Dynamic optimization of routing in a Semiconductor Manufacturing Plant

Hermann Gold

Infineon Technologies AG, Wernerwerkstrasse 2, D-93049 Regensburg, Germany
e-mail: hermann.gold@infineon.com

Abstract. We consider a semiconductor manufacturing facility with multiple products and associated routes, single servers and batch servers, job class dependant service times and external arrivals. The system is modelled as an open queueing network. The aim is to optimize routing such that cycle time constrained capacity is maximized and can be checked in reasonable computation time. We use a decomposition approach based on the connected components of a properly defined fab graph. Taking into consideration arrival rate vectors and service time matrices the routing problem for the network is formulated as a Quadratic Programming Problem (QP) involving averages and variances. The strategy for use of the manifold routing options as they typically occur in semiconductor manufacturing is to distribute load in a way such that each connected component, also called closed machine set (CMS), approaches heavy traffic resource pooling behaviour. In the presence of batch servers, mainly in the furnace area of a fab, results for the batch service queue $M/D^{[r,K]}/1$ with threshold server starting policy are combined with a new result for a batch service system with infinitely many job classes and Round Robin service discipline and applied along with the QP solver.

1 Introduction and Problem Statement

In this paper we consider the problem of routing in a semiconductor front end manufacturing facility (called fab throughout this paper) as part of shop floor scheduling. Material is started into the fab in units of lots, also called jobs, where each lot includes a prespecified number of wafers. There is no defined upper bound on the number of jobs in the fab. To a small degree admission control is imposed by specific batch service operations occurring at some early process step for a given product, but as material flow is reentrant, and since scheduling can generally not be sufficiently tuned with respect to downstream batch operations because of product variety, this sort of admission control is neglected. For the interior of our network the reentrance of flow has the consequence that scheduling policies are taken to be non-anticipating. Therefore the mathematical model for the type of fabs considered is an open stochastic queueing network with external arrivals.

Each type of the multiple products is manufactured by accomplishing a product specific set of tasks in some specific order. The activity of processing a specific task requires a single machine resource, single server or batch server. Alternate machines as a means of accomplishing a task are considered as

routing alternatives. Routing alternatives are memoryless, i. e. the set of machines of which the system manger can choose from to perform a given task does not depend on earlier choices for routing alternatives regarding the associated product. Routing alternatives can be categorized as to whether a routing decision has to be done immediately after a job finishes a previous step and enters the queue for the machine where the next processing step will be done, or each machine can choose from a global queue where all jobs belonging to a job class, which is qualified on that machine, reside. The former is called push routing, the latter is called pull routing (see e.g. [9]). In practice push routing does not exist in its pure sense, but the machines qualified for one and the same process can be located in different halls. Hence a set of machines can be partitioned into subgroups with push routing on the level of the subgroups and pull routing inside the subgroups.

In our mathematical model all the routing-relevant information of a job is included in the number of the closed machine set (CMS) currently visited and the number of the job class as which this CMS is visited by that job. A CMS is a connected component of the fab graph. The fab graph contains all machines of the fab as vertices and an edge between every pair of machines for which a process or task exists which is qualified on both machines. As we deal with Markovian routing, the problem of routing optimization can be decomposed into optimizing routing for each CMS independently.

For the network described we aim at minimizing some linear or convex function of cycle times (waiting plus service) for job classes as well as linear operating costs (e. g. costs for recycling wafers, chemicals, energy and labour). A widely used approach for controlling stochastic multiclass queueing networks, ideally in heavy traffic, is heavy traffic approximation using Brownian network models on the grounds of [5] - applications are typically shown under 90% load. This approach offers valuable insights into the behaviour and most critical aspects concerning a CMS. It uses discretization methods (called BIGSTEP, see [6]) to find the best candidate of discrete review policies for some given problem instance, however only the most basic parameters of such a policy can be found in a straightforward way. Furthermore the so called nominal plan which is input to the Brownian network is suggested to be created by a linear program. In our type of problems for a CMS with m job classes and n servers this nominal plan would select only at most $m + n - 1$ of the totally possible routing alternatives as routing options to be used by a scheduling policy (see e. g. [7]). This is optimal under heavy traffic in the strict sense (Utilization $U \rightarrow 1^-$) but far from optimal under the usual load conditions in a semiconductor wafer fab since for a fully flexible multi-server with k servers cycle time decreases with growing k. In addition there exists a more theoretical problem. The LP solution underlying the nominal plan in [7] might include routing alternatives in its optimal basis with respective value 0. Thus the graph for the corresponding CMS with edges between two machines, only when there exist routing alternatives in the