Abstract: This paper explains the importance of cost-effective volume scalability of existing manufacturing systems, which enables enterprises to meet market demand in a timely manner. The system must be designed at the outset for future scalability in order to enable its rapid and cost-effective volume expansion, exactly when needed. Accordingly, this paper offers a set of principles to guide manufacturing systems design for scalability, and presents several examples.

Keywords: Manufacturing systems capacity, Economic design, CNC, Forecast, System architecture, system design.

1. INTRODUCTION

Scalability of a manufacturing system refers to the ability to adjust its production capacity to adapt to the production needs. It is an important system design characteristic that measures the ease of cost-effectively changing a given manufacturing system’s throughput to meet changes in market demands.

Prior to the early 1990’s, there was no urgent need for scalable manufacturing systems for high-volume manufacturers who enjoyed a stable growing market with long product lifetimes and produced using fixed transfer lines[1]. Driven by global competition, manufacturing companies in the twenty-first century are facing increasingly frequent and unpredictable market changes, including rapid introduction of new products, and constantly varying product demand. To remain competitive, companies must design manufacturing systems that not only produce high-quality products at low cost, but also respond to market changes in an economical way [2].

Reconfigurable manufacturing systems (RMS) have been suggested by Koren et al [3] as a solution to address the needs for meeting the changing product demands. This has been acknowledged and supported later by other researchers [4-7]. From the viewpoint of RMS, a manufacturing system should be designed in such a way that it can be rapidly and cost-effectively reconfigured to the exact capacity needed to match a new market demand.

Researchers at the NSF Engineering Research Center for Reconfigurable Manufacturing Systems (ERC/RMS) began addressing system scalability since the late 1990’s [8, 9], and issued a patent that deals with strategies to change production capacity in reconfigurable manufacturing systems (RMS) [10]. They developed one of the first algorithms that addresses capacity scalability [11], but this early algorithm was limited to upgrading the capacity of serial lines only. A more comprehensive approach was presented in [12] where scalability was analyzed as one of the critical issues in designing large, complex machining systems. Capacity scalability may be also achieved by scaling the capacity of individual pieces of equipment [13, 14]. Nevertheless, the most practical approach to system scalability is adding or removing machines to or from existing manufacturing systems, and in this case the original system layout design is critical for achieving cost-effective scalability [15].

A dynamic model for capacity scalability analysis in reconfigurable manufacturing systems is introduced in [4]. This dynamic model is associated with minimizing the delay in scaling the system’s capacity and thereby improving the RMS performance in response to sudden demand changes. In this current paper we deal with optimizing the original system layout [6] such that adding/removing machines when needed by the market demand will be done quickly and cost effectively. Simultaneously with adding/removing machines, also the material handling system must be adapted to serve the new system layout. There are cases in which Autonomous Guided Vehicles (AGVs) form the material transport system. Although AGVs facilitate the part transfer to and from the new machines, they are expensive and slow, and therefore are not regarded as a cost-effective solution.

With the advancement of machine technologies over the past decade, the production of medium-to-high volume, large size mechanical parts, such as automotive powertrain mechanical

Fig.1. Schematic layout of a reconfigurable manufacturing system
parts, such as automotive powertrain components, has undergone a transformation. Dedicated transfer lines with dedicated machine stations are being replaced with systems composed of flexible CNC machine tools [16]. As shown in Fig. 1, this system architecture is composed of multiple parallel CNC machine tools at each stage, with all machines performing exactly the same machining tasks. Such configurations of parallel identical machines in each stage, with material transfer between the stages (also known as crossover) improve throughput and reduce work-in-process inventories, and are the easiest to be reconfigured.

Each manufacturing system is designed with a specific capacity to fulfill a forecasted demand [17, 18]. However, for example, if the forecast for an annual product sale is between 230,000 and 300,000 units, marketing usually dictates building a capacity for 300,000 units. Therefore, even if a system is optimally designed, capacity may be still wasted when the real demand is significantly lower than the full planned capacity. Comparing with the entire life cycle of a manufacturing system the periods in which the system is operated at the full capacity are usually short [19]. If, however, the investment in the excess capacity (70,000 units in this example) could be delayed until it is actually needed, the system lifetime cost can be significantly reduced. A system designed for scalability means that it is designed in a way that enables a rapid capacity change to meet a larger or smaller demand, exactly when needed.

This paper discusses scalability issues of incrementally changing production capacity of manufacturing systems by adding or removing machines. Section 2 defines the system scalability and describes the concept of incrementally scaling system capacity and identifies several principles for achieving incremental scalability. Section 3 defines inputs needed for scalability planning, together with mathematical formulations to minimize the total number of machines by concurrently reconfiguring and rebalancing the system. Section 4 presents a case study that was done with industry to validate the proposed approach. Conclusions are summarized in Section 5.

2. DEFINITION OF SYSTEM SCALABILITY

Dedicated lines do not have scalable capacity and cannot cope with large fluctuations in product demand [20]. This challenge can only be met by flexible or reconfigurable manufacturing systems, which are composed of CNC machines or reconfigurable machines [21, 22], as these systems are scalable in small increments accomplished by adding or removing individual machines as a need arises. Note that the number of stages always remains unchanged during the scalability process.

We define system scalability, in percentage, as:

\[
100 \text{ – smallest incremental capacity in percentage}
\]

If the minimal capacity increment by which the system throughput can be adjusted to meet new market demand is small, then the system is highly scalable. For example, if a serial line (Fig. 2a) needs to increase its production capacity to satisfy a larger market demand, an entire new line must be added. The step-size of this addition doubles the production capacity of the system. Mathematically, the minimum increment of adding production capacity in a serial line is 100% of the system, i.e., adding a whole new line, making the scalability of a serial line 0%. Doubling the line capacity (when double capacity is not really needed) will be expensive because there is no guarantee that the extra capacity will ever be fully utilized, risking thereby a substantial financial loss. Thus, zero scalability means that in order to increase the system capacity, the entire production line must be duplicated. When markets are volatile, designing a manufacturing system with zero scalability is not a good engineering solution.

Similar scalability calculations for the other systems in Fig. 2 show: Configuration b has a scalability of 50% and Configuration c has 67% (i.e., two machines must be added to increase capacity). Configurations d and e have a scalability of 84% - the highest possible for 6-machine configurations. A minimum increment of 16% - in these cases, one machine - can be added to increase system capacity; for example, a machine can be added to stage 2 of Configurations d as shown in Fig. 2.

In this example, the configuration depicted in Fig. 2c of a two stage system with three machines per stage, might be a compromise between reasonable scalability and investment cost. In this case, if a product requires machining on both the upper and side surfaces, the three machines in the first stage might be 3-axis vertical milling machines, and the three machines in the second stage might be 3–axis horizontal milling machines. Conversely, in a parallel system, all six machines in Fig. 2e must be 5-axis milling machines – making the system much more expensive. In the system in Fig. 2c, capacity scalability must be performed in steps of 33.3% by adding one vertical machine and one horizontal machine, rather than in steps of 16.6% as with the parallel configuration. Adding a step of 16.6% in Fig. 2e means in practice adding one 5-axis machine with a large tool magazine that contains every tool needed for the whole part processing – an expensive addition.

To conclude, in general, the smallest scalability adjustment steps can be accomplished when the original system is purely parallel (e.g., Fig. 2e). However, the initial cost of a parallel system is the highest of all system configurations. In parallel configurations, each machine must perform all the
manufacturing tasks needed to complete the part. Therefore, each machine must have the entire set of tools needed to produce the whole part and should also be able to perform more functions, for which more axes of motion are needed. As a result, the capital cost per additional volume increment added to a parallel configuration is the highest of all configurations.

The following example clarifies the method of adding a small incremental capacity.

**Example:** On a system composed of six machines, as shown in Fig. 3, we have to process a part that requires 21 machining tasks of 30 seconds each, totalling 630 seconds, or 10.5 minutes, needed to machine each part. The required demand is 274 parts per 8-hour shift, namely 480 minutes. Therefore, the required cycle time is 480/274 = 1.75 minutes/part.

1. Design a scalable system configuration.
2. After one year, the demand has grown, and 320 parts per shift are needed, reducing the cycle time per part to 1.5 minutes/part. How many machines should be added, and what is the new configuration?

The cost-effective scalable system configuration is depicted in Fig. 2d, which is shown in detail in Fig. 3a. Here, each machine does seven tasks of 30 seconds each, totalling 210s per machine. When the demand grows to 320 parts/day, seven machines are needed. Only Configuration d in Fig. 2 yields the cost-effective solution by adding the new machine to Stage 2 as shown in Fig. 3b. The last task that was performed on Stage 1 is shifted from Stage 1 to Stage 2, so each machine in Stage 1 operates for 180 seconds on the part, and the first task that was performed on Stage 3 is shifted from Stage 3 to Stage 2. Two tasks were added to Stage 2, so each machine in Stage 2 will now operate for 270s on each part, and the system cycle time becomes 90s per part.

\[ L = 2, M = 3, r = 0 \]

\[ \text{cycle time} = \frac{210 \times 7}{320} = 3.907 \text{ min/part} \]

\[ \text{required system capacity} = \frac{274 \times 1.75}{320} = 1.75 \text{ min/part} \]

In this example, twenty-one equal tasks were needed to complete the part, and the system with three stages and two machines per stage was precisely balanced and perfectly scalable. Similar perfect scalability results are obtained by adding \( r \) machines to a configuration that has \( L \) stages and \( M \) machines per stage, if the number of equal tasks needed to complete the part is \( (LM + r)L \). Note that in the above example \( r = 1 \), and the specific task time (30s) does not affect the solution.

From the example, we can identify the following four principles for a manufacturing system to be scalable:

1. The system has to be reconfigurable. A RMS is designed for cost-effective adjustable production capacity to respond to imminent needs.
2. The RMS capacity is designed to be scalable in optimal increments.
3. To be scalable, the RMS may contain an economic equipment mix of flexible and reconfigurable machines.
4. To be rapidly scalable, the RMS requires additional investment on its infrastructure.

The above example presents an ideal case just for the purpose of demonstrating the concept of scalability. To implement this concept in practice, many constraints need to be taken into consideration, such as task precedence, unequal task times, and stage characteristics. These issues are discussed in the following sections.

3. FORMULATION OF SCALABILITY PLANNING

We propose below a practical method to determine the most cost-effective system reconfiguration to meet a new market demand. To perform system scalability planning, many factors need to be taken into consideration. These include a detailed process plan, setup plan, machine capability, and the number of spots reserved in the original system at each stage for adding machines, if needed. When reconfiguring an existing manufacturing system, simultaneous reconfiguration planning and system rebalancing attempts are required to maximize the system capacity, and minimize the number of machines needed to be added. In this section an optimization model for scalability planning is introduced.

3.1 Assumptions

The following assumptions are made based on the current manufacturing practice in the powertrain industry.

1. A multi-stage system with a configuration similar to that shown in Fig. 1 (or Fig. 3) is considered. Parts are moved from one stage to another through conveyors and delivered to different machines within a stage using gantries.
2. The number of stages must remain unchanged during any reconfiguration process. This keeps the system setup plan unchanged in order to avoid major adjustments of process plans, thereby minimizing the impact of system reconfiguration on product quality.
3. All the machines within the same stage perform exactly the same sequence of tasks.
iv. There are reserved spaces for adding new machines in the stages and material handlers can be extended to deliver parts to the newly added machines.

3.2 Inputs

Scalability planning requires the following inputs:

- **Configuration information**
  
  Number of stages = $L$; Number of machines in each stage = $N_s$ where $i = 1,2,...,L$; Maximum number of machines allowed in each stage $M[i], i = 1,2,...,L$.

- **Stage Characteristics**
  
  Each manufacturing stage has capabilities that are defined by a set of characteristics of the stage. These include machine tool characteristics such as functionality, power, accuracy and machining ranges, as well as fixture features such as face accessibility, which defines the faces that are accessible by the cutting tool. When a set of tasks is assigned to a stage, the necessary capabilities must fall into the set of characteristics of the stage. Assuming that the number of characteristics of each stage is $K$, a capability matrix $S$ stores all possible characteristics of each stage.

$$S[i,j] = d: d = j_k, k \text{ key characteristic of stage } i,$$

where $1 \leq i \leq L$, $1 \leq j \leq K$.

- **Manufacturing tasks**
  
  **Task Precedence**: manufacturing tasks must be performed at a certain order. Each task can only be performed after all its parent tasks have been completed. A two-dimension binary matrix $Pre[i,J \rightarrow N, 1 \rightarrow N]$, where $N$ is the number of tasks to be processed, is used to represent the precedence tree.

$$Pre[i,j] = \begin{cases} 
1, & \text{if task } i \text{ must be performed before task } j \\
0, & \text{otherwise}
\end{cases}$$

**Task Characteristics**: These include task type, access direction, dimension, accuracy and power needed to perform the task. For a task to be assigned to a stage, its characteristics must be included in the set of characteristics of that stage. Assuming the number of characteristic of each task is $R$, a task characteristic matrix $K$ is used to store the characteristic of each task.

$$K[i,j] = f: f = j_k, k \text{ key characteristic of task } i,$$

where $1 \leq i \leq N$ and $1 \leq j \leq R$.

- **Machine reliability information**
  
  Machine reliability can be expressed by two parameters: MTBF (Mean Time Between Failure), and MTTR (Mean Time to Repair).

- **Demand**
  
  The system must be reconfigured so its new capacity will fulfill the new demand, $D_{\text{new}}$.

3.3 Decision variables

Two decision variables need to be determined: (1) machine allocation array, $M[i]$, which determines how many machines are to be added to the system and where to add them, and (2) task allocation array, $T[i]$, which describes how the tasks should be reallocated when the new machines are added to or removed from the system.

$$M[i] = \text{Number of machine being added to stage } i, 1 \leq i \leq L;$$

$$M[i] > 0 \text{ for adding machines, and } M[i] < 0 \text{ for removing machines from systems}.$$

$$T[i] = s, s \text{ is an index of stage to which the task } i \text{ is assigned, } 1 \leq s \leq L.$$

3.4 Optimization Model

The objective of scalability planning is to minimize the number of machines needed to meet a new market demand. This can be expressed by Eq. (1):

$$\text{Minimize } \left( \sum_{i=1}^{L} (N_i + M[i]) \right)$$

Subject to precedence constraints, characteristic constraints, space constraints and throughput constraints.

A genetic algorithm (GA) [23] based optimization tool was developed to solve equation 1 so as to determine the optimal system reconfiguration when market demand changes.

4. CASE STUDY

This section presents a case study that we conducted with our industry partners to examine and validate the proposed approach. The case selected is the rough machining process of an automotive V6 cylinder head provided by an industrial partner of the NSF Engineering Research Center for the Reconfigurable Manufacturing Systems. There are 141 features on the part, which can be grouped into 43 machining tasks, including milling, drilling, boring, spot-facing, and tapping. The total time needed for the rough machining is 1019 seconds. Because of its complexity, this part is ideal for the study as it permits many process design solutions for different system configurations. The machines used for all stages are four-axis CNC machining centers which are capable of completing all the machining tasks. Fig. 4a shows the part and Fig. 4b shows the machine configuration. MTBF and MTTR of the CNC machines are 193 minutes and 16.7 minutes, respectively.

4.1 Baseline Configurations

For each configuration in Fig. 5, reconfigurations for three system configurations, 3x4, 4x3 and 6x2, are used as
baselines to study the scalability planning. Fig. 6 gives three system configurations and their line balancing results.

### 4.2 Scalability Results

Assume a 4x3 configuration (Fig. 5b) is currently being used to fulfill a production demand of 30JPH (jobs per hour). Also assume that a maximum of two machines can be added to each stage while the setup plan remains unchanged. When the new production demand changes to 35JPH, the proposed scalability planning algorithm found that 2 new machines need to be added to the system, as is shown in Fig. 6a. The rebalancing results per machine and per stage are shown in Fig. 6b and 6c, respectively. After adding two machines system capacity increased to 36.6JPH. Compared to duplicating a four machine serial line, the new configuration only needs two new machines to fulfill the new production demand.

For each configuration in Fig. 5, reconfigurations for adding up to 5 machines to existing systems are calculated. Fig. 7 shows the reconfigurations for three-stage system. Again, Fig. 7 shows that for a given case, machines are not evenly added to each stage. Some stages tend to require more machines than others to maintain the work load balance of the system. The number of machines and their locations to be added to the system can be optimized by the proposed method.

From the cost-effective point of view, we suggest scalability planning be performed concurrently with the design of a new manufacturing system. This way, optimal locations where future machines should be installed can be identified in advance. Thus, material handling systems can be optimized for future scalability planning to reduce the life-time investment cost.

Table 1 summarizes the system productivity of each configuration and the new productivity when 1 to 5 machines are added to existing systems. It can be seen that from Table 1 that system of 3x4 configuration gives both the highest original system throughput and the largest throughput gain per new added machine. This is because both system reliability and system balance tend to decline with the increase of number of system stages. When the production demand increases for a given existing system, Table 1 is very convenient for helping to decide how many new machines are needed and where they should be added to.

![Figure 5](image1)

**Figure 5.** Three configurations and their productivity baselines to study the scalability planning. Fig. 6 gives three system configurations and their line balancing results.

![Figure 6](image2)

**Figure 6.** Scalability planning example of increasing productivity by 5JPH

![Figure 7](image3)

**Figure 7.** Reconfigurations for scalability planning for a 3x4 system

<table>
<thead>
<tr>
<th>No.of Machines added</th>
<th>3x4</th>
<th>4x3</th>
<th>6x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>+0</td>
<td>33.1</td>
<td>30.9</td>
<td>27.9</td>
</tr>
<tr>
<td>+1</td>
<td>35.9</td>
<td>34</td>
<td>30.1</td>
</tr>
<tr>
<td>+2</td>
<td>39.1</td>
<td>36.6</td>
<td>32.3</td>
</tr>
<tr>
<td>+3</td>
<td>41.8</td>
<td>38.9</td>
<td>34.6</td>
</tr>
<tr>
<td>+4</td>
<td>44</td>
<td>41.8</td>
<td>37.1</td>
</tr>
<tr>
<td>+5</td>
<td>47.1</td>
<td>44.2</td>
<td>39.3</td>
</tr>
</tbody>
</table>

| Average throughput gain per machine | 2.85 | 2.80 | 2.24 |
5. CONCLUSION AND DISCUSSION

This paper elaborates in detail on the scalability concept as introduced in [24], and presents an original systematic approach for scalability planning – adding or subtracting the exact capacity needed to fulfill market demand. Our approach utilizes a scalability planning process that simultaneously changes the system configuration and rebalances the reconfigured system. An optimal solution based on Genetic Algorithm was developed for the scalability planning process that is subject to realistic constraints.

The proposed approach was validated through a real industrial case. Experimental results showed that the proposed approach can address the scalability planning problem cost-effectively and efficiently. This paper suggests that the scalability planning should be performed concurrently with the design of a new manufacturing system. With this approach, the material handling system can be optimized for future scalability planning to reduce the investment cost.

For the purpose of simplicity, this paper only used the total number of machines as the optimization objective. However, in real production, many other cost factors need to be taken in consideration as well. These include labor, tooling, utility, floor space, operating cost, and material handlers [22]. In addition, since a reconfiguration process usually requires shutting down the production system, an extra cost will occur due to the production loss during the reconfiguration process. From an economic point of view, all these factors should be accounted for when planning for system scalability.

6 ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support of the Engineering Research Center for Reconfigurable Manufacturing Systems (NSF Grant No. EEC95-92125) at the University of Michigan, as well as the valuable input provided by our industrial sponsors and the paper reviewers.

REFERENCES