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

第十章 推式与拉式生产系统

You say yes.
I say no.
You say stop,
And I say go, go, go!
——约翰·列侬与保罗·麦卡特尼《*Hello goodbye*》

10.1 引言

实际中对 JIT 的描述都采用了推式（*push*）与拉式（*pull*）生产系统这两个术语。然而，推和拉的定义常常不是很精确，并因此在美国产生了一些对于 JIT 的困惑。

在这一章，我们从概念水平上提供一个推和拉的正式定义。通过它们的具体实施来区别推和拉的概念，我们说现实中的系统往往是推和拉的混合。更进一步地，对比分别处于两个极端的“纯粹的拉”生产系统与“纯粹的推”生产系统，我们获得了使拉式生产系统运行更有效的因素。这种见解暗示出存在着许多种实现拉式生产的益处的方法。究竟哪一种最好，取决于一系列的环境因素，正如我们在本章讨论，并将在第三篇中进一步探讨的那样。

10.2 定义

JIT 之父，大野耐一（Taiichi Ohno），只在非常广泛的意义上使用拉这个术语：

制造商与生产车间不能再仅仅将生产建立在桌上那一纸计划的基础上，然后再将产品分配，或者推到市场上。它已经成为持有不同价值系统的客户或使用者站在市场前沿的必然结果；他们认为，以需要的数量、在需要的时刻拉出他们需要的产品，理所当然。

Hall（1983，39），在美国关于JIT的最著名的教科书之一中，更加具体地用事实来定义了拉式生产系统，“使用者因为需要而取物料”。（339|340）虽然他承认可能有不同的拉式系统，但是唯一详细描述的系统是我们曾经在第四章中讨论过的丰田的看板。[1] Schonberger（1982），在美国另一本关于JIT的主流书籍中，严格在丰田式的看板系统的背景下谈及拉式系统。因此，拉这种形式被认为与看板（*kanban*）类似也就不足为奇了。

然而，我们不认为这么狭窄的定义是大野耐一先生所希望的。我们认为，将拉仅仅解释为看板是对预期目标的彻底颠覆：赋予拉更多具体说明的同时，它掩盖了拉的精髓。它混淆了概念（拉）与执行（看板）。为了从工厂物理学的角度来讨论拉的概念，给推拉系统一个全面而简单的定义是很重要的。

[1] 霍尔又将拉式系统表述为广播（*broadcast*）系统，在这个系统中总装进度计划（FAS）被广播到产线中所有的起始点来触发加工任务的投放。但是，他提醒到因为 FAS 是外生的，这个系统不能严格限制其中的库存总数。通过提出将 FAS 信号作为宽松的拉式信号（*loose pull signals*），他区别了广播系统与看板系统中的控制。因为广播系统限制 WIP 的失败，我们根本不能确信它是否能被称作拉式系统。

10.2.1 推与拉的关键区别

将推与拉区分开的是引起制品在系统中运动的机制。根本性地，投料触发来自推式系统之外，拉式系统之内。更正式地，我们定义推式和拉式系统如下：

定义：推式系统根据外部需求计划制品投放，而**拉式**系统则根据系统自身的状态授权制品投放。

图 10.1 中简略地描绘了推与拉的对比。严格地说，推式系统是被外生计划要求时精确地将制品投入生产流程（工厂、产线、或工站），投料时间不会因为制程自身发生了什么事情而调整。与之相反，只有接到产线状态改变而产生需要开始的信号时，拉式系统才会允许物料进入流程。典型地，像在丰田的看板系统中，这些授权信号是产线中某些点制品加工完成的结果。值得注意的是，这个定义与实际上操作任务的人没有关系。如果下游作业员自上游流程收到制品，但这一举动是根据外部排配，那么这就是推；而如果上游作业员将制品递交给下游流程，但这一举动是对下游流程状态改变的反应，那么这就是拉。

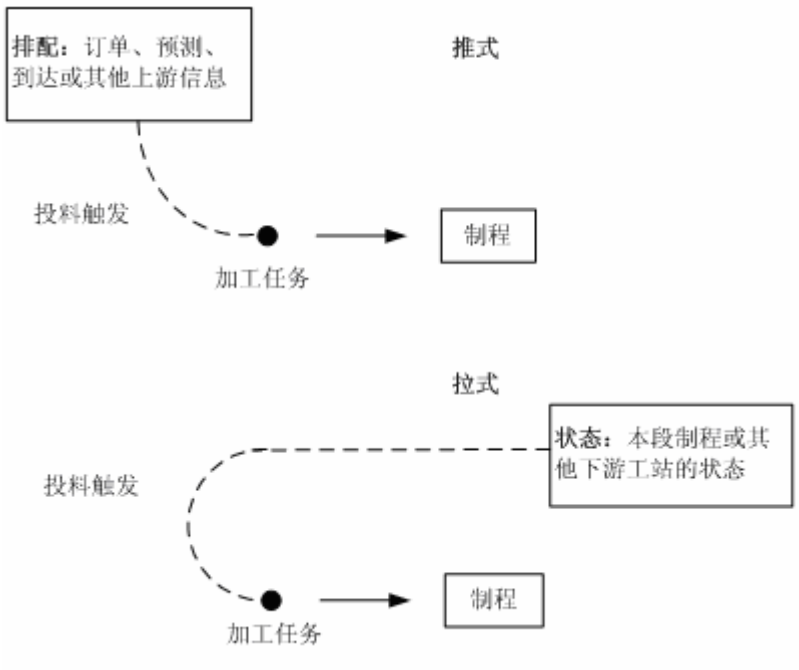

图 10.1 推式与拉式系统的投料触发

另一种有益于考察推与拉区别的方式是，推式系统由其内在属性而决定了是接单生产(*make-to-order*)，而拉式系统则为备货生产(*make-to-stock*)。也就是说，是订单（或预测），而不是系统状态驱动着推式系统的计划。拉式系统则以系统某处库存不足为批准投料的信号。从这个角度看，基准库存点模型（base stock model）当库存降低到某一特定水平之下即触发订单，是拉的方法；MRP 根据客户订单建立规划然后根据规划投放订单，是推的方法。

当然，大多数现实世界中的系统包含了推和拉两个方面。例如，如果一件任务由 MRP 排配投放，但是因为考虑到产线的拥堵又保持了这个任务，那么这种效果就是一个混合式推-拉系统。（340|341）相反地，如果一个看板系统生成批准生产的卡片，但实际的任务触发却因为对部件的预计需求不足（即，没有在主生产计划中体现）而推迟，那么，这也是一个混合式系统（如，见 Wight 1970，Deelersnyder 等 1988，Suri 1998）。我们将在第三篇中讨论

混合式系统的优点，并提供一种实用方法。

我们列出推与拉的显著区别不是建议使用者们严格地选择其中一个或另一个。秉着工厂物理学的精神，我们更愿意用自己的详细说明来解析拉式系统的好处并探究其根源。在某种意义上，我们采用了与物理学类似的方法——将机械系统放在无摩擦的理想环境中去考虑。并不是说理想环境是普遍存在的，而是说在这样一种理想环境下，地球引力、加速度、速度等概念会变得更清晰。就像在理想环境中探究经典力学是现实物理系统分析的基础一样，我们对于纯粹的拉式与推式生产系统的观察将会提供一个分析现实生产系统的基础。

10.2.2 推-拉界面

是否要用、怎么使用拉只是蓝图的一部分，在哪儿用也很重要。甚至在一个单独的生产系统中，部分按照拉式操作都是可能的。推-拉界面（*push-pull interface*）是确定何处放置拉式过程时的一个有用概念，它将生产流程分成推段与拉段[2]。成功地选择这个界面的位置能使系统利用拉的优点而取得战略优势，同时还能保持推式系统的客户驱动的特征。（341|342）

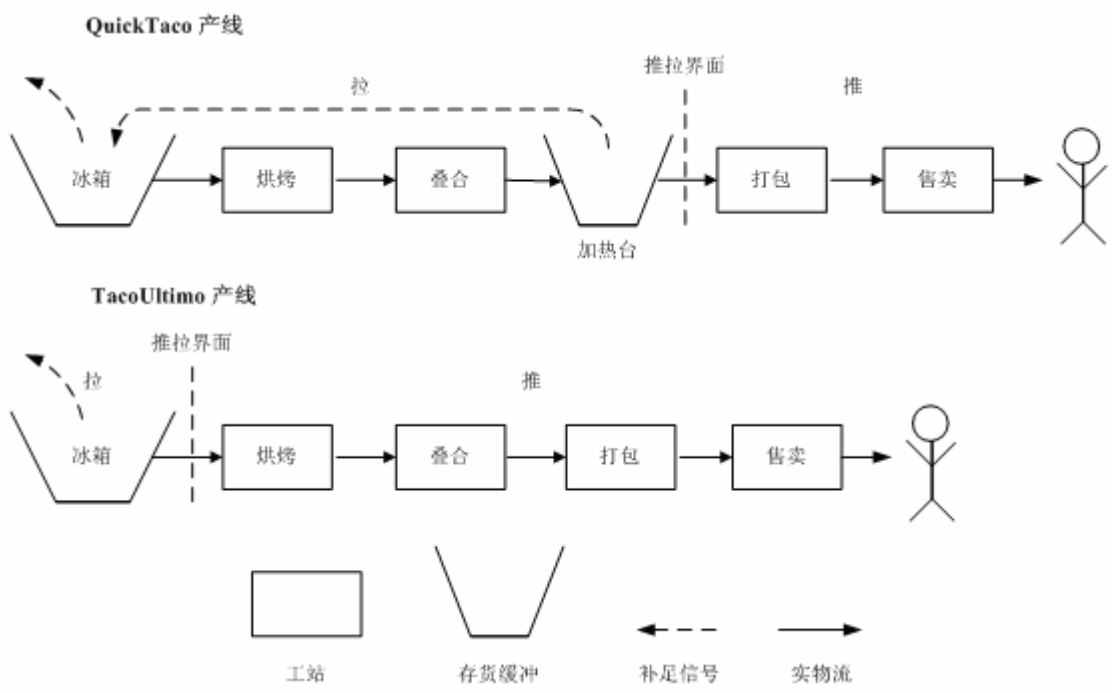

图 10.2 推-拉界面的位置图示

为了理解推-拉界面的概念，不妨将推定义为“接单生产”的形式来，而将拉定义为“备货生产”的形式。为了理解如何将相似的产线有区别地分成推段和拉段，我们考察图 10.2 所示的两个生产系统。在 QuickTaco 的前端，玉米煎饼备货生产以维持加热台的固定库存水平，使得这一部分表现为拉式；产线的后端只有在客户订单触发后才会移动制品，因此这一部分表现为推式。推-拉界面位于加热台处。相反地，在 TacoUltimo 中，玉米煎饼的移动只能由客户订单来触发，因而它完全是一个推式系统。推-拉界面位于冰箱处，在那里按照库存目标来储存原材料。

对比 QuickTaco 与 TacoUltimo 的相对优势，我们可以获得权衡推-拉界面位置的洞察力。

[2] “推-拉界面”这个术语，我们受益于 HP 的 Corey Billington；它被创造来辅助描述 HP“供应链管理纲要”之中的部分实践活动。见 Lee 与 Billington（1995）对 HP 供应链创新的概述。

TacoUltimo，因为它完全由订单驱动并且几乎全部以原材料形式保有库存，所以具有柔性优势（即，它能生产客户想要的任何一种煎饼）；QuickTaco，因为它保有成品煎饼库存，所以具有反应能力优势（即，它能向客户提供更短的提前期）。因此，需要在速度和柔性之间进行权衡。通过将推-拉界面移近客户，我们可以缩短提前期，而代价是降低柔性。（342|343）

所以对于一个给定的系统应该如何来选择推-拉界面的位置呢？既然它取决于客户的偏好以及生产流程的实质细节，那么这就不是一个简单的问题。但是，我们能够提供一些观测资料和现实中的例子。

首先，请注意速度是促使我们将推-拉界面移近客户的首要原因。因此只有在客户看来增加的速度确实显著提高了服务水平时，这么做才是有意义的。例如，一个线上周期时间为两小时却实行日末出货（end-of-day shipments）的生产系统，通过移动推-拉界面缩短了周期时间，但客户却看不出提前期有任何不同。甚至在速度显然很关键的快餐行业，也有一些使用 TacoUltimo 型产线的餐馆。他们这么做是因为确信整条产线的周期时间已经足够地短，能够使系统满足客户的期望。然而，在速度压力特别大的就餐高峰期，许多 TacoUltimo 型餐馆会转向 QuickTaco 型。

其次，我们观察到推-拉界面的选择受到它自身流程的巨大影响。例如，在玉米煎饼产线中我们可能会建议将推-拉界面放在装配线中间的某个地方。那就是说，做好玉米面饼并在里面装上肉，让口开着，等待涂上配料（toppings）。然而，这会导致存储和质量问题（如，部分叠合的煎饼会散开），因而可能是不可行的。

第三，请注意推-拉界面位置的经济性会受到产品通过系统时是如何被定制化的影响。在终端品目（end items）非常少的系统（如，一个只有原木和胶水等原材料，并且生产少数集中不同厚度夹板的夹板制造厂）中，将推-拉界面放在制成品（finished goods）处是相当明智的。然而，在一个有许多终端品目的系统（如，一个零部件可以组合到一系列不同 PC 成品的 PC 装配厂）中，保持制成品库存是非常昂贵的（见 8.8.2 节中安全库存集结（aggregation）的例子）。例如，在玉米煎饼生产系统中，将推-拉界面放在包装之后不是个好主意，因为那需要将煎饼按照所有需要的类型（size）和组合（combination）分装成不同的袋来进行储存。

最后，请注意认识到定制化与变动性汇聚（variability pooling）紧密相关，如我们在第八章中的介绍。在一个越向产线下游产品定制化程度越高的系统中，推-拉界面向上游移动可以降低为应对需求变动性而设立的安全库存的数量。例如，贝纳通（Benetton）使用这样的系统，没有着色的毛衣进行备货生产，然后按订单进行染色。也就是说，他们将推-拉界面从染色工序之后移到之前。通过这样做，他们分担了对不同颜色毛衣的库存，并因此降低了达到给定客户服务水平所需的库存成本。

一些其他的因重新定位推-拉界面位置而提高整体系统绩效的真实例子包括以下这些：

1. *IBM* 有一个印制电路板生产车间，能够将玻璃纤维和一些不同厚度的铜制成 150 种不同的电路板。产线的前面部分生产基板（*core blanks*）——铜片和玻璃纤维，所有的电路板制造都是从这里开始的。只有八种不同的基板，经由固定的成批叠压制程生产出来，很难将其与客户订单匹配。管理层对基板储存进行选择（即，将推-拉界面从原材料处移动到叠压制程之后的库存点）。这样做的结果是从能够被客户感知的提前期中减少一到两天的周期时间，但只需额外增加一点点库存成本。（343|344）

2. 通用汽车引入了一个新的车辆配送系统，从佛罗里达的卡迪拉克开始，在地区配送中心中储存那些受欢迎的型号（《华尔街日报》，1996 年 10 月 21 日，A1）。目标是任何一个特许经销商都能够为用户提供 24 小时“pop cons”送货服务。其他型号的提前期将保持在传统的数周的水平。所以，不同于传统的系统中推-拉界面在装配车间（对于接单生产的车）

或者经销商处（对于备货生产的车），这一新的系统将推-拉界面置于地区配送中心。这么做的愿望是在经销商之间分担库存，通用汽车将能够在总库存成本很低的情况下对大部分销售提供快速送货服务。这个例子向我们说明了同一系统中不同产品的推-拉界面位置不同是可能的更是值得的。

3. 惠普公司面向欧洲市场生产多种不同的打印机，然而，由于不同的电压以及插座，打印机在不同的国家需要不同的电源装置。他们将生产流程修正到停止于电源处，这样就可以将没有差别的打印机运到欧洲的配送中心，在那里完成打印机的客户定制，针对不同的国家安装不同的电源（参阅 Lee、Billington 和 Carter 在 1993 年关于这个系统的讨论）。将推-拉界面放在以欧洲为基础的配送中心而不是以美国为基地的工厂，可以从客户提前期中取消整个的运送周期时间。同时，以电源的形式来推迟客户定制，使得惠普能够在各个国家之间分担库存。这是一个**延迟（postponement）**的例子，这种情况下产品和生产流程以允许尽可能晚的定制为目的来设计。延迟策略可以用来在高度定制的制造环境中建立快速的客户响应，有时候这项技术被称作**大规模定制（mass customization）**（Feitzinger 和 Lee，1997）。

10.3 拉的魔法

什么让日本制造系统如此之好呢？我们希望读者已经由第四章有所归纳：这个问题没有一个简单的答案。二十世纪八十年代那些形象鲜明的日本公司的成功是多种不同实践的结果，从换模时间压缩到质量控制再到产品快速引入。进一步说，这些公司是在一个文化、地理及经济都与美国很不相同的环境中运营。如果我们想理解 JIT 成功的精髓，我们必须缩小关注点。

在宏观水平上，日本的成功是以将符合潮流的优质产品以有竞争力的成本推入市场并获得广泛响应的这种能力为前提的。在微观水平上，这是由一个有效的生产控制系统达成的。这个系统通过高的产出、低的库存和少的重工实现低成本制造。它通过实行内部高质量进而推动外部高质量，通过保持稳定可预期的输出流实现优质的客户服务。这个系统还具有足够的柔性，能够允许产品组合内的变化（只要不是太快或者太明显），从而对变化的需求有着快速的响应能力。

在所有的有价值的特征中，什么是使得日本生产控制系统成为实施商业战略的关键要素呢？（344|345）在美国的 JIT 著作中，拉的作用是基础性的。Hall（1983，39）引用了一位描述拉的精髓的通用汽车领班的话：“不要一直做东西然后不知道送哪儿，得有人来取它（You don’t never make nothin’ and *send* it no place. Somebody has to come get it）”。

我们不同意这个观点。我们将在这一章中展开的观点，是说工站中拉的部分仅仅是一种达到目的的手段。而使得拉式系统具有很多优势的真正的潜在原因是：系统中最大库存的数量是有限制的（*there is a limit on the maximum amount of inventory in the system*）。在单（单卡片）看板系统中，容器的数量由生产卡片的数量限制。不管车间里发生了什么，WIP 的水平是不能超过提前设定的限制的，但是这种结果在看板系统中是不受限制的。因为一个拉式系统是以库存水平空缺为基础授权投料的，或是与备货生产系统等价的。任何一个真正的拉式系统将为 WIP 设置一个上限。向我们在接下来的段落中讨论的那样，JIT 的主要好处可以归因为 **WIP 上限（WIP Cap）**的存在，而不论它是怎么实现的。魔法在于 WIP 上限，而不是拉的过程。

10.3.1 降低制造成本

如果 WIP 是有限制的，那么产线的中断（如，机器失效、质量问题引起的停线、产品

组合变化引起的减速）不会导致 WIP 超过预定的水平。应该注意到在一个单纯的推式系统中，是没有这样的限制的。如果一个由 MRP 成的计划被严格地执行（即，没有依据工厂情况做调整），那么这个计划将会没有节制的提前生产并因此让工厂堆满了 WIP，引起**WIP 爆炸（WIP explosion）**。

当然，我们在现实中从未看过有无限 WIP 的车间。最终在情况坏到一定程度时，管理层不会坐视不理。他们将计划着加班，雇用临时工来提升产能。他向车间推出完工日期和投料限制，换句话说，管理层停止使用纯推式系统。最终一切归于正常……直到下一次 WIP 爆炸（请看第九章关于超时的邪恶循环（overtime vicious cycle）的讨论）。这里的关键点是在一个推的环境中，只有在问题出现之后才会采取纠正措施，而这时 WIP 已经失控了。

在一个建立了 WIP 上限的拉式系统中，投料会在系统过载之前被停止，自然地，产出会下降，无论 WIP 水平是否允许骤升这都是可能发生的。举个例子，如果一台机器坏了，那么这台机器之前所有的 WIP 的生产都不会再继续进行。但是如果将 WIP 放在系统之外，WIP 上限可以保持一定程度的柔性，这种柔性会在触发实施后失去。只要那些加工任务作为一纸命令而存在，它们对于工程和排配优先序变化的适应相对容易。但是一旦这些加工任务被实施，而且被赋予了“个性”（如，印制电路板收到电路）时，优先序的变化要求高成本和剧烈的提速，而且工程变更几乎是不可能的。因此，WIP 上限能够通过减少加速和工程变更成本来降低制造成本。

除了提高柔性，拉式系统能够促进更好的工作投放时机。为了明白这一点，请看一个周期性地允许过量加工任务进入的纯拉式系统（即，在那些新加工任务因为拥塞而不能快速处理的时候）。这会抬高平均 WIP 水平，但产出不会因此而提高。WIP 上限，与实现它的拉的机制无关，能够在实现要求的确定的产出的情况下减少平均 WIP 水平。这能够直接减少与保持库存相关的制造成本。（345|346）

10.3.2 削减变动性

保持高的客户服务水平的关键是产线流量的可预测性。特别地，我们需要低的**周期时间变动性（cycle time variability）**。如果周期时间变动性低，我们就能知道一项加工任务完整地通过车间需要多久。这就使我们能够对客户报出精确的完工日期，然后去满足他们。低周期时间变动性还帮助我们提供给客户更短的提前期。如果周期时间是 10±6 天，那么我们将不得不报出 16 天的提前期，从而确保高的客户服务水平。从另外一个方面，如果周期时间是 10±1 天，那么，就能够报出 11 天的提前期了。

看板系统能够比纯拉式系统达到更短的周期时间。因为周期时间随着 WIP 水平的增加而增加（根据里特定律），而看板能够阻止 WIP 爆炸，同样它能够阻止周期时间爆炸。然而，请注意产生这种效应的原因又一次是 WIP 上限，而不是每个工站的拉式系统。因此，任何一个限制了 WIP 数量的系统都能够阻止可能发生在纯推式系统中的强烈 WIP 回转以及由此而产生的长周期时间。

看板常常能够直接地减少工站中周期时间的变动性。这就是关于 JIT“降低水位暴露礁石”的类比。关键地是，看板限制了系统中的 WIP，使系统更容易受到变动性的攻击，而且能够因此对管理层形成一种持续改善的压力。

我们用图 10.3 中的简单例子展示了这个类比背后的直觉性知识。这个系统由两个机器组成，机器 1 的产品作为机器 2 的原料。机器 1 很快，每秒生产一个部件；而机器 2 较慢，每小时生产一个。假设采用看板体系（单卡片），将机器间的 WIP 限制在五件之内。因为机器 1 很快，在机器 1 运转的时候，这个缓冲区常常是满的。

然而，假设机器 1 常常遭受周期性的失效。如果故障时间长于五小时，那么机器 2，原先的瓶颈，将会感到饥饿。因此，由于机器 1 失效的频率和持续时间，机器 2 在大部分时间

里将会非常饥饿，尽管机器 1 的速度极快。

明显地，如果缓冲区容量（看板系统中卡片的数量）增加，机器 2 的饥饿程度将会减小。举个例子，如果缓冲区增加到 10 件，那么只有超过 10 小时的失效才会引起饥饿。过量的 WIP 能够有效的使系统免于遭受由故障引起的破裂效应。但是像我们之前注意到的那样，纯推式系统需要更高的平均 WIP 水平去完成设定的产出水平。纯推式系统倾向于用这种方式来掩盖机器 1 故障带来的影响。推式系统有着更高的 WIP 水平，因而故障的破坏性也较小。只要管理层愿意维持高的 WIP 水平，提高机器 1 的可靠性的压力也就微乎其微了。

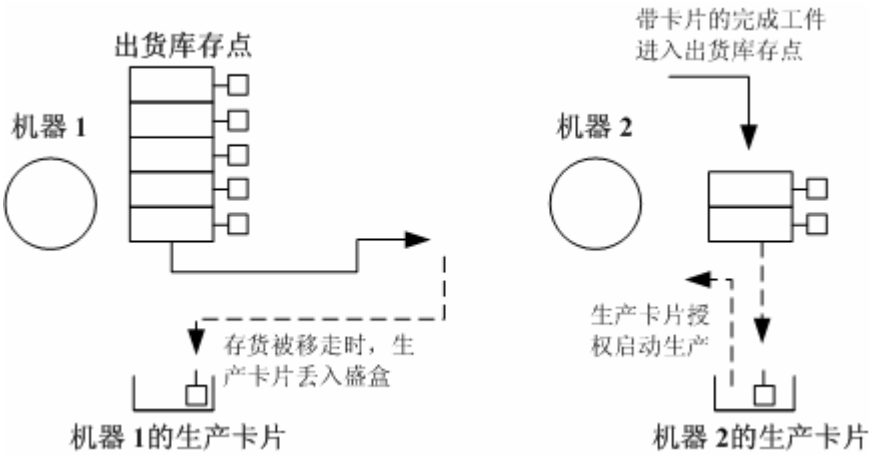

图 10.3 由有限容量缓冲区连接的工站

像 JIT 文献中指出的那样，如果想在低的 WIP 水平（及短周期时间）的情况下保持高的产出水平，就必须减少这些具有破坏性的变动性的来源（失效、换模、重工等等）。我们再次注意到，这种压力的来源是有限的 WIP 水平，而不是每个工站的拉式机制。确切地说，每个工站的拉式机制控制了流程中每个点的 WIP 水平，这不是必须有一个通用的 WIP 上限的情形。然而，通过 WIP 上限来减少整体的 WIP 水平将会减少不同工站间 WIP 数量的平均水平，从而能够造成促进持续改善的压力。通用的 WIP 上限是否能够在产线上合理地分配 WIP 将是我们在稍后会探讨的问题。

10.3.3 提高质量

质量常常是既被看作 JIT 的先决条件，也是 JIT 的好处之一。同样的，JIT 因为绝对需要（sheer necessity）从而促进质量的提高，同时建立使高质量更易实现的条件。

像在第四章中看到的，质量是 JIT 哲学的基本组成部分。原因是，如果 WIP 水平低，那么工站将会在处于流入缓冲区（库存点）的部件没有达到质量要求时感到饥饿。从物流的角度，这种效应与机器故障产生的效应十分相似。一旦 WIP 水平变得足够的低，系统中部件达标率必须高到能够保持合理的产出水平。为了确保这一点，看板系统常常伴随着统计过程控制（SPC）、质量为本的工作人员培训、源质量（quality-at-the-source）程序，以及其他一些提高整个系统质量水平的技术。因为质量水平越高，WIP 水平就越低，在 JIT 系统中为了减少 WIP 而进行的持续努力需要持续的质量提高。

在追求高质量的单一压力之外，JIT 能够更容易地直接促进质量的提高，因为在低 WIP 环境中，检验更加有效。如果 WIP 水品很高而且队列很长，质量保证（QA）检验很可能无法确定流程问题，直到生产出大量有缺陷的产品。如果再制品水平低，那么在质量保证检验前的队列比较短，缺陷将会及时被发现，从而能够在生产出大量不良部件之前去修正。当然，这是 SPC 的目的，即能够实时地检测过程质量。然而，在那些立即检验无法实现的地方，

比如说，在电路板车间，必须进行光学或电子检测从而确保质量，因此低 WIP 水平将会极大地增强质量控制程序的力量。

请注意，我们再次将看板或者 JIT 的好处归因于 WIP 减少的成果。因此，一个简单的 WIP 上限将能够在质量提高方面提供与看板相同的压力，在促进 QA 方面实现与看板同样的队列缩短。

然而，还有一个进一步与质量相关的好处常常被归因于看板系统的拉式活动。它的基本论点是，如果来自于下游工站的工作人员必须到上游工站去取部件，那么他们将能够检验这些部件。如果这些部件的质量不被接受，那么这个工作人员将会立即拒绝它们。这样的结果将会是更快的质量问题得到检验，以及更低的转运和加工这些不良部件的可能性。（347|348）

当物料搬运由一个独立的工作人员，如叉车驾驶员，来执行的时候，这一论点就不那么有说服力了。叉车驾驶员究竟是因为部件已经完成才将它们“推”到下一个工站，还是在看板系统中自发的将这些部件从前一个工站“拉”过来，对于他们执行质量检验的能力来说，并没有任何区别。

在部件很小并且工站很接近从而作业员能够自己移动自己的部件时，这一论点就很有说服力了。这时，假定如果下游作业员去拿到了需要的部件，他们将更有可能去检查这些部件的质量而不是由上游作业员简单地将其下线。但是，这个推论却不必要地将两个单独的问题联系在了一起。

第一个问题是，下游作业员是否会检查接收到的所有部件（推的或者拉的）。我们在工业中看到这样的执行方式，并不一定是拉式系统，作业员只有签了物料表格才能批准其转移。这个批准中隐含的是对质量的检验。

第二个完全独立的问题是在两个相邻工站间是否要限制 WIP 数量。这一章中，我们将在后面探讨这个问题。现在，我们仅仅是指出，每个工站处拉的质量保证一处可以通过为了达到 WIP 需求现在所使用的机制的独立检查处理而获得。

10.3.4 保持柔性

纯推式系统可以在产线拥挤时投放新的加工任务，却使得该任务停滞于产线中途的某个地方。这个结果意味着在很多方面失去了柔性。第一，那些已经部分完工的部件不能很容易整合工程（如，设计）变更。第二，高 WIP 水平妨碍优先序/排配的变化，因为部件将不得不被移出产线，以此来为优先度高的部件让路。最后，如果 WIP 水平高，那么部件的生产必须提前于限定日期被触发。因为随着计划范围的增加，客户的需求越来越不确定，系统只能依靠对未来的预测来决定触发时机。而且因为预测从来不会像我们期望的那样精确，这种依靠会导致系统的表现进一步退化。

建立了 WIP 上限的拉式系统能够阻止这些负面效应并且因此增强系统的整体柔性。通过在工厂过度拥堵时阻止部件的触发，拉式系统能够让订单停留在纸上的时间尽可能地长。这将会使工程或优先序/排配的变化变的容易。还有，尽可能迟地投料将能够使投料建立的基础——客户订单的稳定性具有达到最大程度的可能。实际效果将是提供客户响应服务能力的提升。

我们喜欢用空中交通管制的比喻来展示拉式系统在柔性方面的好处。当我们从 Austin，Texas 飞到 Chicago、Illinois 的时候，我们常常得在 Austin 机场紧张地等待，还可能因为机场的流量控制（*flow control*）而超出预订的启程时间。他们的意思是 Chicago 的 O’Hare 机场过载了（或是将在我们到达的时刻过载）。甚至即便我们能准时离开 Austin，我们也只能紧张地在 Michigan 湖上空盘旋，等待降落的机会。因此，空中交通管制聪明地（虽然很令人恼火）将飞机留在 Autsin 的机场上，直到 O’Hare 的拥塞得到了疏通（或是将在我们到达的时刻得到疏通）。实际结果就是如果我们按时离开了的话，降落的时间（迟了，确实是！）

很准确，但是我们用了更少的燃料同时减少了事故的风险。（348|349）还有很重要的就是，我们对任何其他的选择都持开放态度，比如如果天气变得很危险的话就取消航班。

10.3.5 提前释放加工任务（facilitating work ahead）

之前的讨论暗示拉式系统通过协调触发时机与产线的当前状况（即，在产线拥挤的时候停止投放）来保持柔性。协调的好处还可以延伸到车间状态良好的情况。如果我们严格地遵循拉式机制，并且每当 WIP 低于上限就释放加工任务，就可能在一切运行良好的时候超前于排配。例如，如果处于没有机器失效、人员问题、物料短缺这么一段时间，我们就可能生产出比预期要多的产品。纯推式系统不会有这么好的运气，因为投料是根据计划而不是系统状态来确定的。

当然，实践中的拉式系统对提前投料的程度一般会有一个限制。如果我们开始做一些工作，这些任务的交期是如此之远以至于对有些跟预测不一致，那么现在完成可能就会有风险。需求或工程变更可以在很大程度上否定提早完成的价值。因此，如果我们对需求已经有了一些必要的准备的话，那么减缓工作节拍是很有意义的。我们将在第三篇讨论这么做的机制。

10.4 常量在制品（CONWIP）

我们能想到的建立 WIP 上限的最简单的方式就是*尽管去做*（*just do it*）！那就是说，对于一个给定的产线，建立一个对产线中 WIP 数量的限制，而且在 WIP 临近限制或位于限制之上时不再触发就够了。我们把一个任务离开就会有一个新任务被引入线中的协议叫做**常量在制品（constant work in process）**，因为它能引起一个几乎恒定的 WIP 水平。

记起在第七章中用 CONWIP 协议控制 WIP，所以我们能够确定 WIP、周期时间和产出之间的关系。现在我们把它作为实际 WIP 上限机制的基础，首先进行定性描述，接着给出一个分析 CONWIP 产线绩效的定量模型。

10.4.1 基本结构

我们可以想象一条像图 10.4 那样运行的 CONWIP 产线，其中离开的任务将生产卡片送回到线首，从而授权新任务的投放。请注意这种描述 CONWIP 的方式暗含了两个假设：

1. 这条产线由单条路线组成，所有的部件沿这条路线流动。
2. 任务都是同样的，所以 WIP 能够合理的用一种单位（即，产线中加工任务或部件的数量）来计量。

如果工厂包含共享工站的多条路线，或者不同的任务需要在机器上进行大量不同的加工，情况就不这么简单了。然而，有很多方法可以确定这些复杂的因素。例如，我们可以在不同的路线间建立 CONWIP 水平。我们还可以以根据需要在关键资源上进行处理的数量来进行调整的单位“标准化任务”来表述 CONWIP 水平。（349|350）我们将在第三篇中解决这些实施的问题。现在，为了检查 CONWIP、看板、MRP 的关键不同之处，我们集中关注单产品单路线的产线。

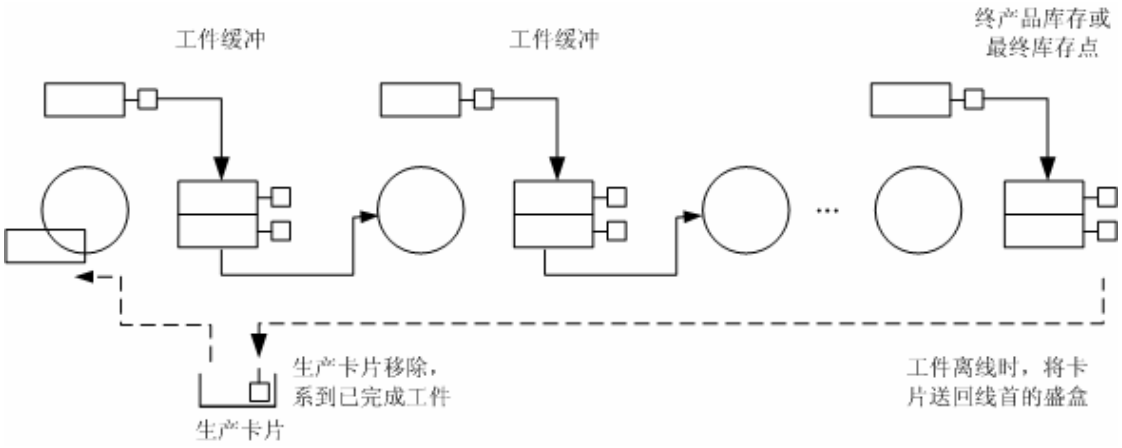

图 10.4 一条 *CONWIP* 产线

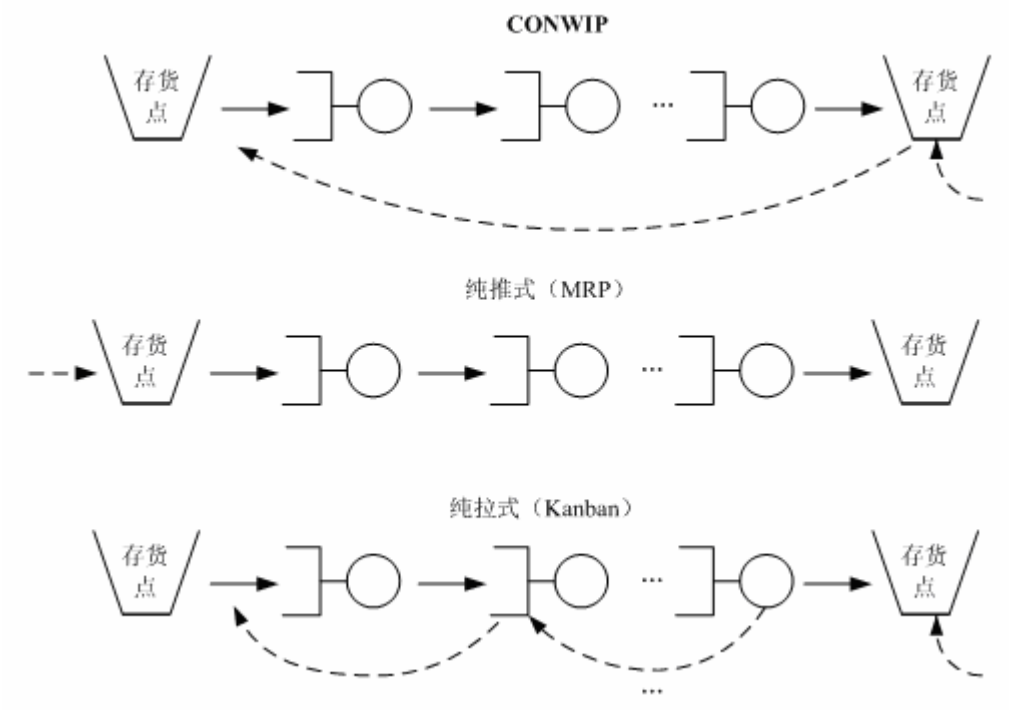

图 10.5 *CONWIP*、纯推式和看板体系

从建立模型的角度，一个 CONWIP 系统看起来像一个**封闭队列网络（closed queueing network）**，其中客户（任务）永不离开系统，但是会无限期地绕着网络循环，像图 10.5 中所展示的那样。当然，在现实中，进入的任务与离开的任务不同。但若是以建模为目的，这就没有任何区别，因为假设是所有的任务都是相同的。

相反，一个纯推式或者 MRP 系统表现的像**开放队列网络（open queueing network）**，其中任务进入产线，然后在通过后离开（图 10.5 中也有展示）。使任务进入产线的触发是由不考虑线中任务数量的物料需求计划引起的。因此，不像封闭的队列网络，任务的数量是随时间不断变化的。

最后，图 10.5 还描绘了一个**有阻塞的封闭队列网络（closed queueing network with blocking）**（单卡片）看板系统。像在 CONWIP 封闭队列网络模型那样，任务不确定的绕着网络循环。然而，不像 CONWIP 系统，看板系统对每个工站可以有的任务数量做出了限制，

因为每个工站生产卡片的数量为这个工站建立了最大 WIP 水平。每个生产卡片所起的作用就像位于工站前的有限缓冲空间那样。如果这个缓冲满了，上游工站就会被阻塞。

10.4.2 均值分析模型（Mean-Value Analysis Model）

为了分析CONWIP产线同时将它与推式系统比较，封闭的（CONWIP）系统的定量模型是很有用的，它类似于我们在第八章为开放（推式）系统提出的Kingman方程模型。对于那种所有工站都是由单台机器组成的情况，我们可以采用**均值分析法（MVA，Mean-value analysis）**[3]来进行。这种方法我们第七章中曾经用过，用来得出实际最差情形的产出和周期时间的曲线，但在当时并没有明确地说明。（350|351）这是一种迭代算法，是一种用WIP水平为ω-1 时产线量度来得出WIP水平为ω时产线量度的方法。它的基本思想是一件任务到达一个有ω件任务的系统中后可以看到根据有ω-1 件任务的系统的平均行为来分配其他ω-1 件任务。这在加工时间服从指数分布的情形下（$c_e = 1$）非常确切。对于一般分布的加工时间，它只是近似成立。这样，它提供了一个近似模型，很像开放系统Kingman的模型。

使用下列符号来描述一个 n 工站的 CONWIP 产线

$u_j(\omega)$ = WIP 水平为 ω 的 CONWIP 产线中工站 j 的利用率

$CT_j(\omega)$ = WIP 水平为 ω 的 CONWIP 产线中工站 j 处的周期时间

$CT(\omega) = \sum_{j=1}^{n} CT_j(\omega)$ = WIP 水平为 ω 的 CONWIP 产线的周期时间

$TH(\omega)$ = WIP 水平为 ω 的 CONWIP 产线的产出

$WIP_j(\omega)$ = WIP 水平为 ω 的 CONWIP 产线中工站 j 处的平均 WIP 水平

我们建立了一个均值分析模型来计算上述各个作为 WIP 水平 ω 函数的各个量度。下面的技术性注释中给出了细节。

技术性注释

像开放系统的 Kingman 模型那样，建立封闭系统的 MVA 模型的基本挑战是计算各个工站处的平均周期时间。我们将各个工站当作 $M/G/1$ 队列——即，有着泊松到达时间和一般（随机）加工时间的单机工站——看待，从而解决这个问题。$M/G/1$ 队列的关键结论如下：

1. 服务台繁忙的长期平均概率是

$$P(busy) = u$$

其中 u 为工站的利用率。（351|352）

2. 在一个随机到达的任务看来，处于被服务状态（即，被加工，而非排队等待）的任

[3] 不幸的是，MVA 不适用于多机情形。我们可以用一台快速机器代替几台并联机器（即，这样产能可以相同）来对工站做出近似估计。但由第七章可知，对于同样的产能，并联机器趋于比单机表现更优。因此，可以预见的是，这种近似将低估并联多机 CONWIP 产线的绩效。

务的平均数目是

$$E[no.of\ \ jobs\ in\ service] = P(busy)\times 1 + [1 - P(busy)]\times(0) = u$$

3. 在一个随机到达的任务看来，一个正在接受服务的加工任务的平均剩余加工时间（如果没有正在接受服务的任务，这个值为零）是（见 Kleinrock 的详细描述）

$$E[remaining\ \ process\ time] = P(busy)E[remaining\ \ process\ time\,|\,busy]$$

$$\approx u\frac{t_e(c_e^2+1)}{2}$$

请注意如果c_e=1（即，加工时间服从指数分布），那么在工站繁忙的给定状态下，剩余加工时间的期望值是t_e（一个刚开始加工的任务的平均加工时间），这显示了指数分部的无后效性。当c_e>1 时，剩余加工时间的期望值大于t_e，因为在具有高度变动性的系统中，随机到达的任务很有可能遇到长期任务。相反，如果c_e<1，那么平均剩余加工时间就小于t_e。

在这三个属性下，我们就可以用当前正在接受服务的人物的剩余加工时间加上处理队列中排在到达任务之前的任务的加工时间，再加上到达任务自身所需的加工时间，来估计一件任务花费在 WIP 水平为 ω 的 CONWIP 产线中工站 j 的平均时间了。因为队列中的任务数等于工站中的任务数减去正在接受服务（如果有的话）的任务数，我们可以把这个写为

$$CT_j(\omega) = E[remaining\ \ process\ time] + (E[no.\ jobs\ at\ station]$$

$$- E[no.\ jobs\ in\ service])t_e(j) + t_e(j)$$

现在，假设有一个任务，到达有 ω 件任务的产线，根据有 ω-1 件任务的产线的平均行为来分配其他 ω-1 件任务，采用上面对于剩余加工时间的表述，我们可以把这个写为

$$CT_j(\omega) = u_j(\omega-1)\frac{t_e(j)[c_e^2(j)+1]}{2} + [WIP_j(\omega-1) - u_j(\omega-1)]t_e(j) + t_e(j)$$

$$= TH(\omega-1)t_e(j)\frac{t_e(j)[c_e^2(j)+1]}{2} + [WIP_j(\omega-1) - TH(\omega-1)t_e(j) + 1]t_e(j)$$

$$= \frac{t_e^2(j)}{2}[c_e^2(j)-1]TH(\omega-1) + [WIP_j(\omega-1)+1]t_e(j)$$

请注意我们有一个替换表述，利用率$u_j(\omega) = TH(\omega)t_e(j)$，有了这个关于工站 j 周期时间的等式，我们可以很容易的计算出产线的周期时间（即，它是各工站周期时间的加和）。知道了周期时间，我们就可以运用里特定律来计算产出（CONWIP 产线中 WIP 水平恒定为 ω）。最终，在里特定律中将产出、各工站周期时间代入，我们就能算出各工站的 WIP 水平。

令$WIP_j(0)=0$，$TH(0)=0$，MVA 算法能够计算出周期时间、产出，还有运用迭代算法将其看作任务数的函数如一站接一站地计算出 WIP 水平，如下：（352|353）

$$CT_j(\omega)=\frac{t_e^2(j)}{2}[c_e^2(j)-1]TH(\omega-1)+[WIP_j(\omega-1)+1]t_e(j) \qquad (10.1)$$

$$CT(\omega)=\sum_{j=1}^{n}CT_j(\omega) \qquad (10.2)$$

$$TH(\omega)=\frac{\omega}{CT(\omega)} \qquad (10.3)$$

$$WIP_j(\omega)=TH(\omega)CT_j(\omega) \qquad (10.4)$$

这些公式在电子表格中很容易就能得出结果，而且可以用来生成表示 CONWIP 产线中除过最佳、最差与实际最差情形的$TH(\omega)$和$CT(\omega)$关系的曲线。Buzacott 和 Shanthikumar（1993）用仿真的方式，设定不同的系统参数，来测试这些式子。结果发现，在$c_e^2(j)$的值在 0.5 到 2 之间时，得到的近似值相当精确。

为了展示（10.1）~（10.4）式的用法，让我们回到第七章中 Penny Fab 的例子，Penny Fab 有四个工站，每个工站的平均加工时间t_e = 2 小时。用第七章的式子，我们可以画出代表最佳、最差以及实际最差情形的$TH(\omega)$和$CT(\omega)$的曲线。然而，假设我们关心提高工站的速度（即，创造一条不平衡的产线）或者相对于实际最差情形（PWC）减少变动性所带来的效果。因为实际最差情形等式只考虑所有工站c_e = 1 的平衡情形，在第七章的公式下没办法做，但是，我们可以用上述 MVA 算法来求解这个问题。

在变动性削减（相对于 PWC）的情况下考虑 Penny Fab，所以对于 j = 1, ..., 4，有$c_e(j)$ = 0.5。从$WIP_j(0)$ = 0 和= 0 开始，我们可以计算出对于 j = 1, ..., 4 有

$$CT_j(1)=\frac{t_e^2(j)}{2}\left[c_e^2(j)-1\right]TH(0)+\left[WIP_j(0)+1\right]t_e(j)=t_e(j)=2$$

既然所有的工站都是同样的，$CT(\omega)=4CT_j(\omega)$，因此$CT(1)$ = 8 小时，TH 为

$$TH(1)=\frac{1}{CT(1)}=\frac{1}{8}$$

每个工站的平均 WIP 是

$$WIP_j(1)=TH(1)CT_j(1)=(\tfrac{1}{8})(2)=\tfrac{1}{4}$$

已经计算出了 $\omega = 1$ 情形下的这些量度，我们下一步要转到 $\omega = 2$ 并且计算每个工站的周期时间如下

$$CT_j(2) = \frac{t_e^2(j)}{2}\left[c_e^2(j) - 1\right]TH(1) + \left[WIP_j(1) + 1\right]t_e(j)$$
$$= \frac{2^2}{2}(0.5^2 - 1)(\frac{1}{8}) + (\frac{1}{4} + 1)2 = 2.313$$

则 $CT(2) = 4CT_j(2) = 9.250$。持续按照这种方法进行，我们能够得到表 10.1 中的数据。

表 10.1 $c_e(j)$ 时 *Penny Fab* 的 *MVA* 计算结果

w	$TH(w)$	$CT(w)$	$CT_j(w)$	$WIP_j(w)$
1	0.125	8.000	2.000	0.250
2	0.216	9.250	2.313	0.500
3	0.280	10.703	2.676	0.750
4	0.325	12.318	3.080	1.000
5	0.356	14.052	3.513	1.250
6	0.348	15.865	3.966	1.500
7	0.395	17.731	4.433	1.750
8	0.408	19.631	4.908	2.000
9	0.418	21.555	5.389	2.250
10	0.426	23.495	5.874	2.500
11	0.432	25.446	6.362	2.750
12	0.438	27.406	6.852	3.000

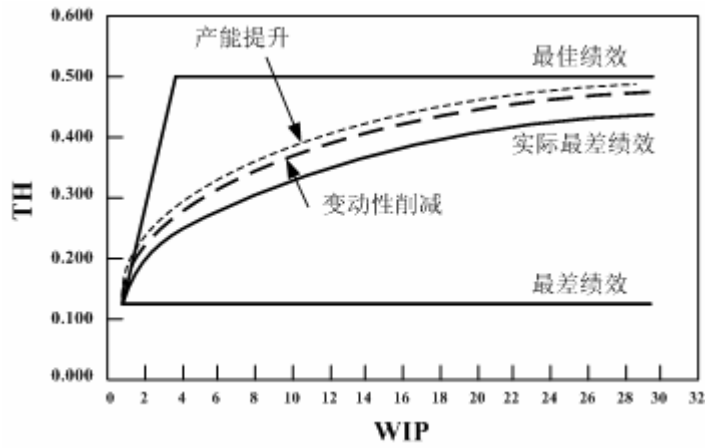

图 10.6 削减变动性和提升产能对 *Penny Fab* 绩效曲线的影响

采用同样的程序，我们还可以得到增加产能情形下的 $TH(\omega)$ 和 $CT(\omega)$，例如，将工站 1 和工站 2 的平均加工时间从 2 小时减少到 1 小时。我们这样做并且将两种情况下的结果都画在图 10.6 中，一种是表 10.1 中的减少变动性情形，另一种是增加产能情形，同时还有最

佳、最差和实际最差情形。请注意，这两种情形对于实际最差情形来说都有改善，因为它们使产线在给定的 WIP 水平下获得了更大的产出。在这个例子中，增加两个工站的速度获得的提高比减少多所有工站的变动性要大。当然，在实际中最终结果决定于具体的系统设定。这里给出的均值分析模型只是一个检查能力和变动性变化对于 CONWIP 产线影响效果的简单粗能力分析工具。（353|354）

现在我们既有推式系统的模型，也有拉式系统的模型，那么我们可以做一些比较来加深我们对拉式系统潜在好处的理解。从比较 CONWIP 和 MRP 开始，接着再比较 CONWIP 和看板。

10.5 CONWIP 和 MRP 的比较

推式系统和拉式系统的一个基本区别是：

推式系统控制产出，观察 WIP；而拉式系统控制 WIP，观察产出。

举个例子，在 MRP 中，我们建立主生产计划，决定计划投入量发，这些依次决定了什么将被触发到系统中。然而，WIP 水平是由产线中发生的情况来决定，这样就会随着时间上下浮动。在一个拉式系统中，WIP 直接由卡片数量的设置来控制。然而，同样由产线中发生情况决定的产出率会随着时间变化。（354|355）哪种方式更好呢？虽然它并不是一个简单的问题，我们却可以做一些观察。

10.5.1 可观测性（Observability）

首先，也是最基本的，我们意识到 WIP 水平可以直接观察得到，但是产出却不可以。因此，在拉式系统中，通过设定 WIP 水平进行控制相对比较简单。我们可以亲自数一下车间里的加工任务并保持对于 WIP 上限的依从性。相反，在推式系统中，设定触发速率必须参考能力来做。如果选择的速率过高，系统将被 WIP 阻塞；太低，就会因为产出不够而造成收入的损失。但是，能力估计不是那么简单的。从机器断供到作业员不可用的一系列扰动，都使得做出精确的估计相当困难。这一事实使得推式系统从本质上来讲要比拉式系统更难优化。

这里我们谈了一些控制理论领域里的一般准则。大体上，我们更倾向于控制稳健性参数（这样过失的破坏性不那么大）与观察敏感性参数（这样反馈会比较及时），而不是采取反过来的另一种方式。因为 WIP 是稳健的且可观察，而产出是敏感的，而且只能通过不可控的能力参数来控制。这是偏好拉式系统的强有力的论据。

10.5.2 效率（Efficiency）

另一个支持拉式系统的论据是它比推式系统更有效率。这里的更有效率是说，对于一个给定的要求的产出，拉式系统所需的 WIP 水平比推式系统要低。为了阐述为什么会这样，不妨考虑像图 10.4 中所示的那样的 WIP 水平固定为 ω 的 CONWIP 系统，然后我们观测产出 $T\tilde{H}(\omega)$。接着我们考虑像图 10.5 中所示的 MRP（纯推式）系统，由和 CONWIP 产线中同样的机器组成，投料速率为 $T\tilde{H}(\omega)$。根据物料守恒，MRP 系统中的产出速率将会等同于输入速率，为 $T\tilde{H}(\omega)$。这样，CONWIP 系统和 MRP 系统的产出就是相等的，效率问题的

关键就在于谁在达到这个产出的同时 WIP 更少。

让我们考虑一个具体的五台单机工站串联问题，每个工站的加工速率为 1 件/小时，加工时间服从指数分布。在这个简单的系统中，第七章中的关于实际最差情形的公式给出了 CONWIP 系统产出的表达式。作为 WIP 水平的函数，产出会降低到

$$\widetilde{TH}(\omega)=\frac{\omega}{\omega+W_0-1}r_b=\frac{\omega}{\omega+4} \qquad (10.5)$$

如果我们将推式系统的触发速率固定为产出，这里触发的间隔时间呈指数分布，每个站都可以看作一个独立的 M/M/1 队列，这样整体的 WIP 水平就是 5 倍的 M/M/1 队列的平均 WIP 水平，这个水平我们从第八章中知道是 $u/(1-u)$，这里 u 是利用率。因此，在这种情形下，加工时间就等于一而到达速率就等于 TH，u = TH。因此，系统的平均 WIP 水平就是（355|356）

$$\tilde{\omega}(TH)=5\left(\frac{u}{1-u}\right))=5\left(\frac{TH}{1-TH}\right) \qquad (10.6)$$

现在，我们假设 CONWIP 系统的 ω = 6。由（10.5）使，产出是 $\widetilde{TH}(\omega)$ = 0.6 件/小时。如果我们在（10.6）式中固定 TH = 0.6，我们看到 MRP 系统的 WIP 是 $\tilde{\omega}(0.6)$ = 7.5。因此，对于同样的产出水平，推式系统需要更多的 WIP。

请注意无论对于 ω 的选择是怎样，推式系统的 WIP 水平都会高于 ω。为了展示这一点，我们在（10.6）式中，设 $TH=\omega/(\omega+4)$ 则

$$\tilde{\omega}\left(\frac{\omega}{\omega+4}\right)=\frac{5[\omega/(\omega+4)]}{1-\omega/(\omega+4)}=\frac{5\omega}{4}$$

所以，在这个例子中，对于任意一个产出水平来说，推式系统中的 WIP 水平会比拉式系统高 25%。

虽然与推式系统相比 CONWIP 的 WIP 的巨大增加量还决定于具体的产线参数，这种定性的效果却具有一般性，如我们在接下来的定律中陈述的那样

定律（CONWIP 效率）：对于给定的产出，推式系统相对于同等 *CONWIP* 系统平均 *WIP* 数量更多。

这条定量有一个直接推论，当 CONWIP 系统和 MRP 系统的产出相同时，里特定律和 MRP 系统中 WIP 更多这一事实暗示出：

推论：对于给定的产出，推式系统相对于同等 *CONWIP* 系统平均周期时间更长。

10.5.3 变动性（Variability）

我们介绍了推式系统相对于同等 CONWIP 系统平均可变周期时间更长，原因如下。从定义上来说，CONWIP 系统的 WIP 水平固定为 ω。这一事实提出不同站之间的 WIP 水平是负相关（*negative correlation*）的。举个例子，如果我们知道站 1 有 ω 件任务，那我们就能

完全确定其它站一件任务也没有。在这种情形下，对于工站 1 的 WIP 水平的了解就给我们提供了关于其他站 WIP 水平的完全信息。然而，即使我们仅仅知道工站 1（如，在 10 工站产线中）有 $\omega/2$ 件任务，我们也能获得其他站的一些信息。例如，其他任何一个站有其余所有 $\omega/2$ 件任务是不大可能的，这种 WIP 水平之间的负相关能够弱化周期时间的波动。

相反，在推式系统中，每一个工站之间的WIP水平是相互独立的（*independent*），[4]工站 1 的高WIP水平不会告诉我们任何关于其它工站WIP水平的信息。因此，若干个工站的WIP水平同时很高（或很低）是可能的。因为周期时间与WIP水平直接相关，这就是说极端（高或低）的周期时间是可能的。这样的结果就是推式系统相对于同等拉式系统周期时间的变动性更大。

周期时间变动性的增加意味着我们必须报出更长的提前期才能达到相同的客户服务水平。这是因为为了达到相同的服务水平，我们的周期时间应该是平均周期时间加上标准差的某一倍数（倍数由所要求的服务水平决定）。例如，图 10.7 展示了两个平均周期时间为 10 天的系统。（356|357）然而，相对于系统 1，系统 2 周期时间的标准差要高出很多。为了达到 90%的服务水平，系统 1 必须报出 14 天的提前期，而系统 2 必须报出 23 天。推式系统变动性的增加使得周期时间的标准差更大。请注意，这是接着上面的事实，对于给定的产出，推式系统相对于同等 CONWIP 系统平均周期时间更长。因此，对于相同的产出和客户服务水平，推式系统的提前期更长有两个原因：平均周期时间更长,周期时间的标准差更大。

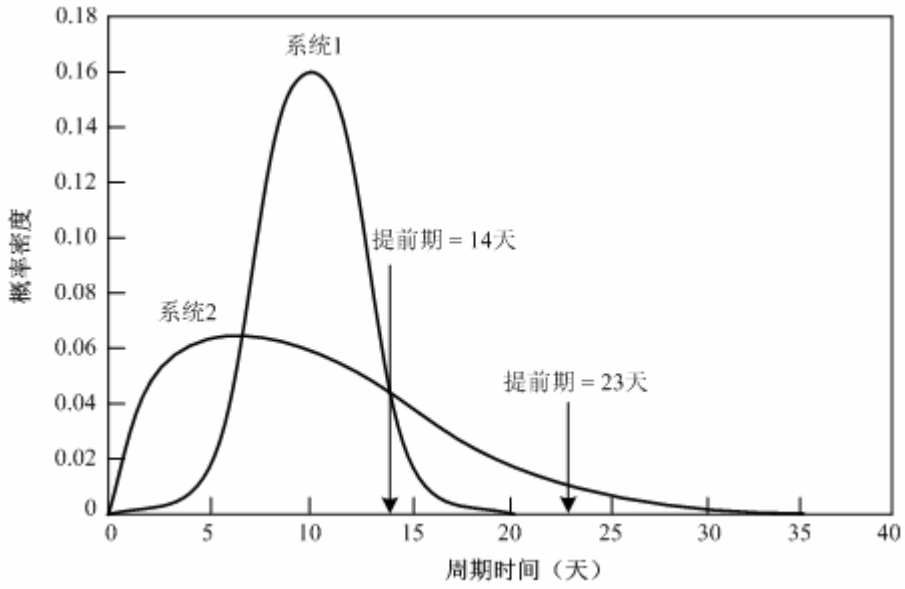

图 10.7 周期时间变动性对客户提前期的影响

10.5.4 稳健性（Robustness）

CONWIP 系统相对于纯推式系统最重要的优势不是 WIP（与平均周期时间）的减少，也不是周期时间方差的减少，虽然它们同样很重要。拉式系统最重要的优势是它们的稳健性，像我们在接下来所陈述的：

定律（CONWIP 稳健性）：*CONWIP 系统遇到 WIP 水平错误时比纯推式系统遇到触发速率错误时更稳健。*

[4] 这个观察结果只在加工时间服从指数分布时才严格成立，但在推式系统比拉式系统中更接近实际，甚至当加工时间不服从指数分布时也是如此。

为了使这条定律的含义更清晰，假设存在一个简单利润公式如下

$$\text{利润} = p*\text{TH} - h*\omega \qquad (10.7)$$

p 是每一件任务的边际利润，TH 是产出，h 是每一单位 WIP 的成本（包括增加的周期时间成本，质量降低成本，等等），ω 是平均 WIP 水平。在 CONWIP 系统中，产出是 WIP 的函数，也就是 $\widetilde{TH}(\omega)$，我们会选择的 ω 值从而使利润最大化。在推式系统中，平均 WIP 是触发速率的函数，$\widetilde{\omega}(TH)$，我们通过选择产出的值使利润最大化。

从先前的定律中可以很清楚地知道，CONWIP 系统的最优利润比推式系统的高（因为对于选定的产出水平 CONWIP 系统的 WIP 较低）。然而，CONWIP 稳健性定律关注的是 CONWIP 系统中 ω 的选择未达最优水平的情况或者推式系统中产出的选择未达最优水平的情况。（357|358）因为 WIP 和产出使用不同的单位来衡量，我们用错误百分比的形式来衡量次最优性。在我们前面提到的例子中进行这项工作，5 台机器，加工时间为 1 小时且服从指数分布，成本系数是 $p = 100$，$h = 1$，如图 10.8 所示。

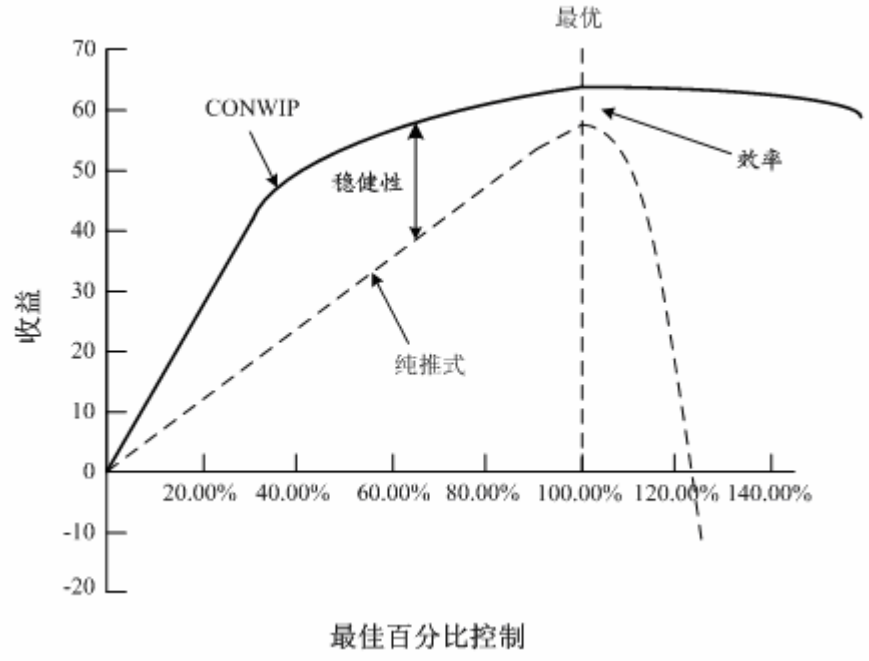

图 10.8 *CONWIP* 和纯推式系统的相对稳健性

我们看到CONWIP系统的最优WIP水平为 16 件，产生的利润为 63.30$/小时。在推式系统中，最佳产出被证明是 0.776 件/小时，产生的利润为 60.30$/小时。因此，像预料的那样，CONWIP系统的最优利润水平比推式系统的最优利润水平要高一点（5%左右）。然而，更重要的是这样一个事实，CONWIP系统中，利润函数在WIP水平在为最优水平的 40%到 160%之间非常平缓。相反，推式系统的利润函数在选择的水平低于最优时逐步下降，当触发速率高于最优水平很小一点时急剧下降。事实上，当触发速率达到最优水平的 120%时利润就是负的，而CONWIP系统直到WIP水平达到最优水平的 600%时仍然是正的。[5]

[5] 尽管我们仅仅给出一个例子,这种稳健性结果相当通用并且不依赖于这里所做的假设。细节见 Spearman 和 Zazanis（1992）。

根据我们早些时候提出的可观察性问题，这些观察资料非常重要。像我们强调过的，推式系统的最优触发速率必须根据系统的真实产能来设定，而真实产能没有办法直接观察得到。人类天性中的乐观，混合一些难以理解的以尽可能多地获得系统的产出从而使收入最大化的愿望，提供了设定过高触发速率的强大动力。像图 10.8 中所显示的，这正是那种代价最大的错误。

另一方面，CONWIP 系统是通过设定易于观察的 WIP 水平参数来控制的。这一特性再加上最优位置附近的平缓的利润曲线，意味着达到最优利润水平比推式系统要来得容易。所有这些的实践结果就是，CONWIP 系统和纯推式系统绩效上的不同可能比我们用等式 10.5 和 10.6 所做的公平比较得出的结果要大得多。因此，稳健性的增加可能是采用拉式系统最显著的原因，比如用 CONWIP 系统代替 MRP 系统。（358|359）

10.6 CONWIP 和 Kaban 的比较

像图 10.5 所示的那样，在产线中的触发是由内部需求引起的意义上，CONWIP 系统和看板系统都是拉式系统。因为两个系统都建立了 WIP 上限，它们相对于 MRP 显示出相似的绩效优势。具体的，与 MRP 系统相比，CONWIP 系统和看板系统都是在 WIP 较少的情况下达到目标产出水平，同时显示了较低的周期时间变动性。进一步地说，因为这两个系统都是通过设定 WIP 水平来控制，而且我们知道 WIP 比触发速率在控制上具有更强的稳健性，所以它们比纯推式系统更易于管理。然而，在 CONWIP 和看板之间也有一些重要的不同之处。

10.6.1 卡片计数问题（Card Count Issues）

最明显的不同是看板比 CONWIP 需要设定的参数更多。在一个单卡片的看板系统，使用者必须为每一个工站建立单卡片计数机制（在双卡片系统里，有两倍的卡片计数需要设置）。相反，在 CONWIP 系统中，只有一个单卡片计数需要设定。因为得到合适的卡片计数需要将分析与持续调整相结合才会得到，这一事实意味着 CONWIP 从本质上更易于控制。因为这个原因，我们将 CONWIP 看作标准，评估其他系统。如果将要采取比 CONWIP 更复杂的拉式系统，那么系统的绩效必须能够证明增加的复杂性是合理的。在第三篇我们会检查一些情况，在这些情况中，更复杂的系统确实是有其存在的价值的。但是，在这一章节中，我们将继续将我们的范围限定在有一系列工站前后排列的简单产线，从而使我们能够在 CONWIP 和看板之间做一些简单的比较。

第二个CONWIP和看板之间的重要不同之处在图 10.5 中不是那么明显，是说卡片是典型的*料号有别*（*part number-specific*），而CONWIP系统中*产线有别*（*line-specific*）。即，看板系统中的卡片确定了用于授权生产的部件。在一个多生产的环境中，这是必要的，因为一个工站必须知道在它的流出库存点必须补充哪一类型的库存。另一方面，CONWIP系统中，卡片并不确定具体的部件数量。相反地，它们来到产线的前端与**延迟清单（backlog）**相匹配，这样就为部件引入产线设定了顺序。这种延迟清单，或者说顺序，必须由CONWIP回路之外的类似于MRP系统中主生产计划方式的模块来生成[6]。因此，根据存货每一个特定的卡片返回到CONWIP产线的前端的时间，就可能授权投放一个不同的部件到产线中。

这一明显的不同之处不是因为工作投放机制的不同，而是因为它对于两个系统的意义。

[6] 开发延迟清单（backlog）和 MPS 的首要区别是，延迟清单是时间与加工任务无关的*排序*（*sequence*），而 MPS 是表明需求时间的*排程*（*schedule*）。我们将在第十五章中讨论使用顺序与计划的区别以及它们的相对优势。

从纯粹形式上看，看板系统必须包含产线上每一个活性料号的 WIP 标准容器。如果不是的话，下游工站就可能产生需求，而这个需求上游工站却无法满足。像我们在实践中所看到的，如果产线生产 40,000 种不同的料号，一个丰田看板系统可能会被 WIP 淹没。问题是 40,000 种料号中的大部分当它是活性部件时仅仅是偶尔生产。（359|360）因此，看板系统不必要的为几个月都不会生产的零件保持了 WIP。但是，如果这些低需求的零件在车间没有库存，那么，线末的需求就会一路生成无法满足的需求而回到产线的起始端。从产线的起始开始一件任务，一直沿产线进行操作的时间将会比储备正常的对于末端需求的反应时间长很多，并且 JIT 协议也就中止了。

CONWIP 系统由于使用了线具体的卡片和作业延迟清单，就没有这种问题。如果 CONWIP 产线中的卡片计数为 ω，那么线中最多会有 ω 件任务。实际上 ω 总是比 40,000 小很多。如果一个零件 6 个月内都没有被需要，那么它将不会在工作订单中出现，因此也不会被触发到线中。当对于低数量零件的需求确实出现的时候，工作订单将会在合适的提前期之下将其投放到产线中以容纳产线中所需要的生产时间。因此，“准时化”的绩效得到了保持，甚至是对一两件的订单。

然而，我们必须指出，CONWIP 和看板的一个基本不同之处是纯看板系统的提前期是零，而 CONWIP 系统的提前期是一个较小的值。这是 CONWIP 系统保持柔性的代价。看板是一个纯备货生产系统，在这个系统中，需要时零件应该在流出库存点。另一方面，CONWIP 通过保持低 WIP 水平从而保持短周期时间。如果周期时间足够的短，就不需要改变零件的顺序，所以，因增加的周期时间而获得的柔性的增加是值得的。

10.6.2 产品组合问题

看板专家清楚地意识到看板不可能在所有的生产环境中都发挥作用。Hall（1983）指出看板只在**重复性制造（repetitive manufacturing）**环境中是可应用的。他所说的重复性制造是指，系统中的物料沿着固定的路径以固定的速率流动。数量上或者产品组合上的大的变化，至少是当零件被当作个体来看，都破坏了这种流，因此严重破坏了看板。CONWIP，虽然同样需要相对稳定的数量（即，平衡策略的 MPS（a level MPS）），但因为生成工作订单这一流程带来的计划能力，对于产品组合的变化却更稳健。

变化了的产品组合会带来更多的微妙变化，而不仅仅是提升看板系统所需要的整体 WIP 水平。如果不同零件的复杂性发生变化（即，部件需要机器上不同次数的处理），那么产线瓶颈会因产品组合不同而变化。例如，请考虑图 10.9 中所示的五个工站的产线。产品 A 需要在机器 2 和 3 上分别进行 3 和 2.5 小时的加工，在其它机器上各进行 1 小时的加工，产品 B 需要在机器 3 和 4 上分别进行 2.5 和 3 小时的加工，在其它机器上各进行 1 个时的加工。因此，如果我们正在加工产品 A，机器 2 是瓶颈。如果我们加工的是产品 B，机器 4 是瓶颈。然而，对于一个包含了 25%到 75%产品 A 的产品组合，机器 3 是瓶颈。（360|361）

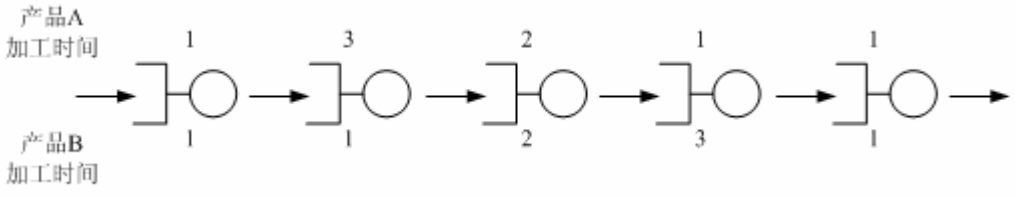

图 10.9 瓶颈漂移的系统

为了明白这一点，请考虑一个 50-50 的产品 A 和 B 的组合，机器 2、3、4 上的平均加工时间是：

机器 2 的平均加工时间 = 0.5（3）+0.5（1）= 2 小时
机器 3 的平均加工时间 = 0.5（2.5）+0.5（2.5）= 2.5 小时
机器 4 的平均加工时间 = 0.5（1）+0.5（3）= 2 小时

只有当产品 A 超过 75%时，机器 2 的平均加工时间才会超过两个半小时。同样的，只有当产品 B 超过 75%时（也就是说，产品 A 小于 25%），机器 4 的平均时间才会超过两个半小时。

在一个理想的看板环境中，我们需要设定 A 和 B 的顺序来达到一个固定的组合；例如，对于一个 50-50 的产品，我们采用的顺序为 A-B-A-B- A-B-……在一个非理想环境中，组合需求不是固定的（例如，需求是季节性的，预测是不稳定的），统一的顺序将没什么实际意义。然而，如果我们让组合根据需求来变化，这将会给看板系统中的卡片计数带来一些问题。我们为了让瓶颈站免于不足或堵塞，一般想在瓶颈站前后放置更多的生产卡片。但是哪个是瓶颈呢——机器 2、3 还是 4？当然，答案决定于我们所要加工的产品组合。这意味着最优卡片计数安置是产品组合的函数。因此，为了在达到高产出的同时有低 WIP 水平，我们必须动态地随时间改变卡片计数。既然我们已经论证了在看板系统中设定卡片计数要花很大的代价，那么这就会成为一个很难的任务。

然而，CONWIP只有一个单卡片计数。因此，只要需要的速率保持相对稳定，就没有必要根据产品组合的变化来变更卡片计数。进一步的，WIP会自然地在瓶颈前累积，刚好是我们需要的地方。[7]在我们的例子中，当我们加工一个偏重于A的产品组合时，机器 2 会是最慢的，而且也得因此累积最大的队列。当组合变得偏重于产品B时，最大的队列会转到机器 4。令人愉快的是，这些都是在没有我们的干预下发生的，因为自然地治理了瓶颈的行为。再一次的，CONWIP系统从基础上来说比看板系统更易于管理。

10.6.3 人的问题

最终，我们在两个以人为导向的观察下完成 CONWIP 和看板的比较。第一，看板系统在每一个站采取拉式的事实给系统造成了一定程度的压力。看板系统中的作业员如果只有原材料但没有生产卡片就无法开始工作。生产卡片到达后，他们必须尽可能快地补充系统中闲置的空间，这样才能阻止线中的某些地方发生不足。正如 Kleina（1989）所指出的，这种类型的节奏压力是作业员压力的主要来源。

相反地，CONWIP 在除了第一个之外的所有站都表现的像推式系统。当中游机器的作业员接收到原材料，他们自动开始在上面工作。因此，作业员就可以提前于原材料的可获得性允许的最大程度工作，也就能减轻一些节奏压力。当然，在 CONWIP 产线的第一个站，作业员只有在生产卡片授权后才能工作，所以实际上他们的工作条件与看板产线中第一站的作业员相同。如果我们要建立 WIP 上限的话，这是不可避免的。（361|362）这样，CONWIP 产线同样带来了一定程度的节奏压力，但是比看板要轻。

我们第二个以人为导向的观察结果是看板产线中在每一个站采取拉式的行动会促进邻近工站作业员建立更为亲密的关系。因为在看板系统中作业员必须拉动所需要的零件，他们将会与上游机器的作业员进行沟通。这就提供了一个检查产品质量问题的机会，同时还提供了识别和讨论任何一个有关生产速率的问题的机会。我们常常听到这一好处被用来作为使用纯看板系统的动机。

虽然我们承认因邻近工站作业员的联系所带来的沟通及学习上的好处是巨大的，但我们的疑问是看板的拉式纪律对于获得这种好处来说是必要的吗？无论两个站之间是不是采用了看板机制，上下游站之间的零件转换都一定会发生。为了阻止坏零件的转换，“买-卖”协

[7] 注意到阻塞不会发生在 CONWIP 体系，因为没有工站之间的卡片计数来限制完成的加工任务转运到下一个工站。

议可以在没有看板的情况下使用，协议规定了下游作业员可以拒绝接受没有达到质量指标的零件。为了促进作业员们协调流水的相关问题，必须培养它们的全线角度。CONWIP 产线需要将焦点放在与合适的生产速率相关的问题，而不是看板系统的保持流出库存点充满问题。如果作业员需要在工站之间移动来促进这一点，那么是好的。有许多工作分配结构化的方式来实现固定产出率的这一总体目标。我们的观点仅仅是虽然看板的拉式机制是一种促进作业员之间协调的方式，但它不是唯一的。从物流和简单化上来考虑，青睐 CONWIP 并且发展这些其他的学习动力而不是执行刻板协议的看板系统是值得的。

10.7 结论

在这一章中，我们指出了以下的基本观点：

1. 推式系统计划（*schedule*）任务的投放，而拉式系统根据系统的状态授权（*authorize*）任务的投放。

2. 拉式的“魔法”是它们建立了 WIP 上限，从而能够避免生产不必要且不能显著提高产出的 WIP。拉是一种达到目的的手段，结果是拉式系统减少了平均 WIP 和周期时间，降低了周期时间的变动性，创造了提高质量的压力，促进了更有效的错误检测（通过削减 WIP），增加了适应变化的柔性。

3. 建立 WIP 上限最简单的机制是 CONWIP（常量在制品，*constant work in process*），其中的 WIP 水平通过协同物料投入与离开产线来保持恒定。

4. 相对于纯推式体系，CONWIP 显示了以下优势：

- WIP 水平是直接可观测的，而纯推式系统中的投料速率必须根据产能（不可观测）来设定。
- 达到相同的产出平均需要的 WIP 较少。
- 对于控制性参数的错误有更强的稳健性。
- 当良性情况允许，它提前于计划开始任务。（362|363）

5. 相对于看板体系，CONWIP 显示了以下优势：

- 从 CONWIP 系统只需设置一个单卡片计数而不是每个工站都有单卡片计数的意义上讲，它更简单。
- 因为采用了产线有别的卡片和作业积压单，CONWIP 能接受变化的产品组合。
- 由于 WIP 能够在最慢的机器前累积的自然倾向，它能接受漂移的瓶颈（视组合而定）。
- 因为采用了更柔性化的节奏协议，它给作业员带来的压力较小。

虽然这些观察是建立在高度简化的纯推式、纯看板及纯 CONWIP 体系的基础上，但它们包含了工厂物理学的基本见解。在第三篇中，我们将转向把这些见解应用于实践，应用到繁杂的真实世界环境中。