Application of Queueing Theory in Bulk Biotech Manufacturing

by

Michael Donohue

B.S. Materials Science and Engineering, MIT, Cambridge, 2003
M.S. Applied Physics, Harvard University, Cambridge, 2005

Submitted to the MIT Sloan School of Management and the Engineering Systems Division in Partial Fulfillment of the Requirements for the Degrees of

Master of Business Administration
and
Master of Science in Engineering Systems

In conjunction with the Leaders for Global Operations Program at the Massachusetts Institute of Technology

June 2011

Signature of Author

Engineering Systems Division, MIT Sloan School of Management

Certified by

Charles Cooney, Thesis Supervisor
Robert T. Haslam Professor of Chemical Engineering

Certified by

Steven Spear, Thesis Supervisor
Senior Lecturer, MIT Sloan School of Management

Accepted by

Professor, Aeronautics and Astronautics and Engineering Systems Division Education Committee

Accepted by

MIT Sloan School of Management

Application of Queueing Theory in Bulk Biotech Manufacturing

by

Michael Donohue

Submitted to the MIT Sloan School of Management and the Engineering Systems Division on May 6, 2011 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Engineering Systems

Abstract

One of the most challenging problems in Amgen's biological manufacturing facility is adhering to the daily schedule of production tasks. Delays in non-time critical tasks have been traced to temporary workload surges that exceed the production staff's capability to handle them. To quantify this effect, a method for creating an M/M/c queueing model that is specific for bulk biologic manufacturing processes was developed. The model was successfully validated by comparing the predicted results to the historical data for each of the five production shifts.

A discussion of how to model different improvement programs is presented, and Amgen-specific data are presented. It was found that across-the-board task duration reductions will reduce the schedule deviation rate by up to 50%. Additionally, it is shown that implementing staff-cross training with other production areas will reduce the schedule deviation rate between 14% and 75%. Implementation aspects of these improvement initiatives in a regulated production environment are discussed.

Thesis Supervisor: Charles Cooney
Title: Robert T. Haslam Professor of Chemical Engineering

Thesis Supervisor: Steven Spear
Title: Senior Lecturer, MIT Sloan School of Management

This page intentionally left blank.

Acknowledgments

First, I would like to thank Amgen, Inc. for its steadfast support of the LGO program and this research. In particular, I would like to thank Dave Bain, my project supervisor, for his insight and advice when I was in Rhode Island. I also want to recognize Leigh Hunnicutt for her invaluable help throughout the entire internship process.

I want to recognize the support I have received from my classmates over the past two years. I am thankful for Noramay Cadena and RJ Lehman for the discussions which kept me sane during the internship. And I am especially grateful for two years with the Steveless Six: Min Hsieh, Kevin Leiter, Diego Mendez De La Luz, Dannielle Sita, and Tim Vasil. I will miss our sangria-soaked team meetings and happy hours at Characters.

I am grateful for the support from my family, and still can't believe they never rolled their eyes or nodded off when I started talking about the details of queue formation and inventory management.

Most importantly, I want to thank my sweet wife, Adie, for all of the emotional support over the past two years. I would never have made it through without you.

This page intentionally left blank.

Table of Contents

List of Figures

1. General Problem Statement

The focus of this research is to develop a model to understand how variability in intrashift manufacturing workload affects scheduling predictability. The relevance of this topic within the biotech industry is discussed in this section and prior research on the topic is presented.

1.1. Project Motivation

Jim, a supervisor in Amgen's Perfusion/Production/Harvest area confronted a challenging decision: which production activity on the schedule did he have to delay? His technicians were supposed to start a CIP, two SIPs, and take two quality samples within the next hour, but he only had enough people to staff four of those activities. He decided to delay the CIP by three hours, to a time when there was less work scheduled, even though he knew this meant the maintenance crew would no longer be able to service that vessel before the next batch. It wasn't a great solution, but with the schedule he had, it was the best he could do.

1.2. Relevant Literature on Queueing Models

Queueing theory is a branch of applied probability that is used to quantify system performance in a variety of industries, from supermarkets, to banks, to amusement parks (1). Queueing theory has also been extensively applied within the field of computer science to improve system performance for Direct Access Storage Devices (DASD) and terminal inquiries (2). These models allow a designer to understand how randomness and variability will affect system service levels like waiting times or line lengths. These models have been used to understand how to design semiconductor fabrication facilities to improve lead times, reduce operating expenses, and pursue improvement projects (2).

However, this author has not been able to find any attempt to apply queueing theory directly to the sequencing of shop floor orders in a biotech manufacturing environment. The primary reason for this is

that queueing systems are generally used to understand performance under demand variability or multiple part processing (3). In the case of the biotech industry, the industry standard is to maintain large stocks of finished goods inventory to buffer against any demand or supply variability, and most facilities produce only a limited variety of products. Additionally, because production levels in a biological manufacturing facility do not fluctuate and shop-floor production tasks are coordinated by a scheduling group, applications of queueing theory to model demand variability have not previously been pursued. However, unique aspects of Amgen's Perfusion/Production/Harvest area within its Rhode Island manufacturing facility allow basic queueing models to be applied to predict the this area's manufacturing performance. These aspects will be discussed in the following sections.

2. Specific Problem and Industry Considerations

Overviews of Amgen's Rhode Island manufacturing facility and recombinant protein production are given in this section and area-specific processing considerations are presented.

2.1.Project Motivation

It was nearing the midpoint of the shift, and Jim and his crew had worked through the backlog of tasks from earlier. Although the next two hours would be a bit slow, this lull was followed by a period in which he had six tasks that were all scheduled to occur near the end of the shift. Unfortunately, because of the shared equipment conflicts and the validated process hold times, he was unable to work ahead on these tasks to level out the production schedule. It looks like the third shift is going to run into the exact same problem he did when they come on in a few hours.

2.2. Recombinant Protein Production Overview

Enbrel® drug production has three distinct stages: master cell cloning, bulk drug production, and formulation/fill/finishing.

As a part of the preclinical research stages for Enbrel, a Chinese Hamster Ovary (CHO) cell was genetically engineered to produce a specific Tumor Necrosis Factor (TNF) inhibitor protein. Pending successful clinical trials, a Working Cell Bank (WCB), consisting of CHO cells cloned from this single master cell, is created. These vials of cloned cells are flash frozen and kept at liquid nitrogen temperatures to prevent any genetic mutations from occurring (4). Vials of cells from this WCB are critical raw materials for the bulk production process.

At the beginning of the bulk drug production process, a single vial of cells is thawed and the cells are introduced into a growth medium that contains the nutrients necessary for them to begin self-replication. At specified time intervals, this batch of cells is transferred into successively larger vessels. During this transfer process, cellular waste is removed and new growth medium is added. This "scale-up" process continues in successively larger vessels until the desired cell density and reactor volume is achieved. At this point, the cells are chemically and physically induced into producing their specific antibodies. When the desired protein volume has been secreted by the CHO cells, the protein is harvested from the cells and purified through a series of chromatography and filtration steps. The bulk drug substance is then loaded into a sterile vessel for refrigerated storage or transportation. An example process similar to the actual Enbrel bulk drug production process is shown in Figure 1.

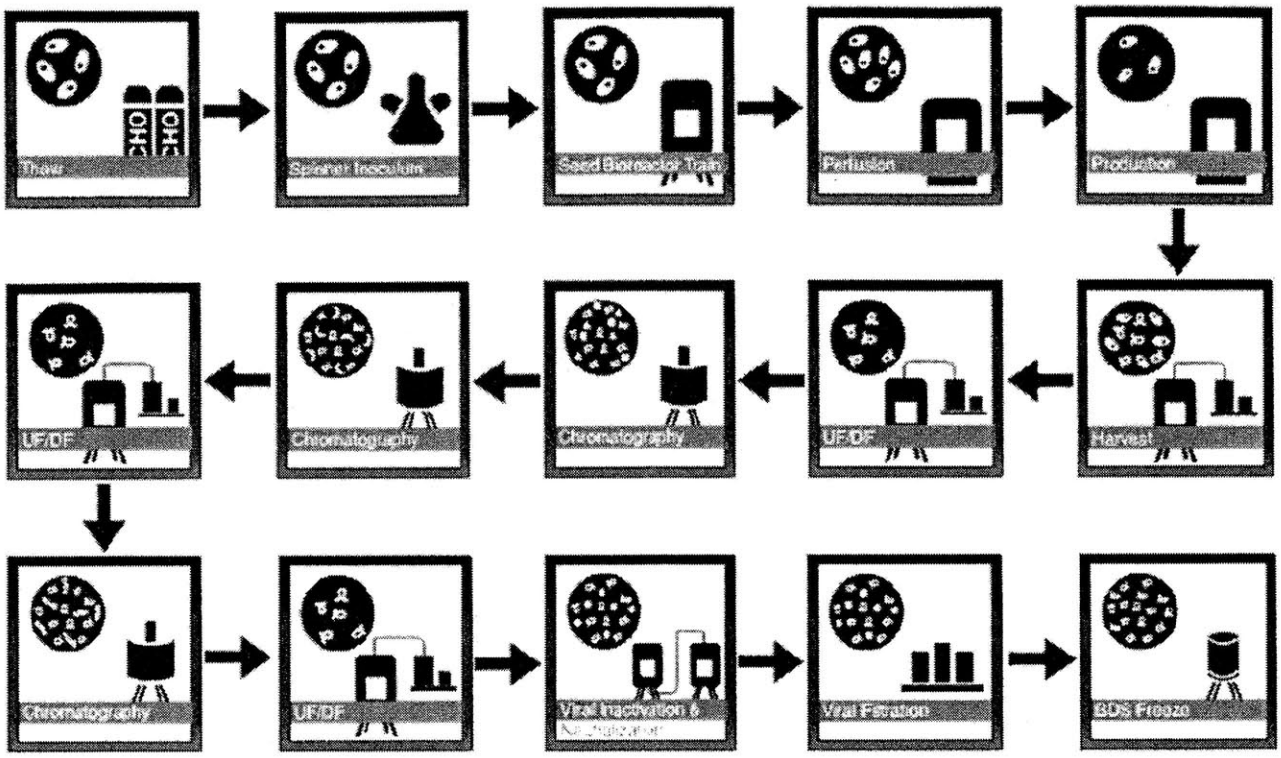

Figure 1: An example process similar to the Enbrel bulk drug manufacturing process (5).

The bulk drug substance may be held in storage for an extended period of time prior to its final formulation processes. In order to fulfill customer demand, the bulk drug substance in these vessels is reconstituted to the desired concentration with various buffer solutions. The final drug product is loaded into its commercially available dosing forms, then labeled and packaged for shipment to distributors or end users.

2.3.Amgen Rhode Island Background

In July 2002, Amgen acquired Immunex and its blockbuster drug Enbrel (6), the first TNF-α inhibitor approved for the treatment of rheumatoid arthritis (7). Within the Amgen production network, the BioNext facility in Rhode Island is responsible for the majority of Enbrel bulk drug production (WCB generation and formulation/fill/finish activities occur at other sites within the Amgen production network). With nine 20,000L production bioreactors, this facility is one of the largest bulk drug

production sites in the world (8). The Enbrel produced here now accounts for over 25% of Amgen's $14B in annual revenues (9).

2.4.Processing Area Specifics

The particular processing area within the Rhode Island facility that is the subject of this study is the Perfusion/Production/Harvest area. This processing area is made up of three distinct processes that occur multiple weeks after the initial working cell vials are thawed, and is shown in the upper-right corner of Figure 1.

Perfusion: This is the final "scale-up" stage of the process. During this process, which lasts multiple days, fresh growth media is being continually fed into the bioreactor while spent media and waste products are being filtered out.

Production: At this stage, media solutions are added to the growth mixture to induce the cells to begin to produce the Enbrel protein, rather than continue dividing. The cells produce the protein until the desired concentration is reached, which takes between one and two weeks.

Harvest: This process is the start of the final filtration steps. Any remaining CHO cells and cellular waste in the reactor are filtered from the remaining mixture of unused growth media and Enbrel protein.

2.5.Additional Process Considerations

Each of these processes occur in oxygen-rich liquid nutrient environments in vessels heated to nearly 37C. These conditions are ideal not only for CHO cell replication, but also for contaminant growth. In order to ensure that there are no viral or microbial contaminates that will grow and deny nutrients to the CHO cells, an extended cleaning cycle are conducted prior to inoculating any bioreactor with a batch of cells. A Clean In Place (CIP) process, consisting of multiple acid, base, and water rinses is flushed through the vessel and transfer lines. These steps are designed to dissolve and remove any organic

contaminants and debris from the inside of the equipment. Following that process, a Steam In Place (SIP) process is conducted. The SIP step, in which specially treated steam is flushed through the vessels, is designed to sterilize the inside of the equipment after it has been cleaned by the CIP process. The operating conditions for these two processes (number of cycles, temperatures, hold times) are validated to ensure that the equipment is clean and sterile prior to the start of each batch. Because of the complex system geometries involved, these processes can take over a day to complete. After the vessel cools, the growth media can be introduced and equalized, and finally, the vessel can be inoculated with the new cells. These steps can be seen in Figure 2.

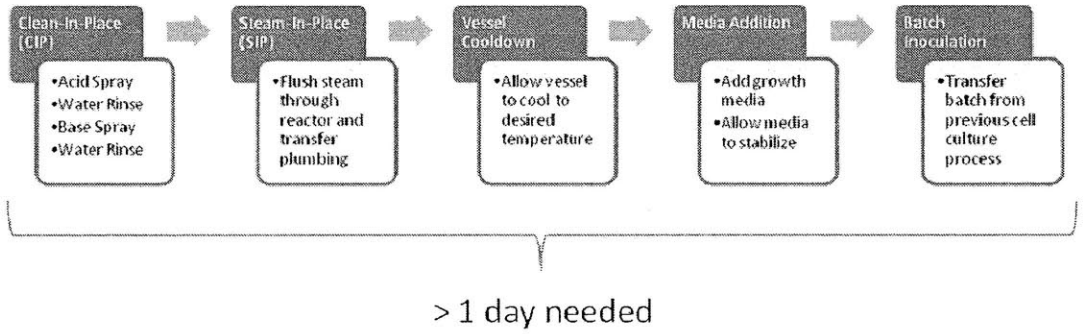

Figure 2: An outline of the process steps necessary prior to inoculating a vessel.

2.6.Role of Central Schedule

Because of the nature of the biological manufacturing process, a centralized production schedule is used to coordinate the activities of the different manufacturing groups. This schedule contains timing information for all quality samples, batch transfers, media feeds, and maintenance tasks.

The complexity of process sequencing, the time scale involved, and the need to resolve shared equipment conflicts requires a central scheduling system. As an example, if a perfusion vessel was just inoculated with a batch, a production vessel will need to be CIP'ed within 24 hours to ensure that this process, a subsequent SIP, and a media transfer can all be completed on that vessel prior to its inoculation

five days later. However, the equipment used for the CIP and SIP may be shared with the media mixing vessel and equipment in other areas. In order to resolve these equipment conflicts, a master schedule which prioritizes needs across multiple areas is used to ensure that the cell growth processes are uninterrupted.

2.7.Task Details

For the purposes of this analysis, all of the production activities that occur in this area are classified into three distinct groups: time critical tasks, maintenance tasks, and non-time critical tasks.

- *Time Critical Tasks*: The timing of these tasks is critical to the quality and yield of the final bulk drug product. Examples of these tasks include quality sampling, media feeds, and batch inoculations. Delays or cancellations in these tasks have very costly consequences, including possibly losing a batch. CIP and SIP processes are time critical if their purpose is to prepare a bioreactor for inoculation.
- *Maintenance Tasks*: As in any other manufacturing area, regular preventive maintenance must occur on equipment to prevent unplanned breakdowns. Examples of these preventive maintenance activities include gasket changes, seal replacements, and measurement device calibrations. Depending on the specific maintenance task, the frequencies of these activities vary between monthly and annually. These activities generally involve accessing the interior of the bioreactors or transfer pipes, so all pressure or chemical inlets must be locked out to ensure the safety of the maintenance technicians. The activities are time-consuming and even simple maintenance tasks can often take an entire shift to complete.
- *Non-Time Critical Tasks*: The timing of these tasks are not essential to the quality or yield of the processes. Generally, these tasks are done to support other activities in the area. For example, prior to conducting any open-vessel preventive maintenance activities, the vessel must be CIP'ed. A delay or cancellation of this CIP only affects the completion of this maintenance activity.

Although delaying or cancelling this activity may be undesirable from a quality or regulatory perspective, it has no impact on any biological processes.

2.8.The Effect of Validated Hold Times

One aspect in particular that separates biotech and pharmaceutical manufacturing processes from most other industries is the effect of validated process hold times. In order to reduce the sampling burden associated with the production process, the FDA allows manufacturers to validate time limits after certain processes are complete in which the vessels are considered to be in a particular state. For example, after a CIP/SIP cycle is completed, the sterility of a particular vessel may be assumed for a limited period of time (generally on the order of one to two days). Likewise, after a vessel has been emptied, there is a validated time period in which it can be assumed that any remnants within it have not dried and adhered to the walls. However, if these validated hold times expire, these assumptions no longer hold. In order to return these tools to active use, multiple additional cleaning and measurement cycles must be performed. The practical effect of this manufacturing regulation is that the production groups are unable to "work ahead" during slow periods when no work is scheduled.

2.9.Schedule Deviation Overview

Because the current equipment capacity does not constrain production activities, and because the current scheduling software notifies the schedule generator about equipment conflicts, deviations from the pre-set schedule have two primary root causes: machine breakdowns or manpower issues.

- *Machine Breakdowns:* These schedule disruptions arise as a result of normal production operations. Examples of these disruptions include gasket leaks, filter failures, and gauge failures. When these breakdowns are detected, they must be remedied prior to processing the next batch on this equipment. These repairs generally can't be completed in a time period that avoids schedule disruption.

- *Manpower Issues:* This facility runs 24 hours a day, seven days a week. Because of the time-sensitive nature of many of the processes, the processing areas are staffed to ensure that all activities scheduled for a particular shift can be completed. However, due to differences in shift staffing levels or employee vacation scheduling, the number of employees available on a particular shift can vary from day to day. Additionally, the scheduled workload (number and type of task scheduled per unit time) varies over the course of the shift. If a peak in the schedule workload occurs that is greater than the capability of the floor staff, schedule deviations will occur as the floor staff shifts non-time critical tasks earlier or later.

For the purposes of this paper, a "schedule deviation" event is defined as any instance in which an activity is started more than a given number of hours away from its scheduled start time. For example, a Clean In Place (CIP) activity that is scheduled to start at 19:00 but is started early at 17:35 would be considered a deviation if the allowable time window is one hour, but not if the time window is extended to 90 minutes.

The PPH area in Rhode Island was selected for this study because it has the highest deviation rate of any area on site. However, it should be noted that the deviation rate is still low enough that time-critical tasks are not affected, and there is no impact on the process yield or product quality. The impact of these deviations can be felt most strongly in the efforts of the maintenance and validation support groups. These groups also rely on the daily production schedule to coordinate their work, and when tasks are delayed, they are often prevented from being able to complete their work.

Other papers have treated the subjects of maintenance scheduling (10), optimizing time spent conducting preventive maintenance relative to unplanned maintenance, and predictive maintenance. However, no work has been published that models the schedule deviation rate due to manpower issues that arise when workload peaks exceed the capability of the production staff. The next section of this thesis will present a theoretical and empirical model of this issue.

3. Model Development

In this chapter, a queueing model is developed to quantify the time that individual processing tasks wait before being serviced. Queueing models incorporate data on task service times, task inter-arrival distributions, and the number of process servers to calculate relevant performance statistics. The empirical data for Amgen's PPH processing area is presented, and a method for validating the model will be discussed.

3.1. Data Collection Plan

Because biologic manufacturing is a tightly controlled process, production parameters for every activity are logged into a production data management system. These parameters include scheduled and actual start times for activities, equipment settings, and if a deviation occurs, a root cause which can be selected from a set of choices. Because this data is gathered under standardized stable conditions with over one year of history, it forms an ideal set for analysis. Extending this analysis into other areas or sites is made less challenging by the ready availability of this data.

3.2. Task Service Time Distribution

The distribution of task service times can be obtained either through direct observation of the actual activities, or through compiling the plant's labor standards if they are available. From this author's experience, direct observation of the tasks is the preferable method. Although data collection using direct observation is more time-consuming, the data set may be more up-to-date and have less distortion than relying on a set generated from current labor standards, which are often compiled to satisfy the desires of different stakeholder groups.

Because of the nature of the production activities that occur in a biological manufacturing facility, it may be appropriate to assume that there is an exponential distribution of these task service times. Some

of the most common activities within a large-scale commercial cell culture facility are CIP and SIP processes. The setups and teardowns for these processes typically involve connecting one or two pipes to the target vessel, running a pressure hold test, then starting an automated recipe. These steps can all be completed by a two-person team in less than 30 minutes. For comparison, vessel transfers or reactor teardowns occur much more infrequently and can take a production team many hours to complete.

When generating the task time distribution, it is essential to "weight" the tasks by the relative frequency with which they occur. If a process must be done more than once each time a batch is processed, it should be counted multiple times. For example, a vessel may be CIP'ed after use, then again after an invasive maintenance procedure, then a third time prior to its next use. This distribution was measured for the Perfusion/Production/Harvest processing area within Amgen's Rhode Island facility and the data are shown below:

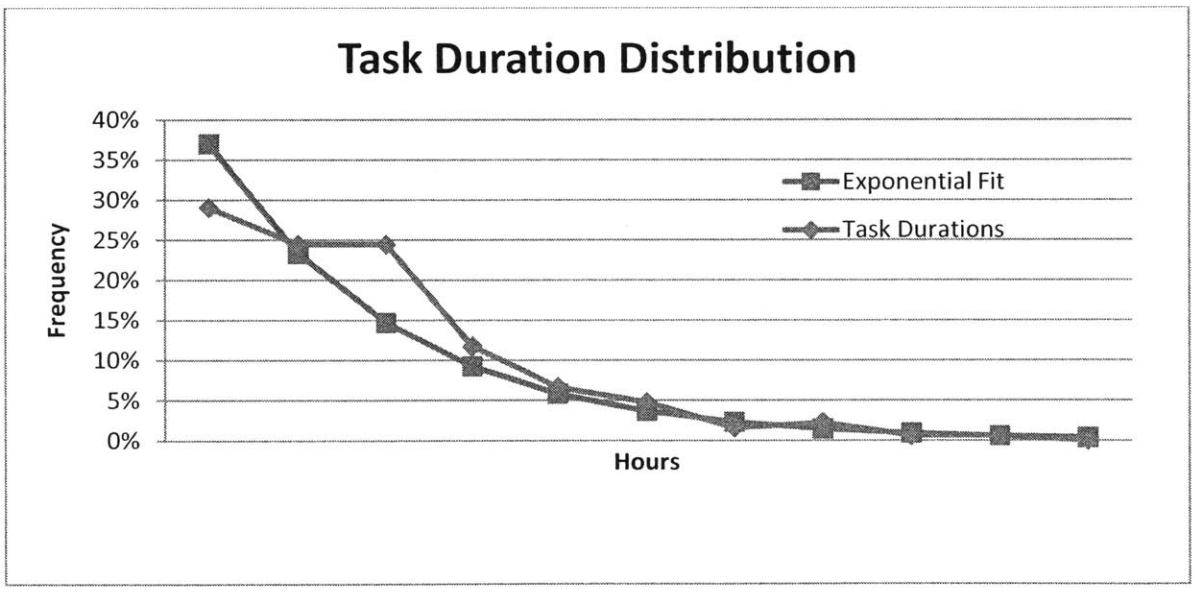

Figure 3: The frequency distribution for processing task durations.

Figure 3 shows the frequency distribution for the processing task durations in the ARI PPH area with an exponential fit overlaid. To protect proprietary processing information, the time scale on the horizontal axis has been removed.

It should be noted that this task time distribution only consists of activities that involve actual work. In a biological process, many of the production steps involve multi-day scale-ups or multi-hour cleaning recipes. Although these activities are monitored by the floor staff to ensure that the automation is running smoothly, these monitoring activities are generally "overflow" activities: while doing other activities, technicians will briefly glance at a computer screen to determine if there are any process alarms. Unless the process is of a critical nature and requires a technician to monitor its progress the entire time, monitoring automated recipes is considered to have a negligible impact on a technician's time and is not included in this analysis.

3.3. Task Inter-Arrival Distribution

Another input to a queueing model is the distribution of interarrival times for tasks. If available, past processing data should be collected from scheduling software. The distribution of interarrival times for the Rhode Island Perfusion/Production/Harvest processing area are shown below.

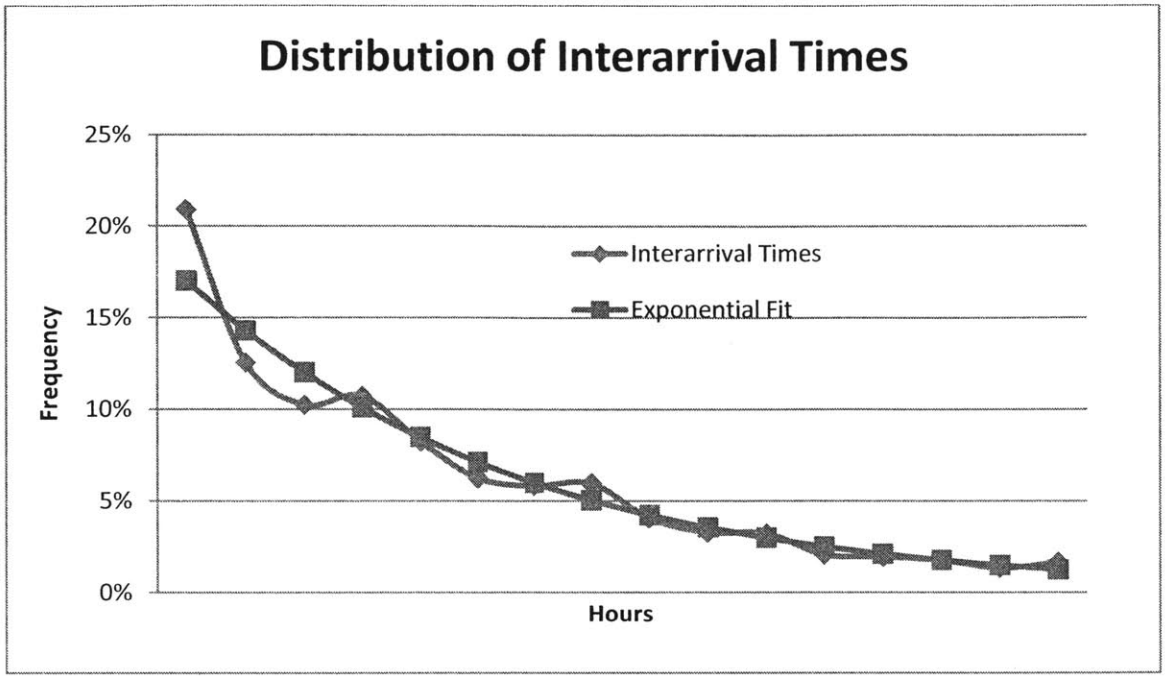

Figure 4: The frequency distribution of task interarrival times.

Figure 4 shows the frequency distribution of the task interarrival times for the AIR PPH processing area with an exponential fit overlaid. The time scale on the horizontal axis is not displayed to protect Amgen proprietary processing information.

One of the fundamental aspects of a queueing situation is that the process servers are not allowed to "work ahead": the queue length of tasks awaiting services is always non-negative. This asymmetry in the probability distribution function of the queue length leads to a positive average wait time, even when capacity utilization is well below 100%. A critical aspect of biotech manufacturing is that the processing vessels are validated to be held in an idle state for only a short period of time. This requirement precludes the production staff from completing future tasks during periods of slow or no work, and the queueing model breaks down without this constraint.

3.4. Memory-Less Arrival Assumption

From Figure 2, it can be observed that the actual data can be approximated with an exponential distribution. A Poisson process is a specific type of process which generates an exponential frequency

distribution of interarrival times. A defining aspect of a Poisson process is its "memory-less" nature: data on the arrival time of one task gives no information on the arrival time of the subsequent task. For example, the likelihood that a customer will enter a fast food drive-through in the next five minute period doesn't depend on whether another customer entered during the last five minute period.

According to this definition, the PPH production area does not meet this standard. For example, if a Day 3 Production Quality Sample was just completed, it is known that the next Day 3 Production Quality Sample will not occur for a few days. Each individual batch has low variability (11) in its arrival rate, which would preclude the use of the memory-less assumption.

However, because there are multiple batches in process within this area "feeding" tasks to the production servers, for the purposes of this analysis, the superposition of each of these low variability sources can be treated as a medium variability, or memory-less process (11). For example, knowing that a Day 3 Production Quality Sample just occurred gives no information on when of what the next production task will be, other than the fact that it will not be another Day 3 Production Sample. Because there are over six batches in process at any given time in this area, requiring roughly 100 tasks necessary to process each batch, knowing that any one particular task just occurred gives no information about when or what the next task is.

Figure 5 shows the interarrival times for the first 100 shop floor orders scheduled for March 2010. The scale on the y-axis has been eliminated to protect proprietary processing data, but the graph is drawn so the x-axis intersection is at 0 hours.

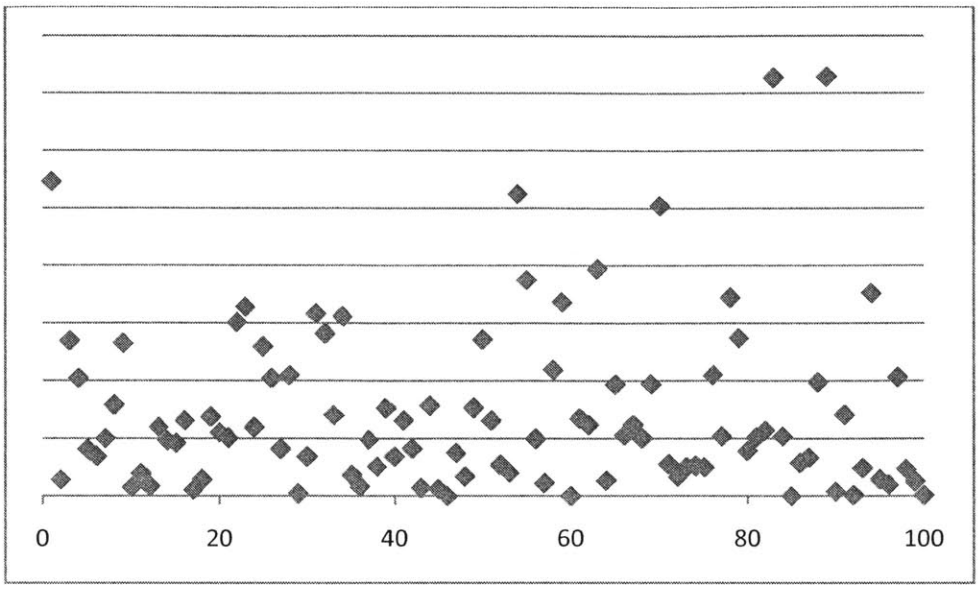

Figure 5: Interarrival times for the first 100 shop floor orders in March 2010.

This chart suggests that there is little serial correlation in the interarrival times for these tasks, which would not preclude the use of the memory-less arrival assumption.

In general, a Poisson process can't be used to model an area's workload when the work for the area is planned by a scheduler: a pre-planned task schedule violates the assumption that the tasks arrive in a memory-less fashion. However, from discussions with the current scheduling group, there is currently no effort made to "smooth" workload peaks by moving non-time critical tasks to "slow" periods during particular shifts. However, the scheduling group does attempt to move non-time critical tasks from shifts that have low staffing levels (overnight shifts) to shifts with higher staffing levels (daytime shifts). This uneven work load across different shifts allows model validation that will be explored in section 3.8.

3.5. Number of Servers

The number of servers in this model is the number of two-person work teams that are on the floor at any given time. This number is half the product of the shift headcount and a utilization constant.

An industry-specific aspect of biotech manufacturing is that all production activities must be completed by two-person work teams. The role of the secondary employee is to follow the primary employee and ensure that all operations are done according to the written procedures. The secondary employee must verify that these operations were done to standard. This work structure is generally mandated by cGMP regulations, and is considered standard for the industry.

The utilization constant is the fraction of time that a manufacturing technician is actually available to do production activities. A hypothetical calculation of this utilization constant is given below:

(52 weeks per year) *(40 hours per week) = 2080 hours possible

(5 weeks vacation/sick time/holidays) * (40 hours) = 200 hours unavailable

(47 work weeks per year) * (2 hours per week on training activities) = 94 hours unavailable

(47 work weeks per year) * (2 hours per week on improvement activities) = 94 hours unavailable

(47 work weeks per year) * (2 hours per week of administrative activities) = 94 hours unavailable

(235 work days per year) * (30 minutes per day of mandatory paid breaks) = 117.5 hours unavailable

(235 work days per year) * (30 minutes per day spent gowning/degowning) = 117.5 hours unavailable

(235 work days per year) * (30 minutes of meetings per day) = 117.5 hours unavailable

Given this calculation, out of the 2080 hours possible, a technician may only be available to perform production activities 1246 hours per year, or 60% of the time. In this hypothetical example, if 10 people are staffed on a shift in a particular area, on average, there will only be three two-person production teams available to handle the incoming work load at any given time.

3.6. M/M/c Model

Given the assumptions made above, an M/M/c model was used to estimate the schedule deviation rate. This model has been treated extensively in the literature, and some of the relevant performance metrics are:

Metric	Equation	Reference
An approximation for the average waiting time in queue	$\frac{\rho^{\sqrt{2(c+1)}-1}}{c(1-\rho)}\ \mu$	(11)
Probability an arriving task must wait in a queue (Erlang C-Function with parameters c and cρ)	$\frac{(c\rho)^c}{c!}\left[(1-\rho)\sum_{n=0}^{c-1}\frac{(c\rho)^n}{n!}+\frac{(c\rho)^c}{c!}\right]^{-1}$	(2)
Probability a task in the queue must wait longer than time T	$e^{-c\mu(1-\rho)T}$	(2)
Capacity Utilization	$\frac{\lambda}{\mu * c}$	(2)

Table 1: Equations giving relevant queueing metrics.

In these equations, μ is the average production task duration (in hours), λ is the average task arrival rate (in tasks per hour), c is the number of production servers, and ρ is the capacity utilization.

Given these equations, the schedule deviation rate, defined by the frequency at which any arriving process task has to wait over T hours is simply the product of the fraction of incoming tasks which must enter a queue and the rate at which tasks in the queue wait over T hours.

3.7. Treatment of Non-Integer Server Values

The example shown in section 3.5 shows a very straightforward calculation of the number of 2-person production teams present in the manufacturing area. However, in the general case, the number of servers will be a non-integer value, and the queueing framework developed in section 3.6 breaks down.

The physical significance of having 2.73 production teams present in an area is that there are 2 teams on the floor 73% of the time, and 3 teams present 27% of the time. Because production supervisors generally attempt to maintain even staffing levels during a shift, it can be assumed that the number of production teams does not drop below this next lowest integer or increase above the next highest integer. Rather than treating the way that the number of servers fluctuates in a time-dependent manner, in cases of a non-integer number of production teams, this author proposes to use the average delay frequencies for the two nearest integer solutions, weighted by their frequency of occurrence. This approximation neglects if production supervisors shift breaks/lunches from time periods of high workload to time periods of low workload, and ignores transient effects in the deviation rate when the number of production teams is increased or decreased. These effects cannot be incorporated easily into this model, so future computer simulation work in this area may be useful to obtain a more accurate understanding.

The calculation of this approximation can most easily be understood by using a hypothetical example. The table below gives the frequencies that a task will be delayed by at least two hours for different integer server values and an average task duration of 75 minutes.

	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5	2.75	3

1	12%	40%	100%									
2	0%	2%	7%	20%	46%	100%						
3	0%	0%	0%	1%	4%	10%	22%	48%	100%			
4	0%	0%	0%	0%	0%	1%	2%	5%	11%	24%	50%	100%
5	0%	0%	0%	0%	0%	0%	0%	0%	1%	3%	6%	12%
6	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%

Table 2: Queueing delay frequencies for given values of c and λ.

Table 2 gives the delay frequency as derived from the formulae in Table 1 for given values of c and λ. μ is selected as 0.75 for this table as an example of a realistic average task processing time. In this example, the delay frequency for λ=1.25, μ=0.75, c=2.73 is approximated as:

$$(0.73)*(46\%) + (0.27)*(4\%) = 35\%$$

3.8. Model Validation

Given the assumptions and approximations discussed in 3.3, 3.4, and 3.7, this model allows a direct calculation of the schedule deviation rate within the manufacturing facility. Table 1 gives the probability that a task must wait T time before being serviced. This value is equivalent to the probability that a production task must deviate outside a window of +/- 0.5T from its scheduled start time.

To validate this model, data from multiple manufacturing shifts may be used. Bulk biotech manufacturing facilities must operate on a non-stop schedule. Biological processes are occurring all day, and there are no stoppages for evenings and weekends. Care must be taken, however, when considering each shift's work load, as given by the inverse of the mean arrival time. In order to match the work load as closely as possible to traditional working hours, many non-time sensitive processing tasks are shifted to weekday mornings and early afternoons. This uneven work balance and shift staffing presents an

opportunity to validate the model's assumptions by comparing deviation rates across different shifts. Although some shifts may be more heavily staffed than others, they may also have the highest workload. If deviation data is available for each processing shift, the model can be validated by comparing predicted deviation rates for each shift with the actual deviation rates for the shifts. Alternatively, if the process is stable and the data is available, the deviation rates for earlier time periods can be used for comparison/assumption validation.

The equations in Table 1 express the deviation rate as a function of the average interarrival rate, the average task duration, and the number of process servers. In order to assess the validity of the model, a production parameter needed to be predicted based on other data sets, then compared to the actual production parameter. Because the task arrivals and durations were studied to ensure that they met the model assumptions, and because the deviation rates were published and discussed in the production areas, the shift headcount was selected as the validation parameter. Table 1 was used to solve for each shift's headcount as a function of that shift's deviation rate, its workload, and the task durations, and this predicted shift headcount was compared to the actual shift headcount. The model's error in predicted headcount as a function of each shift is shown in Figure 6. This chart shows the percentage error between the predicted headcount based on the workload and deviation rate and the actual headcount for the PPH processing area. Negative values indicate shifts with a lower actual headcount than predicted by the deviation data and model.

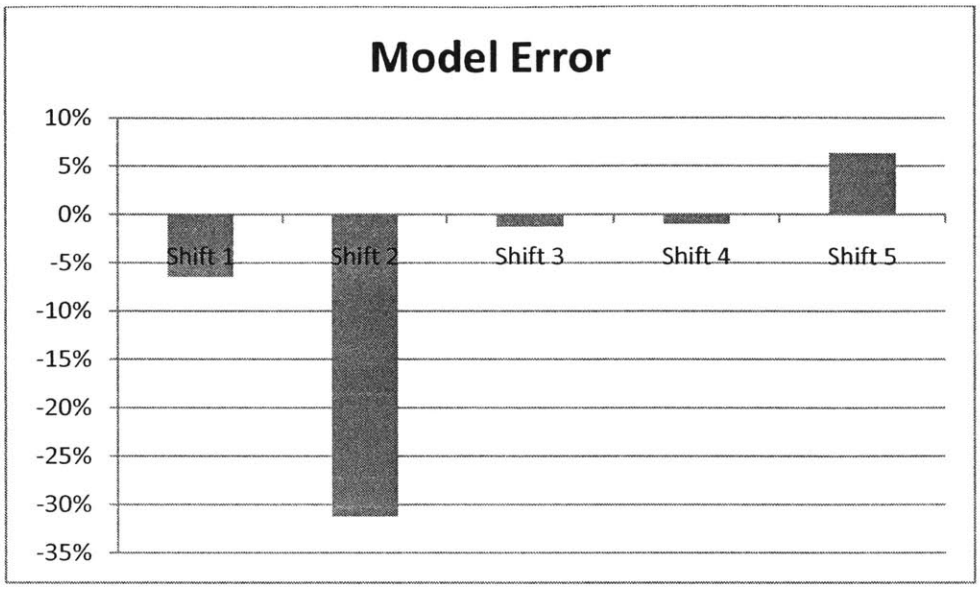

Figure 6: Queueing model error as a function of the production shift.

3.9 Second Shift Model Error

From Figure 6, it can be seen that the model accurately predicts each shift's headcount, with the exception of the second production shift. There are three possible explanations for this discrepancy:

- Because of higher average operator tenure/skill level on this shift relative to the others, the average task time (μ in the queueing equations) may be lower than measured for the other shifts.
- The actual headcount in the area during this shift could be higher than assumed. This could be the case if fewer improvement or training activities were conducted during this shift. However, informal discussions with the shift manager and technicians on this shift did not seem to support this theory.
- Schedule deviations for this area may be getting logged into the computer system incorrectly. The focus of this study is to quantify deviations that occur when the workload exceeds the capacity of the floor staff. However, if an operator logs a manpower-related deviation as

equipment-related or a different alternate cause, examining manpower-related deviations only will not give a true characterization of the actual state of the system.

4. Application of Model

Given the queueing model developed in the previous section, the current state process parameters can be altered to quantify the effect that different changes in the production states will have on scheduling stability. This chapter will explore some potential changes.

4.1. Increasing Production Rate

Because overbuilding capacity in a biotech manufacturing facility is so much less expensive than underbuilding capacity, biological manufacturing facilities often have excess equipment capacity. However, these facilities are only staffed at levels considered economically efficient for the current production rate. As drug demand conditions increase (due to approvals of additional clinical indications, reduction in competition or new geographic market development), a biological manufacturing facility may need to increase its production rate to match the demand increase. The queueing model developed in the previous chapter can be used to estimate staffing needs for increased run rates.

For example, Amgen may consider increasing its Rhode Island facility's Enbrel production rate by 20-50%. These production rate increases can be modeled by increasing λ, the task arrival rate used in section 3.6. Therefore, in order to achieve the same schedule deviation rate as the current baseline at these higher production rates, each shift's headcount needs to be increased by 9% for a 20% increase in production rate, or increased by 27% for a 50% increase in production rate.

Modeling an increased production rate using these equations is an acceptable application of this queueing model, however, care must be taken when trying to extend this model to lower production rates. The M/M/c model presented in section 3.6 relies on an assumption of memory-less task arrivals. Increasing the production rate strengthens this assumption, because there will be a greater number of

batches being processed in the area at any given time. However, decreasing the production rate weakens this assumption. In the limit of only a single batch in process at a time, this assumption is completely violated, and other scheduling models must be used. Other sources have suggested a limit of 4 concurrent batches above which this memory-less assumption may be applied (11).

4.2. Improvements from Kaizen

Kaizen is the process of continuously implementing small, incremental improvements, which can be contrasted with radical process innovation (12). An effort that is currently underway within the ARI PPH processing area is to implement a series of small-scale improvements that reduce the workload burden for every production task. This focus on implementing numerous small improvements, rather than attempting to generate large process that radically streamline the most time-consuming tasks, is due to the nature of the regulatory environment. Because the drugs produced by these processes are injected directly into a patient's bloodstream, regulatory authorities need to ensure that the final product remains unaltered by any process changes. Radical changes to time consuming task often need extensive process validation and may even require refilling and review by the FDA. However, smaller changes do not require the burden and can be completed in a much shorter time period.

The cumulative effect of these changes can be modeled using the queueing model developed in section 3 by changing the value of μ, the average task processing time. For example, if the current average task processing time is one hour, an across-the-board 10% decrease in task durations is an increase in μ from 1 to 1.1. The deviation rates for new values of μ were calculated and shown below in Figure 7. This graph shows the percentage reduction in the deviation rate as a function of decreasing process task durations with constant headcounts. The x-axis intersection is 0%, but the scale was eliminated to protect proprietary data.

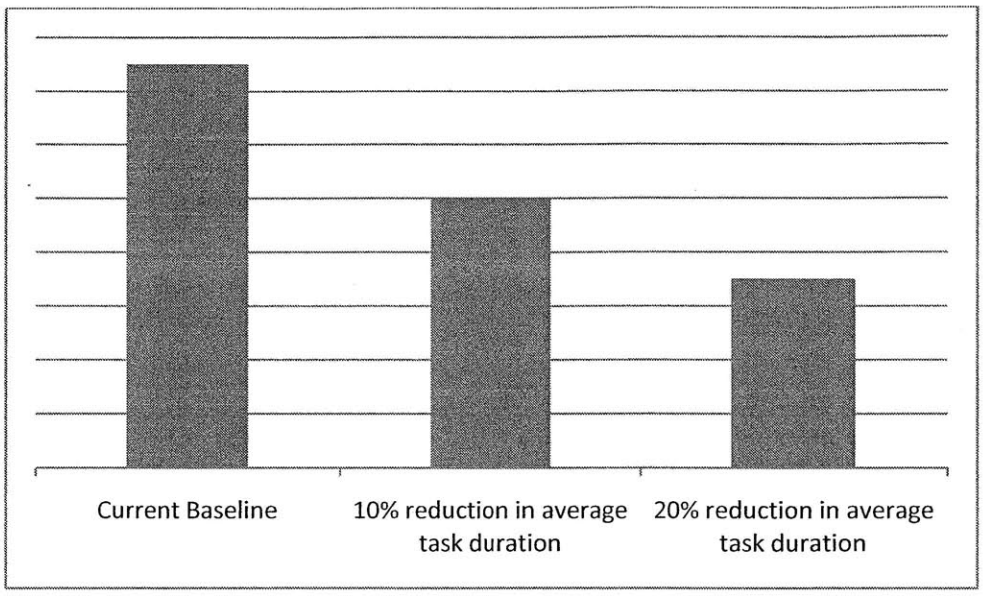

Figure 7: Impact of task duration reduction on deviation rate.

From this figure, it can be seen that a 10% reduction in the process task durations in this area will lead to a 33% decrease in the schedule deviation rate, while a 20% reduction in the task times will result in better than a 50% improvement in this metric.

Once again, it is useful to consider that the task times under consideration are the "active" times (when technicians are actively performing tasks), not the "passive" times (when technicians are monitoring automated processes to ensure that the equipment is running smoothly). Reducing the amount of time spent monitoring the processes would mean transferring batches between vessels faster or running fewer flushes for each CIP. During the process qualification period for the plant, these steps were fully validated, and any changes to these automated recipes are extremely difficult and may require refiling/requalification with regulatory authorities.

However, reductions in the "active" processing times can be achieved through the use of many standard industrial engineering approaches to reduce unnecessary motion or parts searching. Examples of these simple solutions include storing jumper pipes at the point of use rather than in a central location, eliminating the need for an operator to gown/de-gown from a controlled area to complete a task, and

reducing the number of sheets that an operator needs to print out and sign. Each of these tasks taken individually may only account of one or two minutes of processing time, but taken as a whole, they can easily eliminate three minutes from a 30-minute task. None of these changes would involve refiling/requalification with the regulatory authorities, and the process validation necessary to implement them would be minimal.

These changes can also be considered in proportion to the difference in processing time between a novice operator and an experienced operator. From this author's discussions with area managers and production supervisors, teams of novice operators generally take between five and ten minutes longer to perform most standard production tasks than teams of experienced operators. Increasing the level of process standardization and operator training will reduce this difference and will reduce the overall schedule deviation rate.

4.3. Effect of Changing Headcount

A different approach to increasing the schedule adherence is to simply add production personnel. Adding staff to each shift increases production capacity and reduces the probability that a temporary surge in scheduled work will exceed the capability of the floor staff. The model developed in the previous sections can also be used to quantify this effect.

While the schedule deviation rates can be recalculated directly using the equations in section 3.6, a less computationally complex approximation is available. For the conditions used to calculate Table 2, it can be seen that each time the number of process servers, c, is increased by one, the deviation rate drops to between 5% and 10% of its previous value. This *ReductionFactor* can be used to estimate the deviation rate reduction that would occur if one technician is added to the shift. Each staff member added to a shift increases c, the number of two-person production teams on the floor, by *UtilizationConstant* / 2, in which *UtilizationConstant* is the fraction of time available to do work described in section 3.5.

Therefore, using the approximation method developed in 3.7, the decrease in the deviation rate for each additional technician added to a shift will be:

$$Percent\ Decrease\ in\ Deviation\ Rate = ReductionFactor^{\left(\frac{Utilization\ Constant}{2}\right)}$$

When table of the deviation rates using Amgen's PPH processing values is calculated, a *ReductionFactor* from adding an additional server of 0.06 is calculated. Therefore, using this area's empirically determined *UtilizationConstant*, it is found that adding one additional employee to each shift will reduce the deviation rate by 54%.

5 Areas for Future Work

In this section, two potential projects that can improve production schedule adherence are discussed within the framework of the queueing model that has been developed.

5.1. Cross-Training Technicians

Relative to the solution of reducing the overall task durations, hiring new technicians is a relatively expensive and time-consuming solution to reducing the deviation rate. Technicians in these manufacturing areas are generally well-educated and highly compensated, and training a new employee may take months. However, cross-training technicians from one area to conduct operations in another area is a potential solution for adding staff when needed to handle the work load. This solution has a number of advantages:

- Technicians would broaden their understanding of how their area's operations affect the PPH area and may learn how to more effectively manage the processes within their original function.
- The non-time critical tasks most likely to be delayed are CIPs and other cleaning functions. These tasks are very similar between different processing areas within the facility, so the

amount of additional training necessary for the technicians to perform these new tasks is minimal.

Too many assumptions are broken in the queueing model to use it to accurately predict the impact of technician cross-training on the overall deviation rate. However, the model can be used to estimate the upper and lower bounds of this change

- Upper bound of effect: The upper-bound of this effect would be to treat the cross-training as adding one technician to the floor at all times. This would be modeled as an increase in the number of production teams by 0.5 and a corresponding decrease in ρ, the capacity utilization. This overstates the reduction that would actually occur, since this additional technician would not be present on the floor at all times. However, using the approximation developed in Section 4.3, the upper bound of this effect is calculated as a 75% reduction in the deviation rate.
- Lower bound of effect: The lower bound of this cross-training would be to assume that an additional technician would only be called into the area to handle a task that was about to be delayed. This lower bound would be modeled by leaving both c and ρ unchanged, but increasing T, the acceptable time period a task can be processed in without being considered a deviation. Because it takes an average of μ hours for two operators to complete a task, it is assumed that one operator can complete a task in $\mu/2$ hours. Therefore, the addition of one technician to this processing area will push out the time cut-off from T hours to T + $\mu/2$ hours. For Amgen Rhode Island's PPH processing area, this change is a 14% reduction in the deviation rate.

Cross-training is certainly an appealing solution to the challenge of reducing the scheduling deviation rate, and it was being considered for implementation prior to this author's departure from the facility. However, multiple organizational and regulatory challenges need to be addressed in order for this to become a viable long-term strategy for addressing scheduling challenges:

- Organizational challenges: How would the PPH supervisors call on technicians in other areas to help? What new communication channels need to be developed? How can potentially competing staffing needs of different areas be managed to achieve a global schedule optimization?
- Regulatory challenges: How can Amgen ensure that technicians in one area of the plant maintain their training status in tasks they may only perform infrequently in another area? How can their task competency be assessed regularly?

Future projects in this area should address these challenges and should compare the model's predicted effect of this change with the actual improvement.

5.2. Level-Loading the Schedule

The fundamental cause of scheduling deviations that is examined in this research is that the scheduled workload can temporarily exceed the capability of the production staff. The previous sections examined ways to either reduce the overall workload or increase the working capacity in the area. In this section, the scheduling systems themselves are discussed.

Currently, Amgen uses a finite scheduling software package to coordinate production. This software only has functionality that resolves shared equipment conflicts: it offers no ability to schedule tasks based on staffing and workloads. This decision was made when the production facility was being built for a reason: the staffing levels during the plant start-up were very high, and deviations due to manpower constraints were almost non-existent. At that time, the primary function for the scheduling software was to ensure that process lead times were managed to ensure that batch transfers between vessels could be done when needed. However, due to changing production and business conditions, the staffing mix at this plant has evolved over time. The software that was designed to resolve equipment conflicts must now also level-load the schedule to reduce the workload variability within and between shifts. Because this software doesn't have the capability to manage this in a global manner, this level-loading is done by the production supervisors in an *ad hoc* and often inefficient way.

Amgen is currently considering adding this capability to its software system to address the variability in the work schedule. Although this will help to address the root causes of the scheduling deviations, there are other implementation issues that need to be considered:

- How to maintain work standards? The work standards for the different production tasks are constantly evolving as various continuous improvement projects are implemented. A mechanism for updating the standards must be available so the scheduling software reflects the reality of the production area.
- How dynamic is the software? The production supervisors will not know the exact number of staff members they have available until the start of the shift. The software must allow inputs at the start of every shift in order to recalculate a predicted schedule.

Because this type of change to the production environment breaks the assumptions used to develop the current queueing model, its effect cannot be predicted accurately with the current tool that has been developed. However, future studies should focus on the organizational issues that may arise as this software is implemented in the coming years.

6 Conclusions

A model for quantifying the degree of schedule deviation and variability has been generated that incorporates biotech-specific processing considerations into traditional queueing models. The accuracy of the model has been tested on the current production conditions and it has been used to estimate the effects of potential changes in the area. The magnitude of these changes and potential implementation issues are discussed. It is the author's hope that this model will guide future production and improvement decisions in biological manufacturing facilities.

List of Abbreviations

ARI: Amgen Rhode Island manufacturing facility

CHO cell: Chinese Hamster Ovary cell

mAb: Monoclonal Antibody

MCB: Master Cell Bank

PPH: Perfusion/Production/Harvest processing area

WCB: Working Cell Bank

Bibliography

1. **Heizer, Jay and Render, Barry.** *Operations Management, Ninth Edition.* Upper Saddle River : Pearson Education Inc., 2008.

2. **Brown, Steven, et al.** Queueing Model Improves IBM's Semiconductor Capacity and Lead-Time Management. *INTERFACES.* 2010, Vol. 40, 5.

3. *Congestion and Complexity Costs in a Plant with Fixed Resources that Strives to Make Schedule.* **Lovejoy, William and Sethuraman, Kannan.** 2000, Manufacturing and Service Operations Management, pp. 221-239.

4. *Guide to Fermentation and Cell Culture, 3rd Edition.* Duluth : s.n., Mar 1996. Biopharm International. pp. 1-31.

5. **Pasenak, David.** *The Conclusion of a Biologic's Lifecyle: Manufacturing Sourcing Strategies on the Eve of Follow-On Biologics.* Cambridge : MIT, 2008.

6. **Pollack, Andrew.** Amgen Reports its Takeover of Immunex. *The New York Times.* July 17, 2002.

7. **St. Clair, William, Pisetsky, David and Haynes, Barton.** *Rheumatoid Arthritis.* Philadelphia : Lippincott, Williams, and Wilkins, 2004.

8. **Amgen.** Immunex Announces Next Step to Boost ENBREL Production. [Online] August 14, 2001. http://www.amgen.com/pdfs/immunex/pressRelease010814.pdf.

9. —. Amgen 2009 Annual Report and Financial Summary. [Online] 2009. http://phx.corporate-ir.net/External.File?item=UGFyZW50SUQ9MzczNzN8Q2hpbGRJRD0tMXxUeXBlPTM=&t=1.

10. *Optimal Production, Maintenance, and Lockout/Tagout Control Policies in Manufacturing Systems.* **Charlot, E., Kenne, J.P. and Nadeau, S.** 2007, International Journal of Production Economics, pp. 435-450.

11. **Hopp, Wallace.** *Manufacturing Physics.* Chicago : Irwin, 1996.

12. **Imai, Masaaki.** *Gemba Kaizen: A commonsense, low-cost approach to management.* New York : McGraw-Hill, 1997.

13. **Adan, Ivo and Jacques Resing.** Lecture Notes on Queueing Theory. [Online] February 28, 2002. [Cited: January 15, 2011.] http://www.win.tue.nl/~iadan/queueing.pdf.

14. **Supervisor.** [interv.] Michael Donohue. March 15, 2011.

15. **Snyder, Sophia.** *IBISWorld Industry Report: Biotechnology in the US.* Santa Monica : IBISWorld, 2010.