是瓶颈速率和原始加工时间。下面我们定义这两个参数，并从中能计算出第三个参数：临界在制品水平。

瓶颈速率（Bottleneck rate: r_b）：产线的瓶颈速率是保持最高的长期利用率的工站的产出（单位时间的部件数或单位时间的加工任务数）。考虑到“长期”，我们就可以通过平均化把机器故障、作业员休息、质量问题等造成的中断排除出时间范围。这意味着，需要的对策，由于制定计划频率的不同而不同。例如，对于日计划，每天遇到的典型断供（outage）需要考虑在内；但意外的长期中断，如由重大混乱造成的，则不需要考虑。与之相反，对于年计划，由重大混乱造成的时间损失就要考虑在内了，如果这种情况在一年之中不是不可能发生的话。

无产出损失、每个工站只被访问一次的单路线产线中，加工任务到达每个工站的速率是一样的。因此，有着最高利用率的工站将是那些有着最低长期产能（即，最慢的有效加工速率）。有更复杂工艺路线或产出损失的产线中，瓶颈可能不是速率最慢的工站。较快的工站经历较高的到达速率是会有较高的利用率。因为这个原因，如我们在这里做的，应该以利用率定义瓶颈。

原始加工时间（Raw process time，T_0）： 产线的原始加工时间T_0，是产线中每个工站长期平均加工时间之和。（218|219）另外，我们可以定义原始加工时间为一个加工任务通过空的产线（即，不需要在其他加工任务后面等待）的平均时间。再次要注意的是，当决定平均加工时间应包括哪些内容时，我们必须考虑计划期的长度。在长期，原始加工时间应包括不经常出现的随机性的和计划的中断，而在短期则只需包括经常性的延迟。

临界在制品水平（Critical WIP，W_0）：产线的临界在制品水平，是指给定r_b和T_0，无变动性，以最短的周期时间（T_0）达到最大的产出（r_b）时的 WIP。临界在制品水平由瓶颈速率和原始加工时间来定义，关系式如下：

$$W_0 = r_b T_0$$

7.2.3 实例

我们现在用两个简单的例子来解释这些定义。

Penny Fab One. Penny Fab One 是一条简单的产线，生产专用于美国独立纪念日阅兵游

行的一美分硬币。产线由四台机器串联组成，工艺常见而稳定。第一台机器是冲床，切削出空白硬币；第二台给一面印上林肯头像，另一面印上纪念碑；第三台给硬币成边；第四台清理毛刺。每台机器准确地耗时两小时完成操作。（稍后我们将放松确定的加工时间这个要求。）每个硬币加工完后，立即被送到下个机床。产线每天 24 小时运行，休息、午饭等时间由额外的作业员顶班。为了我们的目标，硬币的市场需求可以假定为无限大，这样所有的产品都可以卖出；因此，对于这个系统，产出越多显然会越好。

因为这是一个没有产出损失的串联产线，瓶颈是最慢的工站。然而，每台机器的产能是一样的，都等于两小时一个硬币或是一小时半个。因此，这四台机器中的任意一台都可以被视为瓶颈，且

$$r_b = 0.5 \text{ 个/小时}$$

或是每天 12 个硬币。这样的产线被称为是**平衡的（balanced）**，因为所有的工站产能相同。

下一步，注意到原始加工时间是四个工站加工时间的简单相加，故

$$T_0 = 8 \text{ 小时}$$

临界在制品水平为

$$W_0 = r_b T_0 = 0.5 \times 8 = 4 \text{ 个}$$

我们将说明这的确是使产线 TH 达到 $r_b = 0.5$ 个/小时及 CT 达到 $T_0 = 8$ 小时的 WIP 水平。

注意到 W_0 等于产线中机器的数量。平衡的产线总是如此，因为每台机器执行一个加工任务刚好足够保证所有机器一直运转。然而，正如我们将看到的，不平衡的产线并非如此。

Penny Fab Two. 现在考虑一个复杂一些的 Penny Fab Two，它是一个有着多机工站的不平衡产线。它同样按四步生产硬币：冲压、印刻、成边、清理毛刺；但各工站现在有着不同数量的机器和加工时间，如表 7.2 所示。（219|220）

表 7.2 *Penny Fab Two:* 一条不平衡的产线

工站序号	机器台数	加工时间（小时）	工站能力（件/小时）
1	1	2	0.50
2	2	5	0.40
3	6	10	0.60
4	2	3	0.67

多机工站的出现使产能的计算变得有点复杂。对于单机，产能就是加工时间的倒数（如，如果执行一个加工任务需要半小时，则机器每小时可以完成两个加工任务）。由若干相同机器并联形成的工站，其产能必须用单机产能乘以机器台数来计算。例如，在 Penny Fab Two, 工站 3 中每台机器的产能为

$$\frac{1}{10} \text{ 个/小时}$$

所以工站 3 的产能为

$$6\times\frac{1}{10}=0.6\text{ 个/小时}$$

请注意工站的产能可以用机器台数除以加工时间来直接计算。各工站的计算见表 7.2。

有多机工站的产线的产能仍然是用瓶颈，也即产线最慢工站的产出来定义。在 Penny Fab Two，瓶颈是工站 2，所以

$$r_b=0.4\text{ 个/小时}$$

请注意瓶颈既不是有最慢机器的工站（工站 3），也不是有最少机器的工站（工站 1）。

原始加工时间仍然是加工时间的总和。注意，在一个工站增加机器并未降低T_0，因为一台机器一次只加工一个硬币。因此，Penny Fab Two 的原始加工时间为

$$T_0=20\text{ 小时}$$

不管产线是否有单机或多机工站，临界在制品水平总是定义为

$$W_0=r_bT_0=0.4\times20=8\text{ 个}$$

与 Penny Fab One 一样， Penny Fab Two 中W_0是个整数。当然，并不必然是这样。如果W_0出现小数，它就意味着没有恒定的 WIP 水平使产出精确地达到r_b、周期时间达到T_0。而且，注意到 Penny Fab Two 的临界在制品水平（8）比机器数（11）少。这是因为系统不平衡（即，工站产能不同），一些工站没有被充分利用。（220|221）

7.3 简单的关系

现在，在对制造科学的探索中，我们提出基本问题：产线中在制品、产出、周期时间三者之间的关系是什么？当然，答案要依据我们对产线的假设。在这一节，我们将给出对产线可能行为变化范围的精确（即，定量的）描述。这会使我们更直观地认识产线如何运行，并提供一个评价实际系统的尺度（标杆）。

在刻画在制品和产出等量度之间关系时，一个问题就是在实际系统中它们都是同时变化的。例如，在一个MRP系统中，产线上有可能这个月生产任务很多（因为一个紧张的主生产计划），而下个月任务很少。因此，在制品和产出量在第一个月都很高，而在第二个月都很低。为了表述清楚，我们通过控制产线在制品水平使其一直恒定来消除这个问题。例如，在Penny Fab One/Two里，我们将以确定数量的硬币（加工任务）来启动产线，然后在每一个制成品离开产线时投入一块新的原料[3]。

7.3.1 最佳情形绩效

为了分析和理解在最佳可能环境，也就是当加工时间绝对固定时产线的行为，我们将模拟 Penny Fab One。用一张纸和几个硬币就很容易实现，如图 7.1。

我们先来模拟产线中只允许有一个加工任务的系统。第一个硬币连续经过工站 1、2、3、4，各耗时两小时，总的周期时间为八小时。接着第二个硬币投入产线，重复同样的顺序。（221|222 因为这样导致每八小时生产一个硬币，所以产出是 1/8 件/小时。注意到周期时间

[3] 我们说这样的生产线是在 CONWIP（常量在制品）条件下运行，并将在第十、十四章进行更透彻的研究。

（$T_0 = 8$）等于原始加工时间，产出是瓶颈速率（$r_b = 0.5$）的 25%。

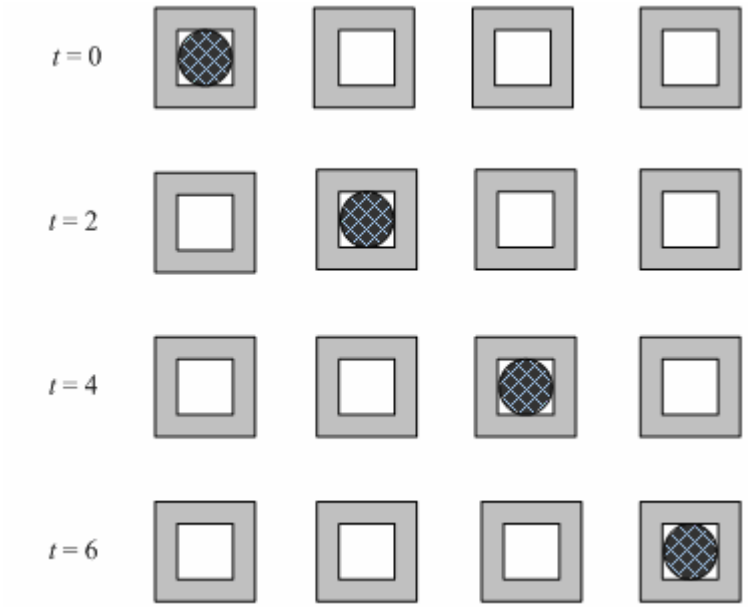

图 7.1 *Penny Fab One (WIP=1)*

现在我们给产线加第二个硬币（都从产线起点开始）。两小时后，第一个硬币在工站 1 加工完后在工站 2 开始加工。与此同时，第二个硬币到工站 1 加工。然后，第二个硬币跟着第一个，每两个小时换一个工站，如图 7.2 所示。在初始的等待后，第二个硬币没有再等待。因此，一旦系统稳定运行，每个投入产线的硬币的周期时间都正好是八小时。而且，因为每八小时生产两个硬币，产出增加到 2/8 件/小时，是 WIP = 1 时的两倍，达到产线产能（$r_b = 0.5$）的 50%。

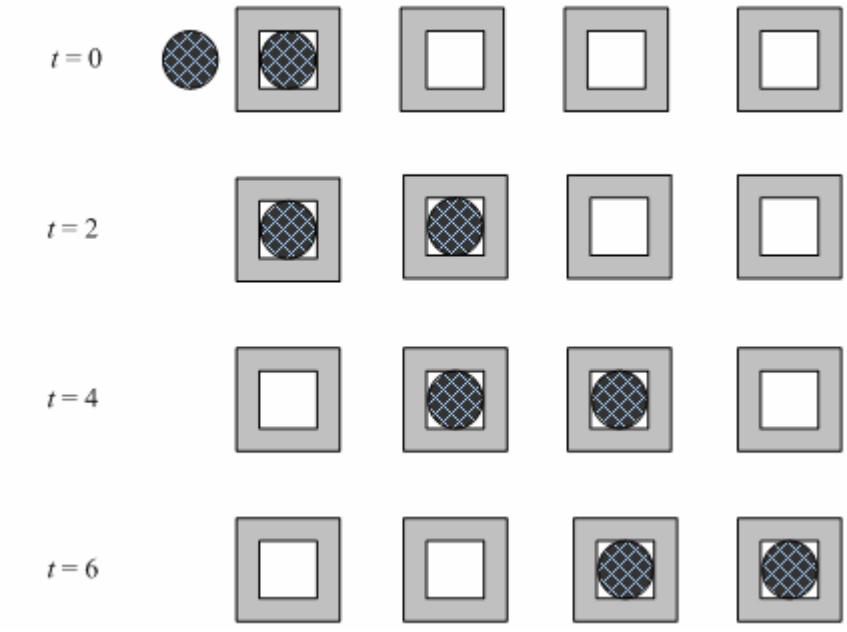

图 7.2 *Penny Fab One (WIP=2)*

我们增加第三个硬币。同样，在第一个工站处短暂的等待之后，就没有等待了，如图 7.3 所示。周期时间保持在 8 小时，而产出增加到 3/8 件/小时，是r_b的 75%。

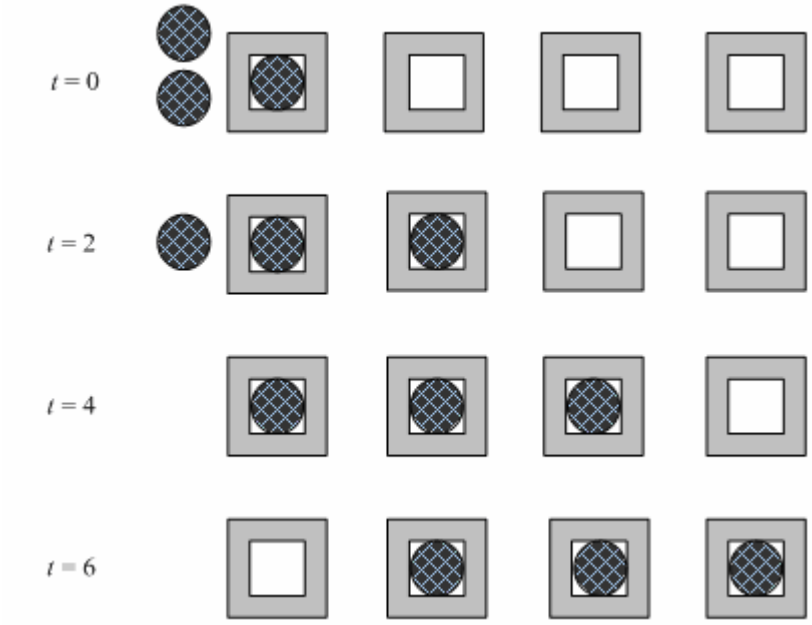

图 7.3 *Penny Fab One*（*WIP* = 3）

当我们增加第四个硬币，我们会看到一旦系统稳定，工站总是处于繁忙状态。因为没有在工站处的等待时间，周期时间仍是八小时。因为最后一个工站总是处于繁忙状态，它每隔一个小时生产出一个硬币，所以产出变成 1/2 件/小时，等于产线的产能。这个只有当 WIP 达到临界值时才出现周期时间达到T_0（最小值）产能达到r_b（最大值）的特殊情形，我们回想起 Penny Fab One，临界 WIP 水平

$$W_0 = r_b T_0 = 0.5 \times 8 = 4 \text{ 个}$$

现在我们增加第五个硬币到产线。因为只有四台机器，所以即使系统达到稳定运行状态，一个硬币仍将在第一个工站处等待。我们计算周期时间从加工任务投放（进入第一个工站队列的时刻）开始到离开产线，加上在第一个工站处额外的两小时等待，周期时间现在变成 10 小时。因此，周期时间第一次比其最小值$T_0 = 8$大。然而，因为所有工站一直处于运转状态，产出保持在$r_b = 0.5$件/小时。（222|223）

最后，考虑将 10 个硬币投入产线会发生什么。在稳定状态，六个硬币排列在工站 1 前面，意味着一个硬币从投入产线到开始在工站 1 加工的时间间隔为 12 小时。因此，周期时间是 20 小时。如前，所有机器保持运转，所以产出仍然是$r_b =$ 0.5 件/小时。此刻我们应该清楚地认识到，每增加一个硬币，周期时间增加两小时，而产出没有提高。

我们在表 7.3 中总结了无变动性时在不同 WIP 水平下 Penny Fab One 的表现，并在图 7.4 中显示出结果。从绩效的角度来看，显然是在只有四个在制品的情况下，Penny Fab One 运行得最好。只有这个 WIP 水平能使周期时间T_0最小、产出r_b最大——任何的减少都会损失

产出而不能减少周期时间；任何的增加都会增大周期时间而不能提高产出。这个特殊的 WIP 水平是我们先前定义的临界在制品水平（W_0）。

表 7.3 *Penny Fab One* 的 *WIP*、*CT* 和 *TH*

WIP	CT	% T_0	TH	% r_b
1	8	100	0.125	25
2	8	100	0.250	50
3	8	100	0.375	75
4	8	100	0.500	100
5	10	125	0.500	100
6	12	150	0.500	100
7	14	175	0.500	100
8	16	200	0.500	100
9	18	225	0.500	100
10	20	250	0.500	100

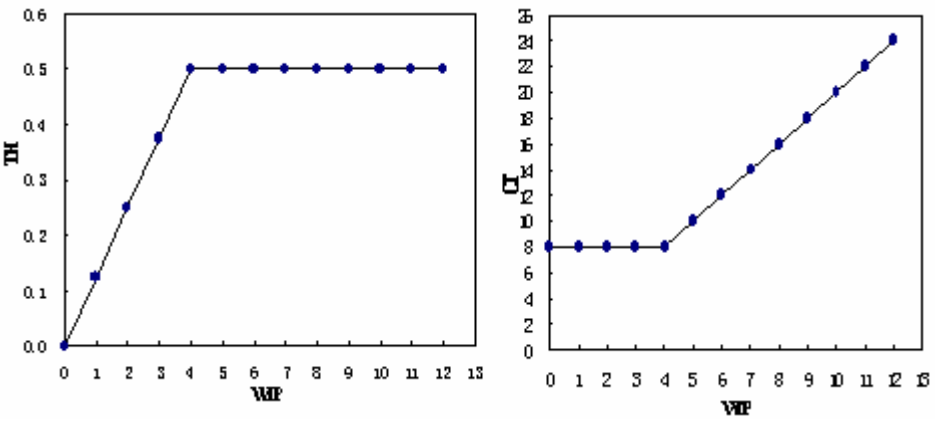

图 7.4 *Penny Fab One* 中 *CT-WIP*、*TH-WIP*

在这个特殊的例子中，临界在制品数等于机器台数。当产线由产能相等的工站组成（即，平衡的产线）时，情况都是这样。对于不平衡的产线，W_0 将小于机器数，但仍具有实现最大化 TH、最小化 CT 的 WIP 的属性，仍然定义为 $W_0 = r_b T_0$。

需要注意的是在无变动性的情况下临界在制品水平最优，在其他情况就不是最优了。的确，一旦有变动性，最优 WIP 水平就不好定义，因为在普遍意义上，增加 WIP 会同时提高产出（好）和增加周期时间（坏）。

里特定律（Little's Law）。 对表 7.3 的仔细坚实，揭示出 WIP、周期时间、产出三者之间一个有趣而根本的关系。在任意 WIP 水平，WIP 等于产出与周期时间之积。这个关系称作里特定律（以给出其数学证明的 John D. C. Little 的名字命名），并代表了《工厂物理学》的第一定律：

定律（里特定律）：（223|224）

$$WIP = TH \times CT$$

实际上，里特定律适用于所有产线，而不只是无变动性的。正如我们在第六章讨论的，里特定律不是定律而是同义反复（*Little's law is not a law at all but a tautology*）。对于特殊情形（如，无限长时期观察系统），这个关系可以通过数学方法证明。然而，除了一些特殊的，它并不能在有限长时期情形（当然包括所有的实际情形）下完全成立。尽管如此，我们会将它作为对制造系统本质的推测，在不确切时用作估计。

里特定律相当有用，因为它可以应用于单一工站，一条产线，或者整个工厂。只要三个量度的单位一致，上述关系将在长期内成立。这使它广泛地应用于实际情形。里特定律的一些直接应用包括：（224|225）

1. 队长计算。因为里特定律可用于单独的工站，我们用它来计算产线上每个工站的期望队长和利用率（运转时间的比例）。例如，考虑表 7.2 中概括的 Penny Fab Two，并假定它以瓶颈速率运行（0.4 个/小时）。依据里特定律，工站 1 的期望 WIP 为

$$WIP = TH \times CT = 0.4 \times 2 = 0.8 \text{ 个}$$

因为工站 1 只有一台机器，这就意味着它的时间利用率为 80%。同样，在工站 3，里特定律预测平均 WIP 为 4 个。因为有六台机器，平均利用率将是 4/6 = 66.7%。注意它等于瓶颈速率与工站 3 速率之比（即，0.4/0.6），正如我们预料的那样。

2. 周期时间削减。因为里特定律可以写成

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

很清楚，假定产出不变，削减周期时间就意味着减少 WIP。因此，长的队列预示着削减周期时间与 WIP 的机会。我们将在十七章中讨论削减的具体方法。

3. 周期时间的测度。直接测量周期时间有时候比较困难，因为它需要纪录系统中每一个部件的进入与离开时间。因为产出和 WIP 可用常规方法追踪，用 WIP/TH 作为周期时间的正确、合理的间接测度比较容易。

4. 计划的库存。在很多系统中，为了确保一个高的客户服务水平，加工任务被计划先于交付期完成。又由于在我们这个库存严苛（inventory consciousness）的时代，客户常常拒绝接受提前发送的货物，这种“安全提前期”使加工任务以制成品的形式等待而不是装货发送。如果**计划的库存（planned inventory）**时间为 *n* 天，那么根据里特定律，制成品库存数量为 *n*TH（这里 TH 的量度单位是件/天）。

5. 库存周转率。回想库存周转率等于产出与平均库存的比值。假设一个工厂的所有库存都是 WIP（即，产品直接从产线上运走，所以没有制成品库存），周转率为 TH/WIP，应用里特定律就是简单的 1/CT。如果加入制成品，则库存周转率为 TH /（WIP+FGI）。里特定律仍然适用，所以这个比率代表了加工任务通过产线及制成品库存的总时间的倒数。因此，直观上看，库存周转率是库存在系统中停留时间的倒数。

从某种意义上讲，里特定律是工厂物理学的“ $F = ma$ ”。它是三个变量之间的普适性公式。同时，它可以被看作是老生常谈。里特定律仅仅指出了一个明显的事实，即我们能以加工任务或时间测度工站、产线或系统中 WIP 水平。例如，一条产线每天生产 100 个曲轴箱、在制品水平为 500，则其中保有 5 天的 WIP。里特定律是对平均 WIP、周期时间、产出的单位有效转换的陈述，或者

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

或者

$$5\text{ 天} = \frac{500\text{个}}{100\text{个}/\text{天}} \qquad (225|226)$$

现在我们可以归纳表 7.3 和图 7.4 所示的结果，以实现我们给出“最优情形”（即，零变动性）产线 WIP 与产出之间关系的精确概括的初始目的。然后用里特定律扩展它从而刻画 WIP 与周期时间之间的关系。因为这些关系是源于无变动性的理想产线，以下表述给出有给定的 WIP 水平以及参数 r_b、T_0 的任何系统最大产出和最短周期时间。它就是我们的下一条工厂物理学定律。

定律（最佳情形绩效 ）：:给定 *WIP* 水平为 ω，最短周期时间为

$$\mathrm{CT}_{\mathrm{best}} = \begin{cases} T_0 & \text{若 } \omega \le \mathrm{W}_0 \\ \dfrac{\omega}{r_{\mathrm{b}}} & \text{其他} \end{cases}$$

给定 *WIP* 水平为 ω，最大产出为

$$\mathrm{TH}_{\mathrm{best}} = \begin{cases} \dfrac{\omega}{T_0} & \text{若 } \omega \le \mathrm{W}_0 \\ r_{\mathrm{b}} & \text{其他} \end{cases}$$

从中我们可以得出一个结论，与流行的口号相反的是，零库存不是一个现实的目标。即使在完美的确定性条件下，零库存造成零产出，因此也是零收入。更现实的理想 WIP 水平是 W_0。

Penny Fab One 代表了一种理想的（零变动性）情况，其中最优措施是维持与机器数量相等的 WIP 水平。当然，现实世界中，没有多少工厂以这么低的 WIP 水平运行。确实，很多产线中 WIP 与机器数量的比率接近 20:1（Bradt 1983）。如果将这个比率应用到 Penny Fab One，周期时间将接近 7 天，WIP 为 80 个。很明显，这比周期时间 8 小时在制品 4 个的情形（最优水平）坏多了。为什么实际工厂运行的情况与理想的临界 WIP 水平差得这么远呢？

不幸的是，里特定律几乎帮不上忙。因为 TH = WIP / CT，我们能用高的 WIP 水平和长的周期时间，或者低的 WIP 水平和短的周期时间得到同样的产出。问题是里特定律仅仅是三者之间的一个关系的表述。在给定一个量的情况下，如果要确定另外两个量（如，通过产出预测 WIP 和周期时间），则需要第二个关系。不幸的是，这三者之间没有通用的第二个关系。我们能做的，最好就是描述具体假定下产线的行为。除了以上的最佳情形，我们再来看两种其他状况，称之为最差情形和实际最差情形。

7.3.2 最差情形绩效(Worst-Case Performance)

考虑过产线的最佳可能行为后，我们再来看最差的。特别地，我们寻找有参数 r_b、T_0 的产线的最长周期时间和最小产出。这使得我们将行为分裂并评估实际产线的绩效。如果一条产线更接近最差情形而不是最佳情形，那么一定存在一些实际的问题（或者机会，取决于你

的想法)。(226|227)

为了便于讨论最差情形,我们假定产线中的加工任务一直保持不变。一个加工任务完成,另一个立即开始。实际中能达到这样的一种方法就是通过载具(*pallets*)在线上传输加工任务。加工任务一旦完成,立刻从它的载具上移除,载具立即回到线首去承接一个新的加工任务。WIP 水平,因而等于载具(固定的)数量。

现在想象你坐在一个载具上,在 WIP 等于临界值(如,有四个加工任务的 Penny Fab One)的最佳情形产线上。每次你到达一个工站,机器可以立即加工。正是因为没有等待(排队)才使产线达到最短的可能周期时间 T_0。

为了得到这个系统的最长的可能周期时间,我们必须在不增加平均加工时间(否则我们会改变 r_b 和 T_0)的情况下增加一些等待时间。我们所能达到的最坏等待时间就是,每次载具到达一个工站,都要在产线中其他所有加工任务后面等待。这种情况是怎样发生的呢?

考虑以下方式。假定在调整后的 Penny Fab One 中有四个载具,你在 4 号载具上。然而,对比于所有的加工任务在每个工站需要正好两小时,假定 1 号载具的加工任务需要八小时,而 2、3、4 号载具需要零小时。每个工站的平均加工时间为

$$\frac{8+0+0+0}{4} = 2 \text{ 小时}$$

像从前一样。因此,我们仍然有 r_b = 0.5 个/时, T_0 = 8 小时。然而,每次你的载具到达一个工站,你发现 1、2、3 号载具总是在你前方(见图 7.5)。 1 号载具的慢加工任务,导致其他加工任务在其后一直堆积。这就是所能引入的最长等待时间,因此它代表了最差情形。

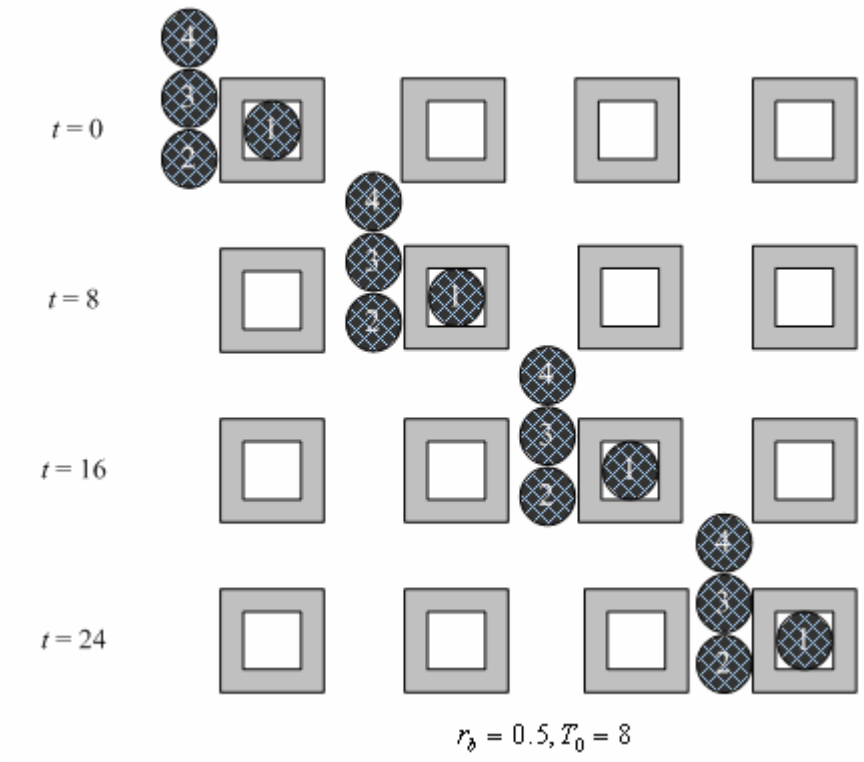

图 7.5 最差情形产线的演进

此系统的周期时间为：

$$8 + 8 + 8 + 8 = 32 \text{ 小时}$$

或者 $4T_0$。因为每当 1 号载具在工站 4 完成加工后，就会产出 4 件，因此产出为

$$4/32 = 1/8 \text{ 个/小时}$$

或者 $1/T_0$ 个/小时。注意到，产出与周期时间之积为 1/8×32 = 4，也就是 WIP 水平，正如里特定律所揭示的。

让我们总结一般产线的此类结果，作为我们的下一条工厂物理学定律。

定律（最差情形绩效）)：给定 *WIP* 水平 ω，最差情形的周期时间为

$$CT_{worst} = \omega T_0$$

给定 *WIP* 水平 ω，最差情形的产出为:

$$TH_{worst} = \frac{1}{T_0}$$

我们注意到一件有趣的事情，系统最佳情形与最差情形中没有随机性。在最差情形下有变动性，因为加工任务有不同的作业时间；但没有随机性，因为所有加工时间都完全可以可预知。（227|228）质量管理的文献强调减低变动性的必要性，但有时却暗示变动性和随机性是同义的。以上的工厂物理学结果却显示并非如此；变动性可能是随机性或不良控制（*bad control*）（或两者兼有）的结果。在第八、九章开发处理变动性的工具之后，我们将更深入地检视这个区别。

最后，读者可能会无可非议地怀疑最差情形的真实性。毕竟在这个例子里，我们使加工时间尽可能多变来达到最长的等待时间（为使周期时间尽可能地长）。为了这样做，我们假定其中一个载具有长的加工时间，而其他载具有零加工时间。当然这在现实中永远不会发生。

但它能，并且（至少在一定程度）的确发生。为了理解其原因，假定 Penny Fab One（当 WIP = 4 时）中承载加工任务的四个载具通过叉车在工站之间自主移动。进一步地，假定因为叉车有其他的职责，不能提供单独运输每个载具所需的次数。相反地，它等一个工站的四件都生产好了再一起集中移到下个工站。类似地，它等到四个载具在产线尾都空了再把它们取回到线首承载新的加工任务。假定每个加工任务在每个工站的的作业时间为两小时（如，原始的 Penny Fab One），并且叉车的移动时间足够短而可以合理地按零处理，系统的进程将与图 7.5 所示的完全一样。因此，最差情形可能产生于**成批移动（batch moves）**。

当然，很少见有如此极端的成批移动、每个加工一起搬运的工厂。（228|229）更常见的是，产线上的在制品以可能在规模上有差异的若干批次搬运。尽管这种更加适度的批次不会产生最差情形行为，它是能将产线推向最差情形而不是最佳情形的一个因素。因此，在许多生产系统中，成批（batching）是一个真正的问题（机会）。

7.3.3 实际最差情形(Practical Worst-Case Performance)

现实世界的产线没有一个是不折不扣地按最佳情形或最差情形运行的。因此，为了更好地理解这两个极端情形之间的行为，考虑中间的状况也是有意义的。我们通过一个不像前面两个的、包含随机性的情形来研究。实际上，从某种意义上说，它代表了最大随机性情形。我们用**实际最差情形(practical worst case)**这个术语来表达我们的信念，它是实际中表现更差的系统改进的目标。

为了描述实际最差情形以及表明为什么它能作为最大随机性情形，我们必须先定义系统状态的概念。系统状态是所有工站的加工任务的完整描述：有多少加工任务，它们已被加工了多长时间。在我们这里假定并在以下描述的具体条件下，所需要的信息仅仅是每个工站的加工任务的数量。因此，我们可以用一个维数与工站数相等的向量给出状态的精确总结。

例如，在一个四工站、三加工任务的产线上，向量（3，0，0，0）代表三个加工任务都在第一个工站的状态，而向量（1，1，1，0）代表工站 1、2、3 各有一个加工任务的状态。对于由四工站、三加工任务构成的系统，有 20 种可能的状态，如表 7.4 所示。

表 7.4　四工站、三加工任务的系统的可能状态。

状态	向量	状态	向量
1	（3，0，0，0）	11	（1，0，2，0）
2	（0，3，0，0）	12	（0，1，2，1）
3	（0，0，3，0）	13	（0，0，2，1）
4	（0，0，0，3）	14	（1，0，0，2）
5	（2，1，0，0）	15	（0，1，0，2）
6	（2，0，1，0）	16	（0，0，1，2）
7	（2，0，0，1）	17	（1，1，1，0）
8	（1，2，0，0）	18	（1，1，0，1）
9	（0，2，1，0）	19	（1，0，1，1）
10	（0，2，0，1）	20	（0，1，1，1）

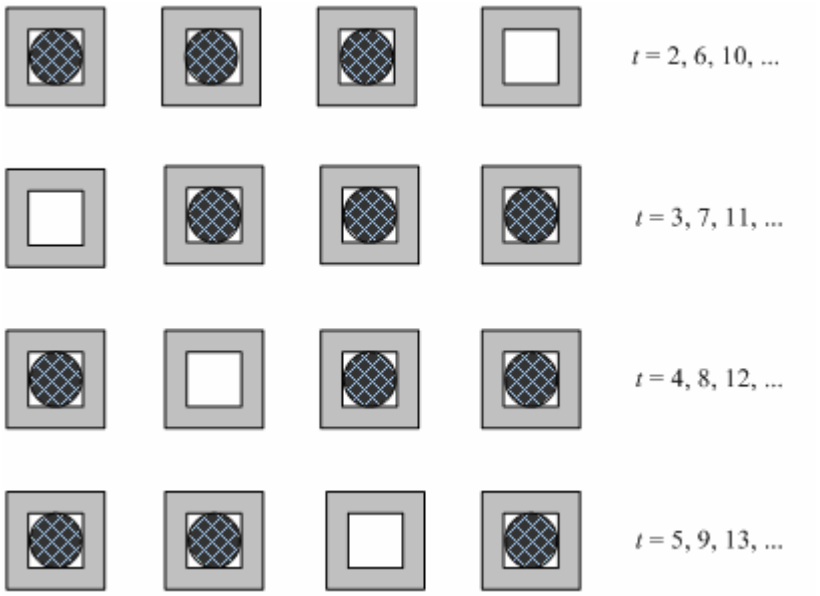

图 7.6　四工站、三加工任务的产线按最佳情形运行的系统状态

根据对产线所作的特定假设，不是所有状态都会发生。例如，假如在四工站、三加工任务的系统中所有的加工时间为一小时，并依照最佳情形运行，则只有四种不同的的状态——（1，1，1，0）、（0，1，1，1）、（1，0，1，1）和（1，1，0，1）不断重复发生，如图 7.6 所示。同样地，如果依照最差情形运行，四种不同的状态——（3，0，0，0）、（0，3，0，0）、（0，0，3，0）和（0，0，0，3）将反复发生，如图 7.7 所示。因为这两种系统没有随机性，

其他的状态永远不会出现。（229|230）

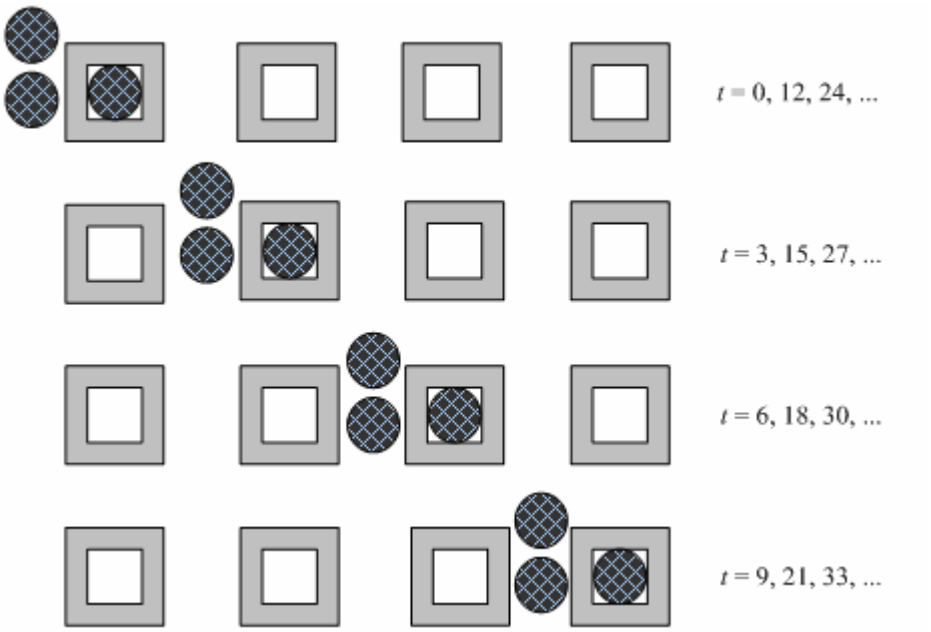

图 7.7　四工站、三加工任务的产线按最差情形运行的系统状态

当随机性被引入产线时，更多的状态成为可能。例如，假定加工时间是确定的，但每隔一段时间一台机器会坏掉（break down）几小时。大多数时间我们会观察到如图 7.6 所示的“摊开”（spread out）状态，但偶尔也会看到如图 7.7 所示的“集群”（clumped up）状态。如果只有一点随机性（如，机器故障非常少），“摊开”状态的频率会很高，然而，如果有大量随机性（如，机器不是这坏就是那坏），表 7.4 所示的所有状态会经常出现。因此，我们定义最大随机性为引起每种可能状态等频率发生的情况。（230|231）

为了使所有状态均等地发生，需要三个特别的条件：

1. 产线必须是平衡的（即，所有工站必须有相同的平均加工时间）。

2. 所有的工站必须由单机构成。（这个假设使我们避免了并联机器作业以及加工任务互超（jobs passing one another）的麻烦。）

3. 加工时间必须是随机的，并且其发生服从**指数分布（exponential distribution）**。指数分布是唯一一个具有**无记忆性（memoryless）**这种特殊性质的连续分布（见附录 2A）。这意味着，如果机器的加工时间服从指数分布，部件已被加工多久不能提供它将于何时完成的信息。例如，如果机器的加工时间服从均值为一小时的指数分布，并且当前的加工任务已进行了五秒，则期望的剩余加工时间为一小时；如果加工任务已进行了一小时，则期望的剩余加工时间为一小时；如果加工任务已惊醒了 942 小时，期望的剩余加工时间还是一小时[4]。当预测未来时，机器好像忘了它以前的工作——因此有术语无记忆性。因此，如果加工时间服从指数分布，完整地定义系统状态时没有必要知道加工任务已进行了多久。

为了理解实际最差情形怎样运行，回到那个想象自己驾着载具在产线上往复循环的思想实验。假设产线中有 N 个每个平均加工时间为 t 的（单机）工站，在制品保持在 ω 个加工

[4] 虽然想象加工时间这样表现可能只是一个延伸，但在日常生活中肯定有此类行为的例子。例如，延误的航班离开之前的时间，某条铁路上火车到站之前的时间，某些承包人完成家居改进任务的时间，等等。

任务的常量水平。则这条产线的原始加工时间 $T_0 = Nt$，瓶颈速率 $r_b = 1/t$。

因为以上三个条件保证所有状态等可能发生，那么从你在载具上的视角来看，每次你到达一个工站，会期望看到平均 ω-1 个其它加工任务等可能地分布于 N 个工站。所以在你之前到达的加工任务的期望值为（ω -1）/N。因为你在工站中花的平均时间为其他加工任务完成的时间与你的加工任务的加工时间之和，可以写成

在一个工站的平均加工时间 = 进行其他加工任务的时间 + 进行你的加工任务的时间

$$= \frac{\omega - 1}{N} t + t$$

$$= (1 + \frac{\omega - 1}{N}) t$$

通过假设在你之前的（ω - 1）/N 件加工任务所需平均完成时间为 $\frac{\omega - 1}{N} t$，我们忽视你到达时处理中的加工任务只是部分完成的情况。正是指数分布的无记忆性使我们可以这样做。

最后，因为所有工站假设为相同的（*identical*），我们可以简单地用每个工站的平均时间乘以工站数 N 来计算平均周期时间，得到（231|232）

$$CT = N(1 + \frac{\omega - 1}{N}) t$$

$$= Nt + (\omega - 1) t$$

$$= T_0 + \frac{\omega - 1}{r_b}$$

为了得到相应的产出，我们简单地应用里特定律：

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

$$= \frac{\omega}{T_0 + (\omega - 1)/r_b}$$

$$= \frac{\omega}{W_0 + (\omega - 1)/r_b}$$

$$= \frac{\omega}{W_0 + \omega - 1} r_b$$

它引出我们对实际最差情形绩效的定义。

定义（实际最差情形绩效）：给定 *WIP* 水平 ω，实际最差情形（*PWC*）的周期时间为

$$CT_{PWC} = T_0 + \frac{\omega - 1}{r_b}$$

给定 *WIP* 水平 ω，实际最差情形（*PWC*）的产出为

$$TH_{PWC} = \frac{\omega}{W_0 + \omega - 1} r_b$$

注意到这种情形的行为对极低或极高的 WIP 水平都是适用的。一个极端情况下，系统中只有一个加工任务（$\omega = 1$），正如我们预计的周期时间变成原始加工时间 T_0；另一个极端情况下，WIP 水平非常高（即，$\omega \to \infty$），产出接近产能 r_b，而周期时间无限制地增长。后者的结果揭示了在高变动性系统中要使产出接近产能，则需要高的 WIP 水平，来确保机器的高利用率。但这也造成了大量的等待并因而有很长的周期时间。

实际最差情形的产出、周期时间总是在最佳情形和最差情形之间。因此，PWC 提供了一个估计许多真实系统行为的有用的中间点。通过收集一条真实产线的平均 WIP、产出、周期时间（实际上，由于有里特定律，三者之中任意两个就够了）的数据，我们可以判断它在最佳情形（BC）与实际最差情形（PWC）之间的区域内，还是在实际最差情形（PWC）与最差情形（WC）之间的区域内。绩效好于 PWC（即，对于一个给定的 WIP 水平，有更大的产出和更小的周期时间）的系统是“优”的，差于 PWC 的系统是“劣”的。我们有理由将改善的努力聚焦于劣的产线，因为它们有改善的空间。因此，我们的三种情形提供了一种**内部标杆比较（internal benchmarking）**的方法论（即，相对于与外部系统进行对比的**外部标杆比较（external benchmarking）**）。

为了获得关于如何改进劣的产线的进一步指导，我们可以看看推导 PWC 所用到的三个假设：（232|233）

1. 平衡的产线；
2. 单机工站；
3. 指数分布（无记忆性）的加工时间。

因为这三个条件被用来最大化产线中的随机性，所以改进任何一个都会趋于改进系统绩效。

首先，我们可以通过在一个工站增加产能来使产线变得不平衡。可以通过增加实体设备，减少由工人休息或机器故障造成的停机时间，或以更有效率的工作方法加速加工等来实现。显然，如果我们增加所有工站的产能，产出将会增加。但即使我们只在一些工站增加产能以致 r_b 没有变化，还是会减少随机性（即，表 7.4 的状态不再等可能发生）并因此引起 TH-WIP 曲线更迅速地上升（即，系统中较少的 WIP 达到同样的产出）。我们意识到这样使产线不平衡与传统工业工程强调的使产线平衡相反。然而，正如我们将在第十八章中看到的，线平衡主要应用在同步装配线，而不是由像我们在这里考虑的独立工站构成的产线。

其次，我们可以在工站中利用并行机代替单机。如果是通过增加额外的机器来实现，那么就可以增加产能并因此有着与上文的讨论本质上相同的重要效果。但甚至用同样产能的并联机器代替单机在某些情况下也能提高绩效。例如，重新考察 Penny Fab One，假定加工时间不是定量而是服从指数分布，每个工站的平均加工时间仍为两小时。假定工站 3 和 4 合并为一个有两台并联机器的工站，每台机器一步就可以成边和去毛刺，加工时间为原来的两倍（即，每个硬币平均用四小时）。因为工站的产能是 0.5 个/小时，所以产线的瓶颈速率仍是 r_b。

原始加工时间仍是T_0 = 8 小时。但在原先的安排中，两个硬币要分别成边和去毛刺，结果就是一个得等待。在修改的产线中，任何时候都有两个硬币在成边或去毛刺，我们保证了两个在同时作业。结果将是，对于给定的 WIP 水平，使用并联机器改进的产线中等待更少，因此有更短的周期时间。

最后，我们减少加工时间的变动性使之低于指数分布所指示的。减少加工任务在工站后堆积的可能性，因此也减少等待，将改善给定 WIP 水平下的产出和周期时间。我们将在第八章讨论与指数分布相关的变化减少的意义，在第三篇讨论达到它的实际方法。

图 7.8 和 7.9，在 Penny Fab Two 所有工站加工时间指数分布的假定下，将周期时间和产出作为 WIP 的函数，解释了其中一些概念。为了比较，我们描出了在同样瓶颈速率和原始加工时间（即，r_b = 0.4，T_0 = 20）下的最佳、最差和实际最差情形的曲线。即使加工时间服从指数分布，因为 Penny Fab Two 有不平衡的产线及并联机器工站，它的绩效优于实际最差情形。如果我们降低减少加工时间的变动性，绩效将会更好。

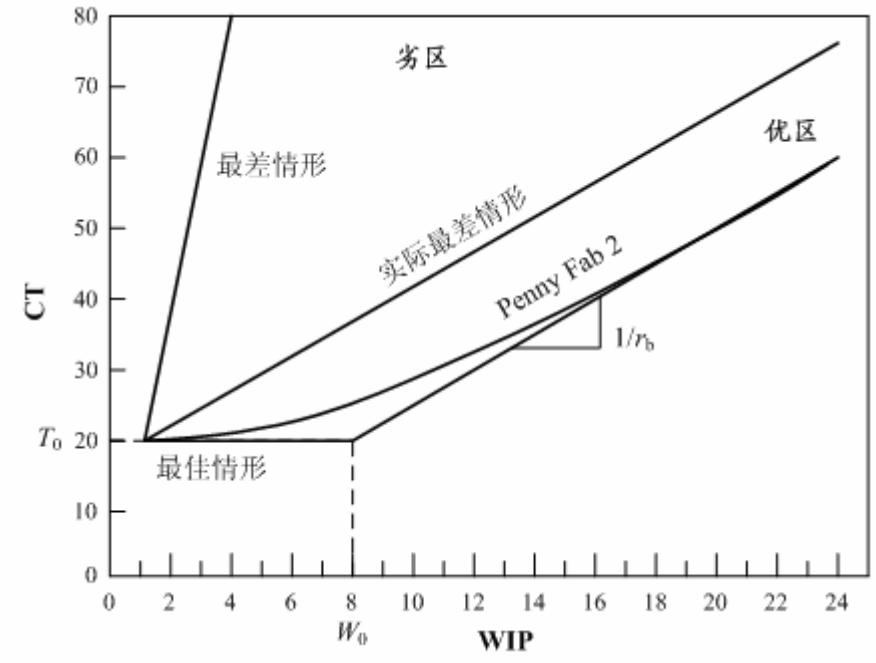

图 7.8 *Penny Fab Two* 中的 *CT-WIP* 关系

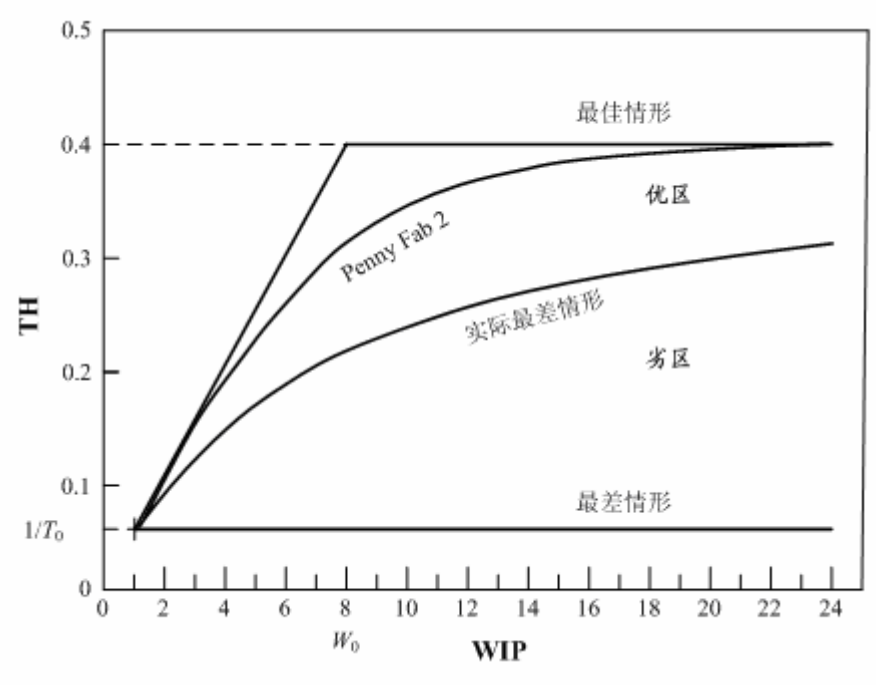

图 7.9 *Penny Fab Two* 中的 *TH-WIP* 关系

7.3.4 瓶颈速率和周期时间

自从 20 世纪 80 年代，许多注意力集中于加工任务系统瓶颈的重要性（参见，如 Goldratt 与 Cox 1984）。（233|234）我们这里的讨论当然赞同瓶颈速率 r_b 是重要的，因为它设定了产线的产能。但工厂物理学定律让我们更深刻地理解瓶颈的作用，而不仅仅是这个明显的结论。

首先，如果我们在运作一条“优”的产线（即，对于任意 WIP 水平，产出大于实际最差情形），则在典型 WIP 水平下（如，5 倍 W_0 到 10 倍 W_0 之间）周期时间将会非常接近 ω / r_b，其中 ω 是 WIP 水平（这可由图 7.8 和 7.9 观察得出）。因此，对于任意给定的 WIP 水平，提高瓶颈速率 r_b 会降低周期时间。

不幸地是，有时由于物理或经济的原因，提高瓶颈速率是不现实的。例如，假设镀铜设备是我们在本章开始时描述的 HAL 工厂的瓶颈。机器的速率由化学工艺控制。因此，如果它已经以每天最大的小时数运行（即，没有需要解决的人员及维护问题来增大有效产能），则增大产能的唯一方法是增加另一台设备。这是一个很可能过分的成本极高的选择，因为它会导致产能增加 100%。在类似状况下，考虑提升非瓶颈资源的产能从经济上来看是划算的。

为了理解这一点，考虑一个有四个单机工站的系统。前三个工站用 10 分钟完成一个加工任务，最后一个（瓶颈）用 15 分钟。因此，瓶颈速率是 4 件/小时。（234|235）

现在，假设我们把瓶颈加速到 10 分钟/件（6 件/小时），从而平衡了产线。图 7.10 说明了对于产线 TH-WIP 曲线的影响。注意到改进的产线有更高的加工速率上限（一个新的 r_b），然而 TH 曲线与它的距离也更远了。原因平衡的产线比不平衡的产线更频繁地趋于饥饿其瓶颈，因此需要更多的 WIP 使产出达到产能。然而，加速瓶颈能在任何 WIP 水平下提高产出。

或者，假设我们加速所有非瓶颈制程，使它们的加工时间仅需要五分钟，但瓶颈仍保持在 15 分钟。图 7.11 这样也会在任何 WIP 水平下提高产出。事实上，较低的 WIP 水平下，

这样比加速瓶颈更多地提高了产出。然而，对于较高的 WIP 水平（六个及以上），加速瓶颈比加速非瓶颈更多地提高了产出。同样要注意到我们对于非瓶颈工站做的改变要大于瓶颈工站（即，我们将三台机器加工时间减半，而只削减另一台机器的 33%）。如果我们可以自由将任意一处的加工时间削减五分钟，则最佳位置是瓶颈，总是这样。但这样未必总是可行的（经济性），所以我们应该知道改进非瓶颈资源也可以带来绩效提升。（235|236）

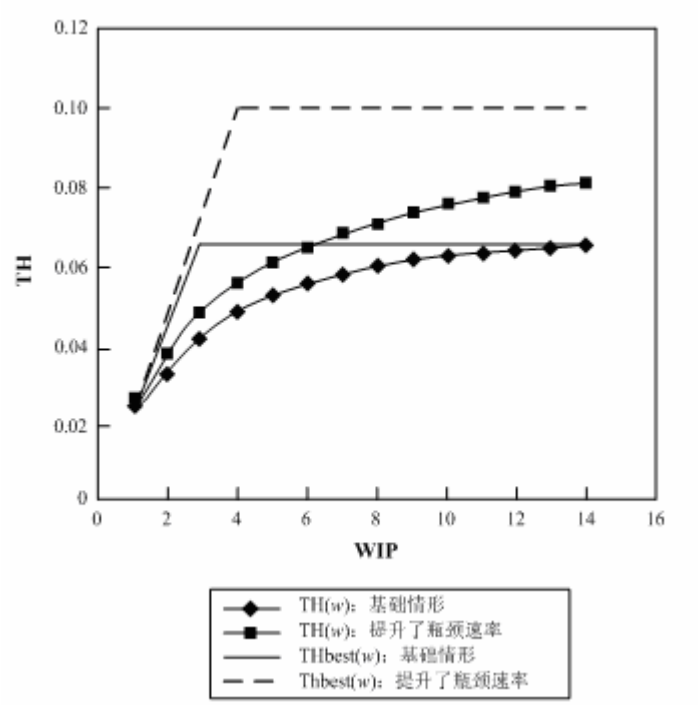

图 7.10 提升瓶颈速率对产出曲线的影响

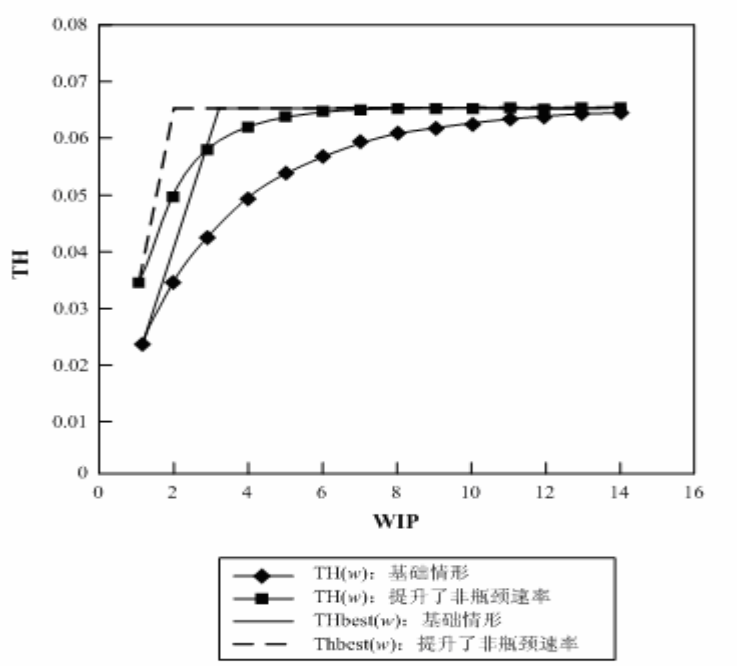

图 7.11 提升非瓶颈速率对产出曲线的影响

7.3.5 内部标杆比较

我们现在有工具可以重新考虑本章开始时的 HAL 案例。通过实际绩效与最佳、最差、实际最差情形的比较，我们可以评价 PCB 产线。为了这样做，我们必须估计瓶颈速率 r_b 和原始加工时间 T_0。如果忽视多次进出一些工站（如，层压）的再次观察，因为这已在速率和时间数据里考虑了，瓶颈就是有着最小产能的工艺。以 r_b=126.5 件/小时来衡量。原始加工时间就是表 7.1 的加工时间之和，T_0=33.1 小时。因此产线的临界 WIP 水平为

$$W_0 = r_b \times T_0 = 126.5 \times 33.1 = 4187 \text{ 件}$$

回忆起实际产出为 45.8 件/时，实际周期时间为 816 小时，实际 WIP 水平为 37000 件，我们立即可以得出一些结论。首先我们用里特定律快速得核对一下数据：

$$TH \times CT = 1100 \times 34 = 37400 \approx 37000 \text{ 件}$$

因为对于里特定律的精确适用只针对长期平均情形，我们并不期待它在此完全成立。然而，上述计算正在数据的精确程度之内，因此没有问题。

其次，通过注意到产出是瓶颈速率的 45.8/126.5 = 36%，周期时间是原始加工时间的 816/33.1 = 24.6 倍，WIP 是临界 WIP 的 37000/4187 = 8.8 倍，我们将这些实际量度与上下文连贯起来。这些看起来都不是很好。然而，我们必须慎重从单一的量度得出结论。例如，仅仅知道 WIP 是临界 WIP 的 8.8 倍，这本身并不意味着产线表现不好。甚至一条非常好的产线都需要一个高的在制品水平来达到接近瓶颈速率的产出水平。但当 WIP 水平很高而产出很低时，这就是差的标志了。到底有多差可以通过与实际最差情形比较而确定。

我们有两种方法可以比较实际绩效与 PWC。一种就是计算与 HAL 产线 r_b、T_0、WIP 相同的 PWC 产线的产出水平，并与实际产出比较。（236|237）用 PWC 定义的公式，可得

$$TH_{PWC} = \frac{\omega}{W_0 + \omega - 1} r_b = \frac{37400}{4187 + 37400 - 1}(126.5) = 113.8 \text{ 件/小时}$$

实际产出 45.8 件/小时比这个水平的一半还少，显示其绩效比实际最差情形差多了。

或者，我们可以计算与 HAL 产线 r_b、T_0 相同的 PWC 产线为了达到 TH 的观测水平而需要的 WIP 水平，即

$$TH_{PWC} = \frac{\omega}{W_0 + \omega - 1} r_b = 45.8 = 0.36 r_b$$

得出

$$\frac{\omega}{W_0 + \omega - 1} = 0.36$$

或

$$\omega = \frac{0.36}{0.64}(W_0 - 1) = 2354 \text{ 件}$$

实际 WIP 比这个水平的 15 倍还多，再一次显示 HAL 产线在把 WIP 转化为产出方面远不如 PWC 有效率。

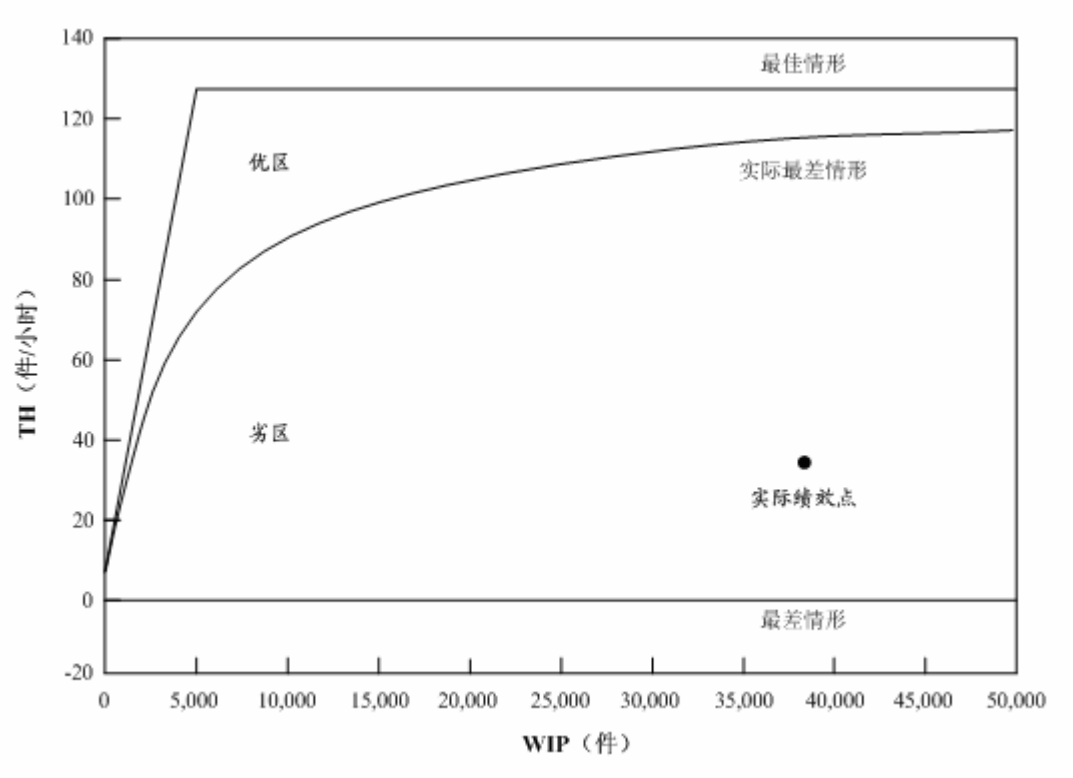

图 7.12 *HAL* 例子中的 *TH-WIP* 曲线

通过描出最佳、最差、实际最差 TH-WIP 曲线，并描出实际绩效点，我们可以把这些计算结果图示出来。这就是图 7.12。从中我们可以戏剧性地看出 WIP 与 TH 组合（37400，45.8）正好落在最差与实际最差之间的劣区。很明显，有这样表现的产线相比于落在最佳与实际最差情形之间的“优”区的产线有更多改进的机会。

这个例子显示本章所讲的模型可以帮助诊断一条产线，并判断它是否有效率地运行。但它们并没有揭示为什么一条产线运行得差，因此并不能帮我们确定怎样去改进。为此，我们需要更深入地研究是什么使一些产线在 WIP 转化为产出方面效率很高而其他的产线效率却很低。（237|238）这将是接下来的两章的主题。

7.4 人力约束的系统

这一章里，我们聚焦于机器是主要约束的产线。我们曾含蓄地假设，如果有操作员，并且他们被指派到机器，则因此可以将他们视为工站的一部分。然而，在某些系统中，工人要执行多个任务或照看多个工站。这种类型的系统比我们目前考虑的简单产线表现出更复杂的行为，因为作业流受机器、操作员的数量和特性的共同影响。

虽然柔性人力这个主题太宽泛了而很难无所不包地描述，但我们可以考察一下有人力约束的产线如何与先前讨论的简单产线相联系。我们考虑以下三种状况。

7.4.1 充足产能情形

我们的讨论始于人力是 TH 唯一约束的情形。也就是，我们假定每个工站都有足够的机器设备来保证工人永远不会因为缺乏设备而阻塞。人们可能会认为这种状况永远不会在实际中发生，但是确实存在接近这种行为的系统。作者遇到的一个例子就是印制图像的生产设施。公司从客户那里接收内容（文本、照片等）并通过一系列步骤（如，扫描、颜色校正、纸产品完成）转化为电子制版数据，然后再送入打印机制成纸质产品。大部分印制步骤需要计算机及外围设备。因为与交货滞后的损失（delay cost）相比计算机设备的价钱不贵，公司在每个工站安装了足够多的设备来确保技师执行多种任务时几乎不需要等待设备空闲。结果就是机器比人多得多，这意味着人是系统的关键约束。

图像公司在工站中配备充足产能的一个主要原因就是便于它的柔性人力政策。相对于为每道操作配备专职人员，公司交叉培训员工，使得几乎每个员工都可以执行所有操作。这样就使得公司指派工人到加工任务而不是工站。工人可以在合适的工站执行操作，并跟随加工任务通过整个系统，如图 7.13 所示。（238|239）额外的计算机使得工人不必在工站处等待设备空闲。让工人一直跟随着一个加工任务通过系统，意味着客户只需联系一个人，并使得这个人清楚地对质量负责。

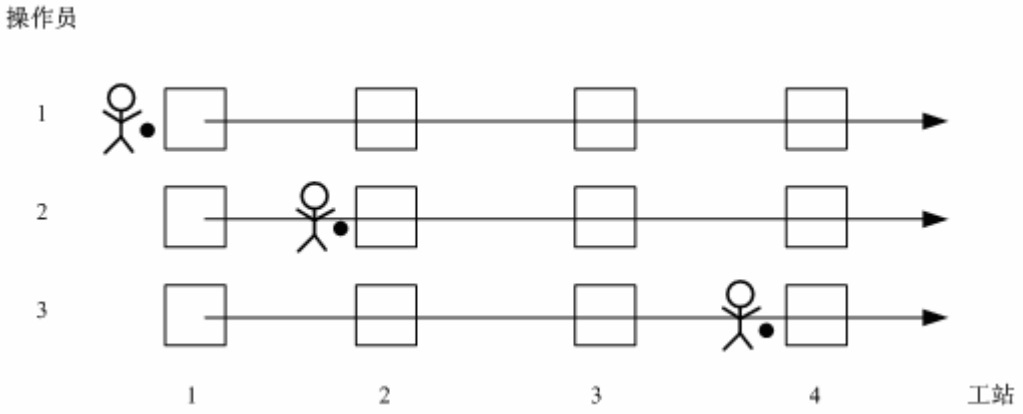

图 7.13 配备完全交叉培训员工的充足产能产线

在这样的系统中，产能是用人力而非设备来定义的。为了刻画产能，我们仍用 T_0 表示、加工任务穿越整个系统的平均时间，并假定它独立于被指派到加工任务上的工人。进一步地，我们假定一旦工人开始一个加工任务，他或她会持续到这个加工任务的完成。加工任务中途的作业停止不会提升产出却只会增加周期时间，所以没有理由那样做，除非一些客户比他人有较高的优先权。在这些假定下，只有工人变得可用时才向系统投放加工任务，并由于没有设备造成的阻塞，周期时间总是 T_0。如果产线上有 n 个以同样的效率工作的工人，则每人每隔 T_0 时间产出一件，意味着产出是 n/T_0。

因为充足产能情形是理想状况，对我们的假定的任何改变都会降低产出。这些改变的例子包括设备不充足从而发生阻塞，加工任务间歇到达可能引起饥饿，不完全交叉培训导致加工任务在某些工站可能不得不等待“专员”，或者其他一些使工人不能保持一直忙碌状态的原因。因此，我们可以陈述以下的工厂物理学定律。

定律（人力产能 Labor Capacity）：配备 n 个经过交叉培训、具有理想工作速率员工的产线

的最大产能为:

$$TH_{\max} = \frac{n}{T_0}$$

这条定律提供了一种把人力引入产能计算的方法。例如，在一个工站数量比工人多的产线上，设备的瓶颈速率 r_b 可能是一个对产线产能的糟糕估计。对于产出受人力约束的产线，n/T_0 可能是个更现实更有用的产能上限。这个上限适用于广泛范围的系统，包括那些配备完全或不完全交叉培训的员工的产线。

然而，一个工人可以同时处理多个加工任务的系统并不适用。例如，一个操作员同时看管多台自动化设备的制造单元的产出超过 n/T_0。这种系统通常被适当地认为是设备约束型的，操作员不可用表现为产能的扰动与变动性的升高。我们将在第八章讨论扰动因素。

7.4.2 完全柔性情形

为了深入理解设备和人力都是怎样影响产能，我们下来考虑具有完全交叉培训员工（即，可以操作产线上每一个工站）的情形。并且，我们假定像充足产能情形一样，工人绑定于加工任务。然而，与那不同的是，设备数量有限，所以工人可能阻塞，如图 7.14 所示。工人一旦在线尾完成加工任务，就回到线首开始一个新的加工任务。

如果图 7.14 所示的工人有理想工作速率，则除了 WIP 水平现在等于工人数量，此产线在逻辑上等同于我们先前讨论过的 CONWIP 产线，。因此，它的表现将在最佳与最差情形之间，用实际最差情形定义最佳与最差之间的部分。（239|240）而且，我们先前所列的所有改进策略——增加产能，降低产线平衡度，使用并联机工站，以及降低变动性——仍适用于这种情形。

图 7.14 完全交叉培训的工人绑定于加工任务的产线

完全交叉培训的工人全程跟随加工任务通过产线的假定可能在许多情况下并不现实。例如，如果工站需要极为不同的技能，可能就应该让工人一个接一个地传递加工任务。一种机制就是蚂蚁件货法（*bucket brigade*）（见 Bartholdi 和 Eisenstein 1996）。在这个系统中，无论最下游的工人何时完成加工任务，他或她就会向产线上游移动并接收紧挨着的那个上游工人的加工任务。那个工人依次向上游接收。持续不断地这样，直到产线最上游的工人接收了一个新的加工任务。如果所有的工人速度一样并且没有因交接加工任务造成延迟，则这个系统与图 7.14 描绘的没有逻辑上的差别。WIP 水平由工人数量设定，产线仍按 CONWIP 方式运行。只有指派给每个加工任务的工人的身份发生了变化。

从逻辑上讲，蚂蚁件货法与工人绑定于加工任务的系统可能没有什么区别，但实际上确有不同。每个工人趋向于在一个区域内操作机器。的确，当所有的加工时间都精确确定（即，最佳情形）时，产线稳定于每个工人通过同样的工站序列完成加工任务的循环状态。交叉培训和加工任务转移使产线可以自平衡，所以每个工人在一个加工任务上花费相等的时间。这种类型的系统已经有效地应用于汽车座椅组装（见第十章对丰田的这种系统的讨论）、仓库

拣货、快餐三明治制作（赛百味）。

注意到在蚂蚁件货发中还是有可能发证阻塞的。无论何时，上游工人一旦追上了紧挨的下游工人，除非有额外的设备，否则他将会被阻塞。因此，有必要安排工人来最小化阻塞的发生频率，把最快的放在下游，最慢的放在上游。Bartholdi 和 Eisenstein（1996）说明了这种从最慢到最快的安排可以显著提高产出，并觉察到它往往成为使用这种系统的行业的惯例。

7.4.3 配备柔性人力的 COPWIP 产线

如果工人绑定于加工任务（或者像蚂蚁件货法中的直接向下一个转交加工任务），则系统中加工任务的数量总是等于工人的数量，并且系统从逻辑上表现为一条 CONWIP 产线。但许多系统，如果不是绝大多数，加工任务的数量明显超过工人数。如果工人能在系统中移动，并在不同的工站作业，则系统绩效将取决于怎样有效地分派工人来促进整个流程。这就可能会变得复杂，因为在系统中动态分配工人的方式有无数种。

将蚂蚁件货法自然延伸到加工任务多于工人的情形，一种方法就是让任何空闲的工人去取上游的下一个加工任务，或者从上游工人那里或者从缓冲区那里（见图 7.15 对机制的图示）。（240|241）无论何时工人因下游工站繁忙而阻塞，他就把加工任务放在工站前面的缓冲区，然后去上游取另一个。这会持续到系统中的加工任务总数不超过某个预设的限制（如果没有这样的限制，线首速度快的工人就会用 WIP 将产线淹没）。

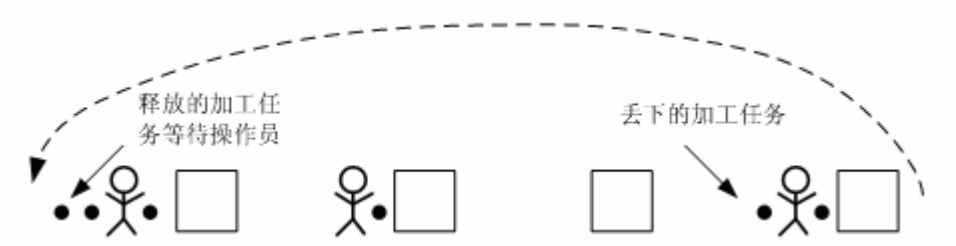

图 7.15 加工任务有丢下的蚂蚁件货法的 *CONWIP* 产线

如果所有工站由单机构成，传递就不可能发生，则在任何时候，工人 *n*（线尾的工人）将会在最下游作业。工人 *n*-1 在次下游没有被 n 阻塞的位置作业。依此类推到所有的工人。如果在多机工站间传递是可能的，工人就会混乱了。但根本目的还是让工人尽可能地执行最下游的加工任务。保持工人繁忙会趋向于最大化产出；执行下游的加工任务会趋向于最小化周期时间。因此，我们有理由期望这项政策运行良好。

当然，其他的柔性人力政策也是可能的。哪一个合适取决于多种因素，如员工交叉培训的程度，不同工站工人的相对速度，加工任务从一个工人传递到另一个的效率。如果工人速度没有差别，则系统产出完全取决于未被阻塞的加工任务由于工人不足而闲置的频率。如果它从不发生，则系统会像规则的 CONWIP 产线一样运行；如果它经常发生以致每个工人就像绑定于一个加工任务一样，则系统会像加工任务只有工人数那么多的 CONWIP 产线一样运行。因此，我们得出有柔性人力的 CONWIP 产线的产出极限，如以下的工厂物理学定律。

定律（有柔性人力的 CONWIP 产线 CONWIP With Flexible Labor）： 在一条有 n 个相同的工人、ω 件加工任务（其中 $\omega \geqslant n$）的 *CONWIP* 产线上，当未阻塞的加工任务可得时，任何不使工人空闲的政策都会使产出水平 $TH(\omega)$ 在以下区间

$$TH_{CW}(n) \le TH(\omega) \le TH_{CW}(\omega)$$

其中 $TH_{CW}(x)$ 表示所有机器配备员工、有 x 件加工任务的 *CONWIP* 产线的产能。

这条定律使我们更深刻地理解系统中交叉培训的价值。例如，在一条工人数至少等于临界 WIP 值、绩效接近最佳情形的有固定工人的产线上，很清楚地可见交叉培训没有多少好处。原因是在任何高于临界值的 WIP 水平下，CONWIP 产线的产出将会接近瓶颈速率，所以 $TH_{CW}(n)$ 近似等于 $TH_{CW}(\omega)$。系统几乎没有变动性，所以没有什么理由看重工人在工站之间移动的能力。

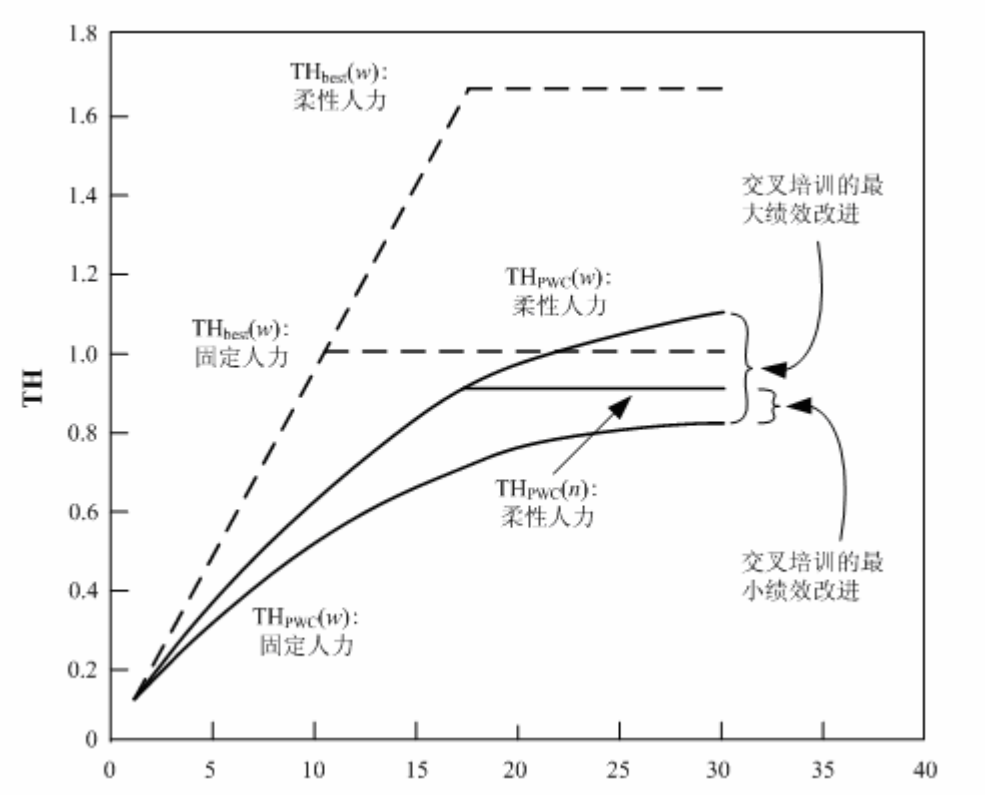

图 7.16 利用柔性人力提高 *CONWIP* 产线的绩效

另一方面，如果产线有显著的变动性，则交叉培训的带来的潜在改进是很重要的。（241|242）为了理解这一点，考虑一个瓶颈速率 r_b = 1 件/小时、原始加工时间 T_0 =10 小时、配备 n = 17 名工人实际最差情形产线。目前，产线按实际最差情形运行（图 7.16 中的 TH-WIP 曲线标记为“$TH_{PWC}(\omega)$：固定人力”）。但是，假定我们培训员工，使他们胜任任何操作；当工作变动需要时，可以使工人们换岗位到缺人的工站。根据人力产能定律，这样会使有效产能高达 n / T_0 = 17/10 = 1.7 件/小时。如果是这样，产线仍按实际最差情形运行，则产出曲线会相应地向上方移动。根据有柔性人力的 CONWIP 产线定律，实际产出将会在 $TH_{PWC}(n)$

和**错误！链接无效。**之间。虽然我们无法准确地说出绩效提高了多少，但它是显著的。结论就是，通过动态平衡产线，交叉培训员工可以增加有效产能从而达到更高的产出。

从先前的分析中，我们可知高变动性/高平衡度系统比低变动性/低平衡度系统表现得更像实际最差情形。因此，这些条件使交叉培训很有吸引力。原因是平衡的、高变动性的系统容易使绑定于工站的工人空闲。因此，允许工人跟随加工任务会避免一些空闲，从而提升产出。

我们应该注意到，虽然单机工站系统也趋向于表现实际最差情形，但交叉培训通常对它们不具吸引力。通过降低因机器不足而引起工人阻塞的频率，并联机工站系统实际上适用柔性人力政策。在极端情形下，有足够的并联能力来防止阻塞，系统可以达到充足产能情形，此时人力成为唯一约束。

7.5 结论

在这一章中，我们通过研究周期时间、在制品、产出和产能之间的关系，检视了单一产线的基本表现，得到如下结论：（242|243）

1. 一条产线可以由两个独立参数来恰当地描述：瓶颈速率r_b和原始加工时间T_0。然而，正如我们所了解的，具有同样r_b和T_0的产线可能有一系列不同的行为表现。我们将在接下来的两章里探究这种差异的原因。

2. 里特定律（$WIP = TH \times CT$）提供了任何工站、产线或系统的三个长期平均绩效量度值之间的基本关系。

3. 最佳情形定义了任何有确切的r_b和T_0值的产线在给定的 WIP 水平下的最大产出和最短周期时间。最差情形定义了任何有确切的r_b和T_0值的产线在给定的 WIP 水平下的最小产出和最长周期时间。实际最差情形提供了一个中间状况，作为“优”、“劣”系统之间的一个有用的界限。

4. 临界在制品水平，定义为$W_0 = r_b T_0$，表示现实的理想 WIP 水平（与不现实的理想水平——零库存相对，后者将导致零产出）。在W_0水平下，一条最佳情形（即，零变动性）产线可以同时达到最大产出（即，r_b）和最短周期时间（即，T_0）。

5. 最佳情形和最差情形都出现在无随机性的系统中。最差情形起因于不良控制造成的高变动性而非随机性。实际最差情形代表了最大随机性的状况。

6. 当 WIP 水平很高时，减少原始加工时间T_0对周期时间几乎没有影响，而增加r_b会有很大的影响。

7. 其他条件都相等（即，和T_0相同）的情况下，不平衡的产线比平衡的产线表现出较少的阻塞。

8. 产线受设备和人力组合的共同约束。设备产能由瓶颈速率 r_b 限制；人力产能由 n/T_0 限制，其中 n 是产线中工人的数量。

9. 有高度加工变动性和平衡工站的系统最适合采用交叉培训和柔性人力政策。此外，并联机工站有助于柔性人力政策的实行。

从工厂动力学基础的分析中浮现出一条线索，就是通过增加产线的产能或者提高产线的效率，可以在较低的 WIP 水平达到同样的产出。正如我们在涉及实际最差情形时暗示的，提高产线效率的首要方法是降低单个工站的变动性。为了评价产能增加与变动性降低之间的相关效果，我们必须更深入地发展工厂物理学来描述涉及随机性的系统的行为。我们将在接下的第八章和第九章做此工作。