文章编号: 1002-0411(2000)05-0425-06

半导体制造系统仿真调度中的优化方法

卫军胡　管晓宏　孙国基
(西安交通大学系统工程研究所, 机械制造系统工程国家重点实验室　西安　710049)

摘　要: 基于仿真的调度方法通常需要进行大量的仿真或者采用好的规则以优化调度结果. 本文建立了以减小平均在制品为优化目标的半导体制造系统的调度模型, 对模型进行分解和简化. 把结论作为一个调度规则直接应用于仿真调度方法. 由于充分利用了系统全局的状态信息, 可以有效地减少仿真的次数, 提高了仿真调度的优化能力. *

关键词: 半导体制造, 仿真调度, 优化调度
中图分类号: TP391.9　　文献标识码: B

1　引言

基于离散事件系统仿真技术的调度方法是仿真技术的最新应用. 其基本原理是, 在调度规则的引导下, 在制造系统的仿真模型上试探性地经历整个加工过程, 记录该过程中系统的状态变化及导致系统状态改变的事件, 产生调度方案并统计性能数据[1]. 这是一种实验性和试探性的方法, 不会出现无解的现象, 而且产生的每一个调度都是调度问题的可行解. 通常采用 2 种方法对调度结果进行优化: (1) 对同一问题分别给定不同的参数, 多次进行仿真调度, 得到多个调度方案, 通过比较目标函数值, 从中选择一个相对好的解. (2) 针对具体的问题, 采用好的调度规则, 直接得到好的调度, 从而减少仿真调度的次数. 所谓好的规则, 是指经过理论推导、仿真或实践证明有利于调度目标的调度规则[2]. 由于通常利用系统当前的和局部的状态进行决策, 因而优化的结果是局部的, 不能保证每个调度都是最优的.

本文建立了以减小平均在制品库存(WIP)为目标的半导体制造系统(SMS)的可分解的调度模型, 把复杂的调度问题分解为一系列子问题, 并进一步简化. 把结果作为一种调度规则直接应用于 SMS 的仿真调度中. 仿真调度模型是实际系统的精确模型, 与实际系统有一一对应的关系, 用于产生调度指令并统计关键的性能参数. 当遇到设备分配和任务选择等决策问题时, 根据系统当前的状态产生调度子问题, 得到某个子区间内的最优解. 在一定条件下, 这些子问题的解综合在一起, 形成原调度问题的解. 这种两种模型不断交互的调度方法, 可以最大限度地利用系统整体的信息, 提高仿真调度的优化能力, 减少优化所需的仿真次数.

2　优化调度模型

2.1　模型定义

设一个系统由集合为 S 的设备组组成, 设备组 s 包含 M_s 个相同的设备, $s \in S$. 有 N 种不同的产品, 产品 i 需要经过 N_i 个操作, 操作 O_{ij} 的加工时间为 o_{ij}, $i = 1, 2, \cdots, N$. $j = 1, 2, \cdots, N_i$. 通常把若干晶片放在同一容器内, 形成一个晶片组(A Lot of Wafers), 作为一个整体

* 收稿日期: 1999-09-14
基金项目: 国家杰出青年基金(6970025)和西安交通大学科研基金资助

在系统中一起加工和运送.

每个设备组都有若干缓冲区(Buffer),用于存放等待该设备组的各类晶片. 所有等待操作 $O_{ij}(i=1,2,\cdots,N,\ j=1,2,\cdots,N_i)$ 的晶片组存放在缓冲区 b_{ij} 中,因此共有 $c=\sum_{i=0}^{N}N_i$ 个缓冲区. 按操作顺序对这些缓冲区统一编号,形成缓冲区序列, $B_1, B_2, \cdots, B_c$. 其中 $B_1=b_{1,1}, B_2=b_{1,2}, \cdots, B_c=b_{N,N_N}$. 用 C 表示所有缓冲区的下标集合, $C=\{1,2,\cdots,c\}$. 相应地,用 s_k 和 u_k 分别表示加工 B_k 的设备及其加工速度,那么, $u_1=1/o_{1,1}, u_2=1/o_{1,2}, \cdots, u_c=1/o_{N,N_N}$. 由于每个设备组可以进行多个不同的操作,所以一个设备组可能具有多个缓冲区,用 C_s 表示设备组 s 上的缓冲区下标集合, $C_s=\{k\mid B_k$ 属于设备组 $s,\ k\in C\}$. 晶片组可以从系统外部投入到系统,用 r_i 表示晶片组进入缓冲区 B_i 的速度,定义为单位时间投入的晶片组数量.

用 $Q_k(t)$ 表示 t 时刻 B_k 中晶片组的数量,包括等待加工和正在加工的晶片,那么 t 时刻系统中在制品(WIP-Work In Process)总量为 $\sum_{k\in C}Q_k(t)$. 用 $c\times c$ 矩阵 P 建立缓冲区之间晶片的转换关系,元素 p_{jk} 表示 B_j 中的晶片在经过 s_j 加工后是否进入 B_k,如果存在这种转换关系则 $p_{jk}=1$,否则 $p_{jk}=0$. 用 $m_k(t)$ 表示 t 时刻设备组 s_k 分配给 B_k 的设备数量, $T_k(0,t)$ 表示在 $[0,t]$ 内设备组 s_k 以 u_k 的速度加工 B_k 的累计时间(折算为在 1 台设备上的累计加工时间),

$$T_k(0,t)=\int_0^t m_k(\tau)\,\mathrm{d}\tau \tag{1}$$

那么截止 t 时刻, s_k 累计加工 B_k 的数量为 $u_kT_k(0,t)$. 有下面的动态平衡方程

$$Q_k(t)=Q_k(0)+r_kt+\sum_{j\in C}u_jT_j(0,t)p_{jk}-u_kT_k(0,t) \tag{2}$$

其中,等式右边第 1 项为 B_k 的初始数量,第 2 项是在 $[0,t]$ 内从系统外部投入到 B_k 的晶片组的数量,第 3 项是上一个操作完成后进入 B_k 的数量,而第 4 项是 B_k 中的晶片在 s_k 上完成加工而离开 B_k 的数量. 把式(1)写成向量的形式,

$$T(0,t)=\int_0^t m(\tau)\,\mathrm{d}\tau \tag{3}$$

其中 $T(0,t)$ 和 m 分别是 $T_k(0,t)$ 和 m_k 组成的 c 维向量. 对向量的积分是指对每个元素进行积分. 可以证明,任意 $t_1\in[0,t]$ 把区间 $[0,t]$ 分为两个子区间 $[0,t_1]$ 和 $[t_1,t]$,都有 $T(0,t)=T(0,t_1)+T(t_1,t)$. 可见 $T(0,t)$ 具有可加性,即设备在整个区间上分配给某个缓冲区的加工时间等于在各子区间上分配给该缓冲区的加工时间之和.

把(2)式写成向量的形式,

$$Q(t)=Q(0)+Rt-(I-P^{\mathrm{T}})UT(0,t) \tag{4}$$

其中, $Q(t)$ 和 R 分别是 $Q_k(t)$ 和 r_k 组成的 c 维向量. I 是单位矩阵, P^{T} 是 P 的转置, U 是 $c\times c$ 矩阵, $u_{kk}=u_k$,而 $u_{jk}=0,\ j\neq k,\ j,\ k\in C$.

如果近似地认为晶片在设备组之间的移动是连续的,那么 SMS 就可以近似地用流体网络模型(FNM-Fluid Network Model)来表示,方程(4)与 FNM 流体平衡方程具有相同的形式[5,6]. 根据文献[5],过程 $\{Q(t)\}$ 具有 Markov 特性,即在任意时刻 $t>t_1$, $Q(t)$ 的变化只依靠 t_1 时刻的状态 $Q(t_1)$,而与 t_1 以前的状态无关.

由式(3)和(4)可以看出,通过对设备的分配,即改变 $m(t)$ 可以改变 $T(0,t)$,从而控制 $Q(t)$,使调度目标最优化. 在决定 $m(t)$ 时,必须保证在任意时刻 $t\geqslant 0$ 满足以下约束:

$$Q(t)\geqslant 0 \tag{5}$$

$$\sum_{k\in C_s} m_k(t) \leqslant M_s, \quad \forall s \in S \tag{6}$$

$$m_k(t) \text{ 为非负整数}, \quad \forall k \in C \tag{7}$$

约束(5)要求 B_k 中的晶片数量不能为负数, (6)是设备组的容量约束, 即设备组分配给每个缓冲区的设备总和不能超过其包括的设备数.

2.2　优化调度问题

在半导体制造环境下, 控制库存是普遍关注的问题, 因为大的库存将造成资金积压、保管费用增加以及系统拥挤和堵塞, 而且晶片长时间暴露在空气中可能导致失效或质量下降. 因此本文以减小调度周期内的平均库存及其费用作为优化目标.

设调度周期为 H, 用 h_k 表示 B_k 在单位时间单位数量的库存费用, h 是由 h_k 组成的 c 维费用向量, 那么在 t 时刻, 平均库存费用为

$$J(t) = \frac{1}{t}\int_0^t hQ(\tau)\,\mathrm{d}\tau \tag{8}$$

$J(H)$ 即为整个调度周期内的平均库存费用. 当所有 $h_k = 1$ 时, $J(t)$ 为 t 时刻的平均库存. SMS 优化调度问题可以表示为整数规划的形式(IP):

$$\min_{m(t)} J(t), \forall\ t \geqslant 0$$

$$\text{s. t. } (5) - (7)$$

即求解以 $Q(0)$ 为初始状态的最优设备分配方案 $m(t)$. 简记为 $J(t)\big|_{Q(0)}^{m(t)}$ 或者 $J(t)\big|_{Q(0)}^{T(0,t)}$.

2.3　分解算法

上述 IP 问题是一个要求极为严格的优化问题, 在每个时刻都求解上述整数规划问题在实际生产控制中是不现实的. 一种简单的方法是采用按时间分解的方法, 把调度周期划分为若干个子区间, 求解每个区间内的最优解.

定理 1　设调度周期为 H, 任意 $t_1 \in [0, H]$ 把区间 $[0, H]$ 分为 2 个子区间 $[0, t_1]$ 和 $[t_1, H]$. $m^1(t)$ 是区间 $[0, t_1]$ 内以 $Q(0)$ 为初始状态的最优分配方法, 其对应的最优累计加工时间为 $T^1(0, t)$, 即

$$J^1(t)\big|_{Q(0)}^{m^1(t)} = \frac{1}{t}\int_0^t hQ(\tau)\,\mathrm{d}\tau = \min, \quad \forall\ t \in [0, t_1]$$

$m^2(t)$ 是区间 $[t_1, H]$ 内以 $Q(t_1)$ 为初始状态的最优解, 其对应的最优累计加工时间为 $T^2(t_1, t)$, 即

$$J^2(t)\big|_{Q(t_1)}^{m^2(t)} = \frac{1}{t - t_1}\int_0^t hQ(\tau)\,\mathrm{d}\tau = \min, \quad \forall\ t \in [t_1, H]$$

其中 $Q(t_1)$ 是在 $m^1(t)$ 的作用下 t_1 时刻的系统状态. 如果整个调度周期的最优解存在且唯一, 那么 $[0, H]$ 上设备的最优分配方案 $m^*(t)$ 为

$$m^*(t) = \begin{cases} m^1(t), & 0 \leqslant t < t_1 \\ m^2(t), & t \geqslant t_1 \end{cases} \tag{9}$$

其对应的最优累计加工时间为

$$T^*(0, t) = \begin{cases} T^1(0, t) & 0 \leqslant t < t_1 \\ T^1(0, t_1) + T^2(t_1, t) & t \geqslant t_1 \end{cases}$$

由于篇幅所限, 这里不作证明. 这一结论可以进一步推广: $[0, H]$ 被任意 $t_1, t_2, \cdots, t_n \in$

$[0, H]$分为 $n+1$ 个子区间$[0, t_1)$, $[t_1, t_2)$, …, $[t_n, H]$, 其中 $t_1 < t_2 < \cdots < t_n$. $m^i(t)$, $i = 1, 2, \cdots, n$ 是每个区间上以 $Q(t_{i-1})$ 为初始状态的最优解, 其中 $t_0 = 0$, $t_{n+1} = H$, $Q(t_{i-1})$ 是上一个子区间结束时的系统状态. 那么在整个调度周期内的最优解可以按照(9)的形式构造. 根据这一结论, 原优化调度问题(IP)可以分解为一系列子区间上的调度子问题. 下面对子问题进一步简化以便于求解.

2.4 子问题的简化

在半导体制造环境下, $m(t)$ 不随时间连续变化, 而是在一些离散的时间点上改变数值. 向量 $m(t)$ 改变了数值是指该向量至少有一个元素改变了数值. 设 $m(t)$ 分别在 t_1, t_2, …, $t_n \in [0, H]$时刻改变了数值, $t_1 < t_2 < \cdots < t_n$, 那么它分别在 $n+1$ 个区间$[0, t_1)$, $[t_1, t_2)$, …, $[t_n, H]$上保持不变, 记为 m^i, $i = 1, 2, \cdots, n+1$. 其中 $t_0 = 0$, $t_{n+1} = H$. 根据定理1, 可以把整个区间$[0, H]$上的调度问题分解为 $n+1$ 个子区间的调度问题(IP^i), $i = 1, 2, \cdots, n+1$.

对于 IP^i, 其调度子区间为$[t_{i-1}, t_i)$, 初始状态为 $Q(t_{i-1})$. 用 t 表示从 t_{i-1} 时刻开始所经历的时间, $0 \leqslant t < t_i - t_{i-1}$. 根据公式(3), $T^i(0, t) = \int_0^t m^i \mathrm{d}\tau = m^i t$. 式(4)的平衡方程可以改写为

$$Q(t) = Q(t_{i-1}) + Rt - (I - P^{\mathrm{T}}) U m^i t$$

代入目标函数

$$J(t) = \frac{1}{t}\int_0^t hQ(\tau)\,\mathrm{d}\tau = h\{Q(t_{i-1}) + \frac{R}{2}t - \frac{(I - P^{\mathrm{T}}) U m^i}{2} t\}$$

因此子问题 IP^i 可以表示为$\min\limits_{m^i} J(t)$, $0 \leqslant \forall\, t < t_i - t_{i-1}$

由于目标函数中的前2项与决策变量 m^i 没有关系, 可以把 IP^i 问题等价地变为$\max\limits_{m^i} h(I - P^{\mathrm{T}}) U m^i t$. 又由于要求在$\forall\, t \geqslant 0$ 时刻都取最优值, 只要保证目标函数中 t 的系数最大即可, 因此, IP^i 可以进一步等价地改写为如下的形式:

$$\max_{m^i} h(I - P^{\mathrm{T}}) U m^i$$

$$\text{s. t.} \quad (5) - (7)$$

采用文献[5]的方法对约束(5)的形式进行变换: 把缓冲区按子区间的初始状态 $Q(t_{i-1})$ 分为3个集合: $q^0 = \{k \mid Q_k(t_{i-1}) = 0, k \in C\}$, $q_1^+ = \{k \mid Q_k(t_{i-1}) > 0$ 且 $Q_k'(t) < 0, k \in C\}$ 和 $q_2^+ = \{k \mid Q_k(t_{i-1}) > 0$ 且 $Q_k'(t) \geqslant 0, k \in C\}$. Q_k' 为 $Q_k(t)$ 的导数, $Q'(t) = R - (I - P^{\mathrm{T}}) U m^i$, 在子区间内为常数. 对于 $k \in q^0$, 只有 $Q_k'(t) \geqslant 0$ 才能满足(5), 因此, 约束(5)的等价形式是 $Q_k'(t) \geqslant 0$. 对于 $k \in q_2^+$, 由于其初始库存大于0, 而且 $Q_k(t)$ 的导数大于等于0, 因此始终满足(5), 可以忽略. 而对于 $k \in q_1^+$, 由于其初始库存大于0, 而且 $Q_k^*(t) < 0$, 因此其库存水平 $Q_k(t)$ 将逐渐减小, 最终达到0. 随着 t 的增加, 当$\forall\, k \in q_1^+$ 的库存水平由正数变为0时, 约束(5)将不再满足. 如果以此刻作为子区间的上限 t_i, 那么在整个子区间$[t_{i-1}, t_i)$内, 约束(5)一直得到满足. t_i 既是该区间的上限, 又成为下一个子区间的初始时刻. 用这种方法可以确定每个子问题 IP^i 子区间的上下限 t_1, t_2, …, $t_n \in [0, H]$.

因此, IP^i 可以等价地写成如下形式(IP1):

$$\max_{m^i} h(I - P^{\mathrm{T}}) U m^i$$

$$\text{s. t.} \ (R - (I - P^{\mathrm{T}}) U m^i)_k \geqslant 0, \quad \forall\, k \in q^0$$

$$\sum_{k\in C_s} m_k^i \leqslant M_s, \quad \forall s \in S$$

m_k^i 为非负整数, $\forall k \in C$

其中$(\cdot)_k$表示向量的第 k 个元素.

3　基于仿真模型的调度及优化方法

上面的结论可以当作一个调度规则应用于仿真调度方法中. 采用事件调度法(ES-Event Scheduling)和进程交互法(PI-Process Interaction)容易建立系统的仿真调度模型[1,3]. 该模型用于描述系统的资源配置、原料的投入、产品的加工工艺以及系统的初始状态等信息, 是实际系统的精确模型, 与实际系统有一一对应的关系. 运行该模型, 在整个调度周期内经历一次加工过程, 跟踪系统的状态变化过程和导致状态变化的各种事件, 产生调度方案并统计关键的性能参数. 调度方案由一个调度指令序列组成, 确定在什么时刻(When), 用哪台设备(Where), 对哪组晶片进行操作(Which).

离散事件系统仿真的关键是对导致系统状态发生变化的事件的定义和处理. 在这里, 操作的开始和结束是两类最基本的事件. 操作的开始事件比较容易处理, 不再详述. 当晶片的某个操作结束时(假设该操作对应的缓冲区为 B_k), 执行结束事件函数: 首先释放占用的设备, 安排该晶片的下一个操作的开始事件, 然后检查 B_k 中是否有晶片等待操作, 如果有, 说明 B_k 库存水平 Q_k 大于 0, 当前的设备分配方案继续是最优方案, 因此从中选择一组晶片继续同样的操作; 否则, 说明库存水平 Q_k 已经由正值变为 0, 当前的设备分配方案不再是最优方案, 需要重新求解新的设备分配方案, 即进入下一个调度子区间. 以当前的库存水平为初始状态, 形成调度子问题 IP1, 求解后得到新的最优分配方案 m. 在仿真调度中根据 m 进一步确定把哪一个设备分配给哪一组晶片. 在整个调度周期内, 仿真模型需要反复多次调用优化模型.

4　举例

引用文献[7]中的经过简化的实例对本文方法进行验证. 一个 SMS 有 4 台设备 D, I, E 和 P, 分别对晶片进行扩散(Diffusion)、注入(Implantation)、刻蚀(Etching)和曝光(Photolithography)操作. 设备 D 是容量为 4 的"批设备"(Batch Machine), 即每次可以同时加工 4 组同类晶片. 系统生产两类产品 A 和 B, 均有 6 个操作, 依次为 $P1$, E, $D1$, $P2$, I, $D2$, 详细参数见表 1, 其中"编号"一行是指每个操作对应的缓冲区的编号. A 和 B 类晶片投入系统的时间间隔分别为 10 和 5 个时间单位. 调度周期为 80 个时间单位, 调度目标是减小平均库存, 因此, 优化模型中的费用向量 h 的每个元素都为 1.

表 **1**　A 和 B 的加工参数

操　作	$P1$	E	$D1$	$P2$	I	$D2$
设　备	P	E	D	P	I	D
加工时间(A)	1	2	4	1	1	5
缓冲区编号	1	2	3	4	5	6
加工时间(B)	1	1	4	1	1	5
缓冲区编号	7	8	9	10	11	12

表 **2**　三种方法的调度性能比较

项　目	*SimF*	*FIFO*	*FSVCT*
Avg. WIP	6.36	7.80	6.55
Avg. Cycle Time	23.83	29.54	24.67
Avg. Util. (%)	55.63	49.06	49.69

分别用 FIFO, FSVCT 和本文的方法(SimF)进行调度, 其中 FSVCT 用于缩短每组晶片

的加工周期和减小加工周期方差[4]. 调度过程中, 对调度周期内的平均库存(Avg. WIP), 晶片的平均加工周期(Avg. Cycle Time)和关键设备 D 的平均利用率(Avg. Util.)等表示调度性能的指标进行统计, 结果如表 2 所示. 可以看出, 在 3 个指标上, SimF 好于 FSVCT, 与常用的 FIFO 相比则有较大提高. 其它实例也证实了这一结论.

5 结论

本文在基于离散事件仿真技术的调度方法中引入了优化方法, 其产生最优解的基础是定理 1. 在整个调度周期内, 对每一个调度子区间, 如果都满足其实际结束状态等于根据公式(4)计算的结束状态, 那么最终的调度结果将是最优的. 但是, 由于 SMS 固有的离散特性, 本文采用的优化模型仅仅是一种近似描述, 这一条件不是总能得到满足. 仔细分析本文实例的调度过程后也发现这一点, 因此不能保证得到的解是最优解. 但是由于仿真模型总是用当前最新的系统状态形成优化模型 IP1, 模型的近似性不断得到修正, 防止了误差的积累, 因此调度的性能得到了很大的提高.

参 考 文 献

1 熊光楞, 高 红. 基于规则的工厂仿真调度环境. 信息与控制, 1994, **23**(4): 193~ 199

2 高 红, 熊光楞. 决策规则在仿真调度中的应用. 控制与决策, 1995, **10**(2): 114~ 118

3 熊光楞, 肖田元, 张燕云. 连续系统仿真与离散事件系统仿真. 北京: 清华大学出版社, 1991

4 Kumar P R. Scheduling Semiconductor Manufacturing Plants. IEEE Control Systems, 1994: 33~ 40

5 Chen H, Yao D D. Dynamic Scheduling of a Multiclass Fluid Network. Operation Research, 1993, **41**(6): 1104~ 1115

6 Connors D, Feigin G, Yao D D. Scheduling Semiconductor Lines Using a Fluid Network Model. IEEE Transactions on Robotics and Automation, 1994, **10**(2): 88~ 98

7 Liao D Y, Chang S C, Yen S R, Chien C C. Daily Scheduling for R&D Semiconductor Fabrication. Proceeding of IEEE Conference on Robotics and Automation, 1993: 77~ 83

OPTIMIZATION METHODOLOGY IN SIMULATION-BASED SCHEDULING FOR SEMICONDUCTOR MANUFACTURING

WEI Jun-hu GUAN Xiao-hong SUN Guo-ji

(*Institute of Systems Engineering, National Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an* 710049)

Abstract: Extensive simulations and good scheduling rules are usually necessary to obtain the optimal schedule in the simulation- based scheduling methods. We present a methodology to build a decomposable model for scheduling the semiconductor manufacturing systems, with the objective of minimizing the average WIP. The model is decomposed and further simplified in the context of semiconductor manufacturing. The conclusion can be directly applied to simulation- based scheduling as a scheduling rule. Because it uses the global state information in decision, the optimization ability and performance of simulation-based scheduling are improved without increasing the simulation runs.

Keywords: semiconductor manufacturing, simulation based scheduling, optimal scheduling

作者简介

卫军胡(1966-), 男, 博士生. 研究领域为复杂系统的建模、仿真与优化控制.

管晓宏(1955-), 男, 博士, 教授, 博士导师. 研究领域为电力系统的优化调度、进化优化方法、基因算法、大系统的优化理论及其应用等.

孙国基(1936-), 男, 博士, 教授, 博士导师. 研究领域为复杂系统的建模与仿真、虚拟制造、虚拟现实技术等.