Flexible Manufacturing Systems and Quantitative Analysis of Flexible Manufacturing Systems

Slobodan Jovanović
PhD Student University Singidunum
Belgrade, Serbia

ABSTRACT
A flexible manufacturing system is a high automated GT machine cell, consisting of a group of processing workstations (usually CNC machine tools), interconnected by an automated material handling and storage system, and controlled by a distributed computer system. The reason the FMS is called flexible is that it is capable of processing a variety of different part styles simultaneously at the various workstations, and the mix of part styles and quantities of productions can be adjusted in response to changing demand patterns. To develop the concept of flexibility in an automated manufacturing system, consider a machine cell consisting of two CNC machine tools are loaded and unloaded by in industrial robots.

Keywords
flexible manufacturing systems, automated

1. INTRODUCTION
A more appropriate term for FMS would be flexible automated manufacturing system. The use of the word "automated" would distinguish this type of production technology from other manufacturing systems that are flexible but not automated, such as a manned GT machine cell. The word "flexible" would distinguish it from other manufacturing systems that are highly automated but not flexible, such as a conventional transfer line.

In that discussion, we identified three capabilities that a manufacturing system must possess in order to be flexible:

1) the ability to identify and distinguish among the different incoming part or product styles processed by the system,
2) quick changeover of operating instructions, and,
3) quick changeover of physical setup.

To develop the concept of flexibility in an automated manufacturing system, consider a machine cell consisting of two CNC machine tools that are loaded and unloaded by an industrial robot from a parts carousel [1]. One might argue yes, it is flexible since the cell consists of CNC machine tools, and CNC machines are flexible because they can be programmed to machine different part configurations. However, if the cell only operates in a batch mode, in which the same part style is produced by both machines in lots of several hundred units, then this does not qualify as flexible manufacturing. To qualify as being flexible, a manufacturing system should satisfy several criteria. The following are four reasonable tests of flexibility in an automated manufacturing system:

1) Part variety test. Can the system process different part styles in a non-batch mode?
2) Schedule change test. Can the system readily accept changes in production schedule, that is, changes in part mix and/or production quantities?
3) Error recovery test. Can the system recover gracefully from equipment malfunctions and breakdowns, so that production is not completely disrupted?
4) New part test. Can new part designs be introduced into the existing product mix with relative ease?

1.1 Types of FMS
Having considered the issue of flexibility, let us now consider the various types of flexible manufacturing systems. Each FMS is designed for a specific application, that is, a specific family of parts and processes (is unique). Two other ways to classify flexible manufacturing systems are by number of machines and level of flexibility.

Number of Machines. Flexible manufacturing systems can be distinguished according to the number of machines in the system by following categories:

1) single machine cell,
2) flexible manufacturing cell, and
3) flexible manufacturing system.

A single machine cell consists of one CNC machining center combined with a parts storage system for unattended operation, as in Figure 1. Completed parts are periodically unloaded from the parts storage unit, and raw work parts are loaded into it. The cell can be designed to operate in a batch mode, a flexible mode, or a combination of the two. When operated in a batch mode, the machine processes parts of a single style in specified lot sizes and is then changed over to process a batch of the next part style. When operated in a flexible mode, the system satisfies three of the four flexibility tests. It is capable of (1) processing different part styles, (2) responding to changes in production schedule, and (4) accepting new part introductions. Criterion (3), error recovery, cannot be satisfied because if the single machine breaks down, production stops.

A flexible manufacturing cell (FMC) consists of two or three processing workstations (typically CNC machining centers or turning centers) plus a parts handling system. The parts handling system is connected to a load/unload station. The handling system usually includes a limited parts storage capacity.

A flexible manufacturing system (FMS) has four or more processing stations connected mechanically by a common parts handling system and electronically by a distributed computer system.
1.2 Level of Flexibility

Another way to classify flexible manufacturing systems is according to the level of flexibility designed into the system. This method of classification can be applied to systems with any number of workstations, but its application seems most common with FMCs and FMSs. The two categories of flexibility are:

1) dedicated, and
2) random-order.

We see in these two system types the trade-off between flexibility and productivity. The dedicated FMS is less flexible but capable of higher production rates. The random-order FMS is more flexible but at the cost of lower production rates. A comparison of these two FMS types is presented in Figure 2. Table 1 presents a comparison of the dedicated FMS and random-order FMS in terms of the four flexibility tests.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedicated FMS</td>
<td>-Limited</td>
<td>-Limited changes can be tolerated</td>
<td>-Limited by sequential processes</td>
<td>-No</td>
</tr>
<tr>
<td></td>
<td>-All parts known in advance</td>
<td></td>
<td></td>
<td>-New part introductions difficult</td>
</tr>
<tr>
<td>Random-order FMS</td>
<td>-Yes</td>
<td>-Frequent and significant changes possible</td>
<td>-Machine redundancy minimizes effect of machine breakdowns</td>
<td>-Yes</td>
</tr>
<tr>
<td></td>
<td>-Substantial part variations</td>
<td></td>
<td></td>
<td>-System designed for new part introductions</td>
</tr>
</tbody>
</table>
2. FMS COMPONENTS
As indicated in our definition, there are several basic components of a flexible manufacturing system [3]:

1) workstations,
2) material handling and storage system, and
3) computer control system

In addition, even though a FMS is highly automated, (4) people are required to manage and operate the system.

2.1 Workstations
The processing or assembly equipment used in a flexible manufacturing system depends on the type of work accomplished by the system. In a system designed for machining operations, the principal types of processing station are CNC machine tools. Following are the types of workstations typically found in a FMS:

- load/unload stations,
- machining stations,
- odder processing stations,
- assembly stations,
- odder stations and equipment.

2.2 Material Handling and Storage System
The second major component of a FMS is its material handling and storage system. In this section we discuss the functions of the handling system, material handling equipment typically used in an FMS, and types of FMS layout.

Functions of the Handling System. The material handling and storage system in a flexible manufacturing system performs the following functions:

- Allows random, independent movement of workparts between stations. Parts must be capable of moving from any machine in the system to any other machine, in order to provide various routing alternatives for the different parts and to make machine substitutions when certain stations are busy.
- Enables handling of a variety of work part configurations. For prismatic parts (nonrotational parts), this is usually accomplished by using modular pallet fixtures in the handling system.

- Provides temporary storage. The number of parts in the FMS will typically exceed the number of parts actually being processed at any moment. Thus, each station has a small queue of parts, perhaps only one part, waiting to be processed, which helps to maintain high machine utilization.
- Provides convenient access for loading and unloading work parts. The handling system must include locations for load/unload stations.
- Creates compatibility with computer control. The handling system must be under the direct control of the computer system which directs it to the various workstations, load/unload stations, and storage areas.

- Material Handling Equipment. The types of material handling systems used to transfer parts between stations in a FMS include a variety of conventional material transport equipment, and industrial robots. The material handling function in a FMS is often shared between two systems: (1) a primary handling system and (2) a secondary handling system. The primary handling system establishes the basic layout of the FMS and is responsible for moving parts between stations in the system. FMS layouts are discussed later in this section.

- FMS Layout Configurations. The material handling system establishes the FMS layout. Most layout configurations found in today’s flexible manufacturing systems can be classified into five categories:

1) in-line layout,
2) loop layout,
3) ladder layout,
4) open field layout, and
5) robot-centered cell.

The first categories of FMS as shown in Figure 3 (a), (b).
Fig 3 FMS in-line layouts: a) one directions flow similar to a transfer line and b) linear transfer system with secondary parts handling system at each station to flow in two directions.

Fig 4 FMS ladder layout

Fig 5 FMS open field layout
2.3 Computer Control System
The FMS includes a distributed computer system that is interfaced to the workstations, material handling system, and other hardware components. A typical FMS computer system consists of a central computer and microcomputers controlling the individual machines and other components. The central computer coordinates the activities of the components to achieve smooth overall operation of the system. Functions performed by the FMS computer control system can be grouped into the following categories:

1) Workstation control,
2) Distribution of control instructions to workstations,
3) Production control,
4) Traffic control,
5) Shuttle control,
6) Workpiece control,
7) Tool control. Tool control is concerned with managing two aspects of the cutting tools:
   • Tool locations and
   • Tool life monitoring.
8) Performance monitoring and reporting, and
9) Diagnostics.

3. MATHEMATICAL MODEL FPS
3.1 Quantitative analysis of Flexible Manufacturing System
Flexible manufacturing systems have constituted an active area of interest in operations research, and many of the important contributions. FMS analysis techniques can be classified into:

1) deterministic models,
2) queuing models,
3) discrete event simulation, and
4) other approaches, including heuristics.

3.1.1 Deterministic model
Deterministic models are useful in obtaining starting estimates of system performance. We present a deterministic model that is useful in the beginning stages of FMS design to provide rough estimates of system parameters such as production rate, capacity, and utilization. Deterministic models do not permit evaluation of operating characteristics such as the build-up of queues and other dynamics that can impair system performance. Consequently, deterministic models tend to overestimate FMS performance. On the other hand, if actual system performance is much lower than the estimates provided by these models, it may be a sign of either poor system design or poor management of FMS operations.

3.1.2 Queuing Model
Queuing models can be used to describe some of the dynamics not accounted for in deterministic approaches. These models are based on the mathematical theory of queues. They permit the inclusion of queues, but only in a general way and for relatively simple system configurations. The performance measures that are calculated are usually average values for steady-state operation of the system.

3.1.3 Discrete Event Simulation Model
In the later stages of design, discrete event simulation probably offers the most accurate method for modeling the specific aspects of a given flexible manufacturing system [4]. The computer model can be constructed to closely resemble the details of a complex FMS operation. Characteristics such as layout configuration, number of pallets in the system, and production scheduling rules can be incorporated into the FMS simulation model. Indeed, the simulation can be helpful in determining optimum values for these parameters.

3.2 Bottleneck Model
Important aspects of FMS performance can be mathematically described by a deterministic model called the bottleneck model, developed by Solberg [5]. Although it has the limitations of a deterministic approach, the bottleneck model is simple and intuitive. It can be used to provide starting estimates of FMS design parameters such as production rate, number of workstations, and similar measures. The term bottleneck refers to the fact that the output of the production system has an upper limit, given that the product mix flowing through the system is fixed. The model can be applied to any production system that possesses this bottleneck feature, for example, a manually operated machine cell or a production job shop. It is not limited to flexible manufacturing systems.

Terminology and Symbols. Let us define the features, terms, and symbols for the bottleneck model, as they might be applied to a flexible manufacturing system:

- **Part mix.** The mix of the various part or product styles produced by the system is defined by \( p_j \), where \( p_j \) the fraction of the total system output that is of style \( j \) [1]. The subscript \( j = 1, 2, \ldots, P \), where \( P \) the total number of different part styles made in the FMS during the time period of interest.

\[
\sum_{j=1}^{P} p_j = 1.0
\]  

- **Workstations and servers.** The flexible production system has a number of distinctly different workstations \( n \). In the terminology of the bottleneck model, each workstation may have more than one server, which simply means that it is possible to have two or more machines capable of performing the same operations. Using the terms **stations** and **servers** in the bottleneck model is a precise way of distinguishing between machines that accomplish identical operations and those that accomplish different operations. Let \( s_i \) the number of servers at workstation \( i \), where \( i = 1, 2, \ldots, n \). We include the load/unload station as one of the stations in the FMS.

- **Process routing.** For each part or product, the process routing defines the sequence of operations, the workstations where operations are performed, and the associated processing times. The sequence includes the loading operation at the beginning of processing on the FMS and the unloading operation at the end of processing. Let \( t_{ijk} \) — the processing time, which is the total time that a production unit occupies a given workstation server, not counting any waiting time at the station.

\[
t_{ijk}
\]

where:

- \( i \) – refers to the station,
- \( j \) – refers to the part or product,
- \( k \) – refers to the sequence of operations in the process routing.

- **Work handling system.** The material handling system used to transport parts or products within the FMS can
be considered to be a special case of a workstation. Let us designate it as station \( n + 1 \), and the number of carriers in the system (e.g., conveyor carts, AGVs, monorail vehicles, etc.) is analogous to the number of servers in a regular workstation. Let \( s_{n+1} \) the number of carriers in the FMS handling system.

- **Transport time.** Let \( t_{n+1} \), the mean transport time required to move a part from one workstation to the next station in the process routing. This value could be computed for each individual transport based on transport velocity and distances between stations in the FMS, but it is more convenient to simply use an average transport time for all moves in the FMS.

- **Operation frequency.** The operation frequency is defined as the expected number of times a given operation in the process routing is performed for each work unit. For example, an inspection might be performed on a sampling basis, once every four units; hence, the frequency for this operation would be 0.25. In other cases, the part may have an operation frequency greater than 1.0, for example, for a calibration procedure that may have to be performed more than once on average to be completely effective.

\[ f_{ijk} \]

where:
- \( i \) – refers to the station,
- \( j \) – refers to the process plan,
- \( k \) – refers to the operations frequency.

### 3.2.1 Operational Parameters for FMS

Using the above terms, we can now define certain average operational parameters of the production system. The average workload for a given station is defined as the mean total time spent at the station per part. It is calculated as

\[ WL_i = \sum_j \sum_k t_{ijk} f_{ijk} p_j \]  \hspace{1cm} (2)

where:
- \( WL_i \) – average workload for station \( i \),
- \( t_{ijk} \) - processing time for operation \( k \) in part \( j \) at station \( i \),
- \( p_j \) - part mix fraction for part \( j \).

The workload of the handling system is the mean transport time multiplied by the average number of transports required to complete the processing of a work part. The average number of transports is equal to the mean number of operations in the process routing minus one. That is,

\[ n_t = \sum_i \sum_j \sum_k f_{ijk} p_j - 1 \]  \hspace{1cm} (3)

where:
- \( n_t \) - mean number of transports

### 3.2.2 System Performance Measures

Important measures for assessing the performance of a flexible manufacturing system include production rate of all parts, production rate of each part style, utilization of the different workstations, and number of busy servers at each workstation. These measures can be calculated under the assumption that the FMS is producing at its maximum possible rate. This rate is constrained by the bottleneck station in the system, which is the station with the highest workload per server. The workload per server is simply the ratio \( WL_i / s_i \) for each station. Thus, the bottleneck is identified by finding the maximum value of the ratio amongst all stations. The comparison must include the handling system, since it might be the bottleneck in the system.

Let \( WL^* \) and \( s^* \) equal the workload and number of servers, respectively, for the bottleneck station. The maximum production rate of all parts of the FMS can be determined as the ratio of \( s^* \) to \( WL^* \). Let us refer to it as the maximum production rate, because it is limited by the capacity of the bottleneck station,

\[ R_p^* = \frac{s^*}{WL^*} \]  \hspace{1cm} (4)

where:
- \( R_p^* \) - maximum production rate of all part styles product by the system, which is determined by the capacity of the bottleneck station (pc/min),
- \( s^* \) - number of servers at the bottleneck station,
- \( WL^* \) - workload at the bottleneck station (min/pc).

The value of \( R_p^* \) includes parts of all styles produced in the system. Individual part production rates can be obtained by multiplying \( R_p^* \) by the respective part mix ratios. That is,

\[ R_{pj}^* = p_j \frac{s^*}{WL^*} \]  \hspace{1cm} (5)

where:
- \( R_{pj}^* \) - maximum production rate of part style \( j \) (pc/min),
- \( p_j \) - part mix fraction for part style \( j \).

The mean utilization of each workstation is the proportion of time that the servers at the station are working and not idle. This can be computed as:

\[ U_i = \frac{WL_i}{s_i} \frac{R_p^*}{\bar{R}_p} \]  \hspace{1cm} (6)

where:
- \( U_i \) -utilization at station \( i \),
- \( WL_i \)-workload of station \( i \) (min/pc),
- \( s_i \) -number servers at station \( i \)
- \( \bar{R}_p \) -overall production rate, (deo /min).

The utilization of the bottleneck station is 100% at\( R_p^* \). To obtain the average station utilization, one simply compures the average values for all stations, including the transport sxstems. This can be calculated as

\[ \bar{U} = \frac{\sum_i U_i}{n} \]  \hspace{1cm} (7)

where:
- \( \bar{U} \) -unweighted average of the workstation utilizations.

A more useful measure of overall FMS utilization can be obtained using a weighted average, where the weighting is based on the number of servers at each station for the \( n \) regular stations in system, and the transport system is omitted from the average. The argument for omitting the transport system is that the utilization of the processing stations is the important measure of FMS utilization. The purpose of the transport system is to serve the processing stations, and therefore its utilization should not be included in the average. The overall FMS utilization is calculated as

\[ \bar{U}_s = \frac{\sum_{i=1}^n \frac{U_i}{s_i}}{\sum_{i=1}^n s_i} \]  \hspace{1cm} (8)


where:
\[ U_i \] - overall FMS utilization,
\[ s_i \] - number of servers at station \( i \),
\[ U_i^* \] - utilization of station \( i \).

Finally, we are interested in the number of busy servers at each station. All of the servers at the bottleneck station are busy at the maximum production rate, but the servers at the other stations are idle some of the time. The values can be calculated as
\[ BS_i = WL_i \left( R_p^* \right) = WL_i \frac{\lambda^*}{W_L^*} \] (9)
where:
\[ BS_i \] - number of the busy servers on average at station \( i \)
\[ WL_i \] - workload at station \( i \).

### 3.3 Extended Bottleneck Model
The bottleneck model assumes that the bottleneck station is utilized 100% and that there are no delays due to queues in the system. This implies on the one hand that there are a sufficient number of parts in the system to avoid starving of workstations and on the other hand that there will be no delays due to queueing. Solberg [5] argued that the assumption of 100% utilization makes the bottleneck model overly optimistic and that a queuing model which accounts for process time variations and delays would more realistically and completely describe the performance of a flexible manufacturing system.

An alternative approach, developed by Mejabi [6], addresses some of the weaknesses of the bottleneck model without resorting to queueing computations (which can be difficult). He called his approach the extended bottleneck model. This extended model assumes a closed queueing network in which there are always a certain number of work parts in the FMS. Let \( N \) this number of parts in the system. When one part is completed and exits the FMS, a new raw work part immediately enters the system, so that TV remains constant. The new part may or may not have the same process routing as the one that just departed. The process routing of the entering part is determined according to probability \( p_j \).

\( N \) plays a critical role in the operation of the production system. If \( N \) is small (say, much smaller than the number of workstations), then some of the stations will be idle due to starving, sometimes even the bottleneck station. In this case, the production rate of the FMS will be less than \( R_p^* \) calculated in Eq. (4). If \( N \) is large (say, much larger than the number of workstations), then the system will be fully loaded, with queues of parts waiting in front of the stations. In this case, \( R_p^* \) will provide a good estimate of the production capacity of the system. However, work-in-process (WIP) will be high, and manufacturing lead time (MLT) will be long.

In effect, WIP corresponds to \( N \), and MLT is the sum of processing times at the workstations, transport times between stations, and any waiting time experienced by the parts in the system. We can express MLT as
\[ MLT = \sum_{i=1}^{n} WL_i + WL_{n+1} + T_w \] (10)
where:
\[ \sum_{i=1}^{n} WL_i \] - summation of average workloads over all stations in the FMS (min),
\[ WL_{n+1} \] - workload of the part handling system (min).

Little's formula establishes the relationship between the mean expected time a unit spends in the system, the mean processing rate of units in the system, and the mean number of units in the system. It can be mathematically proved for a single-station queueing system, and its general validity is accepted for multi-station queueing systems. Using own symbols, Little's formula can be expressed as
\[ N = R_p \left( MLT \right) \] (12)
where:
\[ N \] - number of parts in the system (pc),
\[ R_p \] - production rate of the system (pc/min), and
\[ MLT \] - manufacturing lead time (min).

Now, let us examine the two cases:

**Case 1:** When \( N \) is small, production rate is less than in the bottleneck case because the bottleneck station is not fully utilized. In this case, the waiting time \( T_w \) of a unit is theoretically zero, and Eq. (11) reduces to
\[ MLT_1 = \sum_{i=1}^{n} WL_i + WL_{n+1} \] (13)
where the subscript in \( MLT_1 \) is used to identify case 1. Production rate can be estimated using Little's formula
\[ R_p = \frac{N}{MLT_1} \] (14)

and production rate of the individual parts are given by
\[ R_{p_j} = p_j R_p \] (15)

As indicated waiting time is assumed to be zero
\[ T_w = 0 \] (16)

**Case 2:** When \( N \) is large, the estimate of maximum production rate provided by Eq. (4) should be valid: \( R_p^* = \frac{s^*}{W_L^*} \), where the asterisk (*) denotes that production rate is constrained by the bottleneck station in the system. The production rates the individual products are given by
\[ R_{p_j}^* = p_j R_p^* \] (17)

In this case, average manufacturing lead time is evaluated using Little's formula.
Despite their limitations, the bottleneck model and extended bottleneck model can provide the following practical guidelines on the design and operation of flexible manufacturing systems:

- For a given product or part mix, the total production rate of the FMS is ultimately limited by the productive capacity of the bottleneck station, which is the station with the maximum workload per server.
- If the product or part mix ratios can be relaxed, it may be possible to increase total FMS production rate by increasing the utilization of nonbottleneck workstations.

- The number of parts in the FMS at any time should be greater than the number of servers (processing machines) in the system. A ratio of around two parts per server is probably optimum, assuming that the parts are distributed throughout the FMS to ensure that one part is waiting at every station. This is especially critical at the bottleneck station.
- If work-in-process (number of parts in the system) is kept too low, production rate of the system is impaired.
- If work-in-process is allowed to be too high, then manufacturing lead time will be long with no improvement in production rate.
- As a first approximation, the bottleneck model can be used to estimate the number of servers at each station (number of machines of each type) to achieve a specified overall production rate of the system.

### 5. REFERENCES


